PHYSICAL SCIENCE: A 2-SEMESTER COURSE



Machele Kindle
Bridge Valley Community and Technical College

LibreTexts™

Physical Science

Machele Kindle
Bridge Valley Community and Technical College

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden borne from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and were supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 10/20/2026

TABLE OF CONTENTS

Licensing

1: Physics

Chapter 1: Mechanics I - Motion and Forces

1.1: Kinematics

1.1.1: Introduction to One-Dimensional Kinematics

1.1.2: Displacement

1.1.3: Vectors, Scalars, and Coordinate Systems

1.1.4: Time, Velocity, and Speed

1.1.5: Acceleration

1.1.6: Motion Equations for Constant Acceleration in One Dimension

1.1.7: Falling Objects

1.1.8: Projectile Motion

1.1.9: Centripetal Acceleration

1.1.10: Kinematics (Exercise)

1.2: Dynamics

1.2.1: Introduction to Dynamics- Newton's Laws of Motion

1.2.2: Development of Force Concept

1.2.3: Newton's First Law of Motion- Inertia

1.2.4: Newton's Second Law of Motion- Force and Acceleration

1.2.5: Newton's Third Law of Motion- Symmetry in Forces

1.2.6: Normal Force and Tension

1.2.7: Spring Force- Hooke's Law

1.2.8: Friction

1.2.9: Newton's Universal Law of Gravitation

1.2.10: Centripetal Force

1.2.11: Dynamics (Exercise)

Chapter 2: Mechanics II - Energy and Momentum, Oscillations and Waves, Rotation, and Fluids

2.1: Work and Energy

2.1.1: Introduction to Work and Energy

2.1.2: Work- The Scientific Definition

2.1.3: Kinetic Energy and the Work-Energy Theorem

2.1.4: Gravitational Potential Energy

2.1.5: Conservative Forces, Potential Energy, and Conservation of Energy

2.1.6: Spring Potential Energy

2.1.7: Power

2.1.8: Work and Energy (Exercise)

2.2: Impulse and Momentum
2.2.1: Introduction to Linear Momentum and Collisions
2.2.2: Linear Momentum and Force
2.2.3: Impulse
2.2.4: Conservation of Momentum
2.2.5: Elastic Collisions in One Dimension
2.2.6: Inelastic Collisions in One Dimension
2.2.7: Impulse and Momentum (Exercise)

2.3: Oscillations and Waves
2.3.1: Introduction to Oscillatory Motion and Waves
2.3.2: Period and Frequency in Oscillations
2.3.3: Simple Harmonic Motion- A Special Periodic Motion
2.3.4: Forced Oscillations and Resonance
2.3.5: Waves
2.3.6: Wave Interference- Standing Waves and Beats
2.3.7: Sound
2.3.8: Speed of Sound, Frequency, and Wavelength
2.3.9: Doppler Effect and Sonic Booms
2.3.10: Oscillations and Waves (Exercise)

Chapter 3: Classical Physics - Thermodynamics, Electricity and Magnetism, and Light

3.1: Electricity
3.1.1: Introduction to Electricity
3.1.2: Static Electricity and Charge- Conservation of Charge
3.1.3: Coulomb's Law
3.1.4: Electric Field- Concept of a Field Revisited
3.1.5: Electric Field Lines
3.1.6: Electric Potential and Potential Energy
3.1.7: Conductors and Applications of Electrostatics
3.1.8: Current
3.1.9: Ohm's Law- Resistance and Simple Circuits
3.1.10: Electric Power and Energy
3.1.11: Resistors in Series and Parallel
3.1.12: Electric Hazards and the Human Body
3.1.13: Electricity (Exercise)

3.2: Magnetism
3.2.1: Introduction to Magnetism
3.2.2: Magnets
3.2.3: Ferromagnets and Electromagnets
3.2.4: Magnetic Fields and Magnetic Field Lines

3.2.5: Magnetic Field Strength- Force on a Moving Charge in a Magnetic Field
3.2.6: Magnetic Force on a Current-Carrying Conductor
3.2.7: Motors and Meters
3.2.8: Magnetic Fields Produced by Currents- Ampere's Law
3.2.9: Induced Voltage and Magnetic Flux
3.2.10: Faraday's Law of Induction- Lenz's Law
3.2.11: Transformers
3.2.12: Alternating Current versus Direct Current
3.2.13: Magnetism (Exercise)

3.3: Light

3.3.1: Introduction to Light
3.3.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed
3.3.3: Production and Properties of Electromagnetic Waves
3.3.4: The Electromagnetic Spectrum- an Overview
3.3.5: The Electromagnetic Spectrum- Application Notes
3.3.6: Reflection
3.3.7: Refraction
3.3.8: Dispersion- The Rainbow and Prisms
3.3.9: Polarization
3.3.10: Image Formation by Lenses
3.3.11: Image Formation by Mirrors
3.3.12: Light (Exercises)

3.4: Thermal Physics

3.4.1: Introduction to Thermal Physics
3.4.2: Temperature
3.4.3: The Ideal Gas Law
3.4.4: Heat
3.4.5: Heat Transfer Methods
3.4.6: Temperature Change and Heat Capacity
3.4.7: Phase Change and Latent Heat
3.4.8: The First Law of Thermodynamics
3.4.9: The First Law of Thermodynamics and Heat Engine Processes
3.4.10: Introduction to the Second Law of Thermodynamics- Heat Engines and Their Efficiency
3.4.11: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated
3.4.12: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy
3.4.13: Statistical Interpretation of Entropy and the Second Law of Thermodynamics- The Underlying Explanation
3.4.14: Applications of Thermodynamics- Heat Pumps and Refrigerators
3.4.15: Thermal Physics (Exercises)

2: Astronomy

Chapter 2.1: Introduction to Astronomy

2.1.1: The Basics

2.1.1.1: Module Introduction

2.1.2: Basic Objects in the Universe

2.1.1.3: Scale of the Cosmos

2.1.2: Patterns in the Sky

4.2.1: The Constellations

4.2.2: Other Patterns

Chapter 2.2: Motions of the Moon, Sun, and Stars

2.2.1: The Celestial Sphere

2.2.1.1: Motions of Objects

2.2.1.2: Our Night Sky

2.2.1.3: The Sun and the Moon

2.2.2: Equinoxes, Solstices, and Eclipses

2.2.2.1: Seasons and Seasonal Changes

5.2.2: The Moon in our Skies

2.2.2.3: Eclipses

Chapter 2.3: Light and Spectra

2.3.1: Light

2.3.1.1: Looking at a rainbow, what do you see?

2.3.1.2: Thermal Radiation

2.3.2: Spectra

2.3.2.1: The Electromagnetic Spectrum

2.3.2.2: Spectra

2.3.2.3: Moving Objects, Spectra, and the Doppler Effect

Chapter 2.4: Telescope and Observing

2.4.1: The History of the Telescope

2.4.1.1: The Telescope

2.4.1.2: Early Telescopes

2.4.1.3: Telescopic Observations

2.4.2: The Workings of the Telescope

2.4.2.1: Telescope Optical Types

2.4.2.2: Telescope Mounts

2.4.2.3: Images from the Telescope

2.4.2.4: Light Pollution

2.4.3: Space-based Telescopes

2.4.3.1: Space-based Telescope

2.4.3.2: Hubble Space Telescope

Chapter 2.5: Solar System Formation and Other Stellar Systems

2.5.1: Formation

2.5.1.1: Our Solar System

2.5.1.2: The Nebular Theory

2.5.1.3: What is a Planet?

2.5.1.4: The Nebular Theory - Other Important Evidence

2.5.1.5: The Nebular Theory - Proplyds

2.5.1.6: Systems and Extrasolar Planets

2.5.1.7: The Discovery of Extrasolar Planets

2.5.1.8: The Kepler Mission

2.5.1.9: What We Have Learned about Stellar Systems and Extrasolar Planets?

2.5.1.10: The Circumstellar Habitable Zone

2.5.1.11: Some Interesting Exoplanets

2.5.1.12: Extrasolar Planets

2.5.2: The Rocky Planets

2.5.2.1: What do you think?

2.5.2.2: Comparing Planets

2.5.2.3: Rocky Planets

2.5.2.4: Craters

2.5.2.5: Planet Mercury

2.5.2.6: Planet Venus

2.5.2.7: Planet Earth

2.5.2.8: Earth's Moon

2.5.2.9: Planet Mars

2.5.2.10: Snow on Mars

2.5.2.11: A Martian Controversy

2.5.2.12: Martian Moons

2.5.2.13: Characteristics of the Solar System's Rocky Planets

2.5.3: The Gas Giant Planets

2.5.3.1: What do you think?

2.5.3.2: Gas Giant Planets

2.5.3.3: Rings

2.5.3.4: Planet Jupiter

2.5.3.5: Jupiter's Satellites

2.5.3.6: Planet Saturn

2.5.3.7: Saturn's Rings

2.5.3.8: Saturn's Satellites

2.5.3.9: Planet Uranus

2.5.3.10: Planet Neptune
2.5.3.11: Gas Giant Planets and Select Satellites Overviews

2.5.4: Minor Bodies of the Solar System

2.5.4.1: Trans-Neptunian objects, the Kuiper Belt, and the Oort Cloud
2.5.4.2: The Oort Cloud and Kuiper Belt
2.5.4.3: What do you think?
2.5.4.4: Minor Bodies
2.5.4.5: Asteroids
2.5.4.6: Brightest Asteroids Visible from Earth
2.5.4.7: Asteroid Classification
2.5.4.8: Meteoroids, Meteors, and Meteorites
2.5.4.9: Meteorites
2.5.4.10: The Impact of Meteorites
2.5.4.11: Comets
2.5.4.12: Comet Characteristics
2.5.4.13: Images of Comets
2.5.4.14: Comets in History
2.5.4.15: Scientific Advances
2.5.4.16: Halley's Comet
2.5.4.17: A Comet Impacts a Planet
2.5.4.18: The Centaurs- Are they Asteroids or Comets?
2.5.4.19: Pluto
2.5.4.20: Eris

2.5.5: The Sun

2.5.5.1: What do you think?
2.5.5.2: Fission and Fusion
2.5.5.3: Our Star, the Sun
2.5.5.4: How the Sun Works
2.5.5.5: A "Quick Guide" to Solar Fusion—The Proton-Proton Cycle
2.5.5.6: Solar Regions
2.5.5.7: Major Solar Features
2.5.5.8: Studying the Sun
2.5.5.9: The Dark Side of the Sun
2.5.5.10: The National Solar Observatory
2.5.5.11: NASA's Heliophysics Science Division

Chapter 2.6: Stellar Properties

2.6.1: Stellar Objects

2.6.1.1: Module Introduction
2.6.1.2: Variable Stars

2.6.1.3: Cepheids
2.6.1.4: Cataclysmic Variables
2.6.1.5: Star Clusters
2.6.1.6: Consider this...
2.6.1.7: The Unknown in Astronomy
2.6.1.8: Star and Celestial Object Characteristics

2.6.2: Stellar Distances
2.6.2.1: Distances to the Stars and other Celestial Objects
2.6.2.2: Stellar Distances
2.6.2.3: Everything is Moving
2.6.2.4: How Close are the Close-by Stars?
2.6.2.5: Light-Years
2.6.2.6: Parsec Vs. Light-Year Measurement

2.6.3: Stellar Cartography
2.6.3.1: Stellar and Celestial Object Brightness
2.6.3.2: Magnitude System
2.6.3.3: Celestial cartography
2.6.3.4: Planispheres
2.6.3.5: Star Colors
2.6.3.6: Spectral Type
2.6.3.7: Sample Spectral Types

2.6.4: The Hertzsprung-Russell Diagram
2.6.4.1: The Hertzsprung-Russell Diagram
2.6.4.2: The Four Hertzsprung-Russell Diagram Stellar Groups
2.6.4.3: Types of Star Systems and Stars

Chapter 2.7: Stellar Evolution

2.7.1: Birth
2.7.1.1: Star Life
2.7.1.2: Stellar Birth
2.7.1.3: Stellar Mass
2.7.1.4: Stellar Evolution

2.7.2: Life Cycle - Low-Mass Stars
2.7.2.1: Low-Mass Stars
2.7.2.2: White Dwarf
2.7.2.3: Planetary Nebula
2.7.2.4: White Dwarfs and Neighbors
2.7.2.5: Nova, Novae, and Supernova

2.7.3: Life Cycle - High-Mass Stars
2.7.3.1: High-Mass Stars

2.7.3.2: Neutron stars
2.7.3.3: Neutron Star and Companion Star Scenario
2.7.3.4: High-Mass Star Stellar Endings
2.7.3.5: Gamma-Ray Bursts (GRBs)
2.7.3.6: Checking Out the Theory

Chapter 2.8: Galaxies
2.8.1: The Practice of Studying Galaxies
2.8.1.1: What do you think?
2.8.1.2: Galaxies
2.8.1.3: The History of Studying Galaxies
2.8.1.4: Hubble's Major Contributions
2.8.1.5: Hubble's Law
2.8.2: Galactic Classification
2.8.2.1: Classification of Galaxies
2.8.2.2: Elliptical Galaxies
2.8.2.3: Spiral Galaxies
2.8.2.4: Irregular Galaxies
2.8.2.5: Active Galaxies
2.8.2.6: Quasar-stellar Objects
2.8.2.7: Blazars- Quasar-like Object
2.8.3: Galactic Modern Theory
2.8.3.1: Galaxies, Distance, and Age
2.8.3.2: The Ages of Galaxies and What that Reveals
2.8.3.3: Galactic Groups
2.8.3.4: Dark Matter
2.8.3.5: WIMPs and String Theory
2.8.3.6: Gravitational Lensing
2.8.4: The Milky Way Galaxy
2.8.4.1: What do you think?
2.8.4.2: The Milky Way — Our Home Galaxy
2.8.4.3: Milky Way Galaxy Satellites
2.8.4.4: Milky Way Galaxy Research
2.8.4.5: Galileo Galilei, First to See the Milky Way Galaxy
2.8.4.6: The Milky Way Galaxy's Shape
2.8.4.7: Andromeda Galaxy
2.8.4.8: X-Ray and Gamma-Ray of the Milky Way Galaxy
2.8.4.9: Earth's Position within the Milky Way Galaxy
2.8.4.10: A Spinning Spiral

Chapter 2.9: Cosmology

2.9.1: History of Cosmology
2.9.1.1: What do you think?
2.9.1.2: In the beginning...
2.9.1.3: Physical Cosmology
2.9.1.4: The Big Bang in Eras
2.9.1.5: The Planck Era

2.9.2: "Big Bang" Cosmology
2.9.2.1: Big Bang Evidence
2.9.2.2: Recent Data on the Universe
2.9.2.3: Issues with the Big Bang
2.9.2.4: Beyond the Big Bang
2.9.2.5: What is Next for the Big Bang Theory?

2.9.3: Philosophy of Cosmology
2.9.3.1: Questions Regarding the Eventual Fate of the Universe
2.9.3.2: Religious or Philosophical Cosmology
2.9.3.3: Views on Creation

3: Chemistry

Chapter 3.1: Matter energy and their measurements
3.1.1: Matter and energy
3.1.2: What is chemistry?
3.1.3: Measurements
3.1.4: Significant Figures
3.1.5: Unit conversions
3.1.6: Equations and graphs
3.1.7: Density and specific gravity measurements
3.1.8: Heat and its measurements
3.1.9: Heat and changes in physical states of matter

Chapter 3.2: Elements
3.2.1: Dalton’s atomic theory
3.2.2: Subatomic particles and a modern view of an atom
3.2.3: Atoms of elements
3.2.4: The periodic table
3.2.5: Electrons in atoms
3.2.6: The periodic trends in properties of the elements

Chapter 3.3: Compounds
3.3.1: Bonding in compounds
3.3.2: Naming binary ionic compounds
3.3.3: Polyatomic ions and their compounds
3.3.4: Naming acids

3.3.5: Naming binary covalent compounds
3.3.6: Lewis structures of molecules
3.3.7: Molecular shapes –Valence shell electron pair repulsion (VSEPR) theory
3.3.8: Polarity of molecules
3.3.9: Intramolecular forces and intermolecular forces

Chapter 3.4: Stoichiometry –the quantification of chemical reactions

3.4.1: Stoichiometry
3.4.2: The mole
3.4.3: Chemical reaction
3.4.4: Patterns of chemical reactions
3.4.5: Oxidation-reduction reactions
3.4.6: Energetics of chemical reactions
3.4.7: Stoichiometric calculations

Chapter 3.5: Solutions

3.5.1: Introduction to solution
3.5.2: Solubility
3.5.3: Electrolytes
3.5.4: Concentration of solutions
3.5.5: Osmosis

Chapter 3.6: Acids and bases

3.6.1: What is an Acid and a Base?
3.6.2: Brønsted–Lowry acids and bases
3.6.3: Strength of acids and bases
3.6.4: Acid-base equilibrium
3.6.5: Dissociation of water
3.6.6: The pH
3.6.7: Acid-base reactions
3.6.8: pH Buffers

Chapter 3.7: Gases

3.7.1: Characteristics of gases
3.7.2: The pressure-volume relationship
3.7.3: The temperature-volume relationship
3.7.4: The pressure-temperature relationship
3.7.5: The combined gas law
3.7.6: The volume-amount relationship
3.7.7: Ideal gas law
3.7.8: Dalton's law of partial pressure

Chapter 3.8: Organic Chemistry

3.8.1: Introduction

3.8.2: Hydrocarbons
3.8.3: Alcohols and Ethers
3.8.4: Aldehydes, Ketones, Carboxylic Acids, and Esters
3.8.5: Amines and Amides
3.8.6: Key Terms
3.8.7: Summary
3.8.8: Exercises

Chapter 3.9: Nuclear chemistry
3.9.1: Introduction to nuclear chemistry
3.9.2: Radioactivity
3.9.3: Half-life of radioisotopes
3.9.4: Radiation measurements
3.9.5: Ionizing radiation exposures
3.9.6: Medical uses of radioisotopes
3.9.7: Making radioisotopes for medical uses
3.9.8: Nuclear fusion and fission

4: Geology

Chapter 4.1: Science of Geology
4.1.1: What is Geology?
4.1.2: The Science of Geology
4.1.3: The Branches of Geology
4.1.4: Scientific Processes in Geology
4.1.5: The Nature of Science
4.1.6: Scientific Tools
4.1.7: Geologic Tools
4.1.8: Maps
4.1.9: Location and Direction
4.1.10: Scientific Models
4.1.11: Geologic Modelling

Chapter 4.2: Rock Forming Minerals
4.2.1: The Building Blocks of Matter
4.2.2: Atomic Bonds
4.2.3: Physical Characteristics of Minerals
4.2.4: Classifying Minerals
4.2.5: Abundance of Elements in Earth's Crust
4.2.6: Silicate Minerals and the Silicate Tetrahedron

Chapter 4.3: Rocks and the Rock Cycle
4.3.1: Characteristics of Igneous Rocks
4.3.2: How Are Igneous Rocks Classified?

4.3.3: Identifying Igneous Rocks
4.3.4: Uses of Igneous Rocks
4.3.5: Characteristics of Sedimentary Rocks
4.3.6: Characteristics of Metamorphic Rocks

Chapter 4.4: Weathering and Soil Formation

4.4.1: What Is Weathering?
4.4.2: Mechanical Weathering
4.4.3: Chemical Weathering
4.4.4: Influences on Weathering
4.4.5: Soil Texture and Composition
4.4.6: Soil Horizons and Profiles
4.4.7: Types of Soils
4.4.8: Causes of Soil Erosion
4.4.9: Avoiding Soil Loss

Chapter 4.5: Plate Tectonics

4.5.1: Continental Drift
4.5.2: Wegener and the Continental Drift Hypothesis
4.5.3: Theory of Plate Tectonics
4.5.4: Developing the Theory
4.5.5: Supercontinents
4.5.6: Volcanic Arcs
4.5.7: Hot Spots

Chapter 4.6: Earth's Interior

4.6.1: Characteristics of the Crust
4.6.2: Characteristics of the Mantle
4.6.3: Characteristics of the Inner and Outer Core
4.6.4: The Lithosphere and Asthenosphere
4.6.5: Studying the Earth's Interior
4.6.6: Understanding the Earth's Interior

Chapter 4.7: Crustal Deformation

4.7.1: Stress In Earth's Crust
4.7.2: Strain
4.7.3: Folds
4.7.4: Faults
4.7.5: Mountains

Chapter 4.8: Earthquakes

4.8.1: The Nature of Earthquakes
4.8.2: Causes of Earthquakes
4.8.3: Common Locations of Earthquakes

4.8.4: The Ring of Fire
4.8.5: Body Waves
4.8.6: Surface Waves
4.8.7: Measuring Earthquakes
4.8.8: Magnitude versus Intensity
4.8.9: Earthquake Hazards
4.8.10: Earthquake Prediction
4.8.11: Seismicity and Earthquake Prediction

Chapter 4.9: Volcanoes

4.9.1: Types of Volcanoes
4.9.2: Types of Eruptions
4.9.3: Supervolcanoes
4.9.4: Volcanoes at Plate Boundaries
4.9.5: Volcanoes Hotspots
4.9.6: Pyroclastic Deposits
4.9.7: Dikes and Sills
4.9.8: Volcanic Landforms
4.9.9: Igneous Features
4.9.10: Volcanic Gases, Pyroclastic Flow, and Tephra
4.9.11: Lahars, Landslides, and Lavaflow
4.9.12: Signs of an Eruption
4.9.13: Tools to Monitor Volcanoes

Chapter 4.10: Mass Movement

4.10.1: Slides
4.10.2: Falls
4.10.3: Slump and Creep
4.10.4: Natural Mass Wasting
4.10.5: Human Impact on Mass Wasting

Chapter 4.11: Hydrology

4.11.1: Phases of the Hydrologic Cycle
4.11.2: Types of Streams and Rivers
4.11.3: Drainage Basins
4.11.4: Geologic Processes and Flowing Water
4.11.5: Groundwater
4.11.6: Porosity and Permeability
4.11.7: Aquifers and Well-Potential
4.11.8: Groundwater Withdrawal
4.11.9: Quality of Groundwater
4.11.10: Karst Topography
4.11.11: Geysers

4.11.12: Springs
4.11.13: Geologic Features
4.11.14: Oceans and Shorelines
4.11.15: Ocean Movement
4.11.16: Shorelines
4.11.17: Coastal Hazards
4.11.18: Human Modifications of Coastal Processes

Chapter 4.12: Geological Resources

4.12.1: Ore Genesis
4.12.2: Petroleum
4.12.3: Natural Gas
4.12.4: Coal
4.12.5: Mining
4.12.6: Effects of Mining
4.12.7: Earth's Energy
4.12.8: Energy Consumption on Earth
4.12.9: Use and Conservation of Resources

Index

Glossary

Detailed Licensing

This page titled Physical Science: A Two-Semester Course is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

Licensing

A detailed breakdown of this resource's licensing can be found in ***Back Matter/Detailed Licensing***.

This page titled Licensing is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

1: PHYSICS

Chapter 1: Mechanics I - Motion and Forces

1.1: Kinematics

1.1.1: Introduction to One-Dimensional Kinematics

1.1.2: Displacement

1.1.3: Vectors, Scalars, and Coordinate Systems

1.1.4: Time, Velocity, and Speed

1.1.5: Acceleration

1.1.6: Motion Equations for Constant Acceleration in One Dimension

1.1.7: Falling Objects

1.1.8: Projectile Motion

1.1.9: Centripetal Acceleration

1.1.10: Kinematics (Exercise)

1.2: Dynamics

1.2.1: Introduction to Dynamics- Newton's Laws of Motion

1.2.2: Development of Force Concept

1.2.3: Newton's First Law of Motion- Inertia

1.2.4: Newton's Second Law of Motion- Force and Acceleration

1.2.5: Newton's Third Law of Motion- Symmetry in Forces

1.2.6: Normal Force and Tension

1.2.7: Spring Force- Hooke's Law

1.2.8: Friction

1.2.9: Newton's Universal Law of Gravitation

1.2.10: Centripetal Force

1.2.11: Dynamics (Exercise)

Chapter 2: Mechanics II - Energy and Momentum, Oscillations and Waves, Rotation, and Fluids

2.1: Work and Energy

2.1.1: Introduction to Work and Energy

2.1.2: Work- The Scientific Definition

2.1.3: Kinetic Energy and the Work-Energy Theorem

2.1.4: Gravitational Potential Energy

2.1.5: Conservative Forces, Potential Energy, and Conservation of Energy

2.1.6: Spring Potential Energy

2.1.7: Power

2.1.8: Work and Energy (Exercise)

2.2: Impulse and Momentum

2.2.1: Introduction to Linear Momentum and Collisions

2.2.2: Linear Momentum and Force
2.2.3: Impulse
2.2.4: Conservation of Momentum
2.2.5: Elastic Collisions in One Dimension
2.2.6: Inelastic Collisions in One Dimension
2.2.7: Impulse and Momentum (Exercise)

2.3: Oscillations and Waves

2.3.1: Introduction to Oscillatory Motion and Waves
2.3.2: Period and Frequency in Oscillations
2.3.3: Simple Harmonic Motion- A Special Periodic Motion
2.3.4: Forced Oscillations and Resonance
2.3.5: Waves
2.3.6: Wave Interference- Standing Waves and Beats
2.3.7: Sound
2.3.8: Speed of Sound, Frequency, and Wavelength
2.3.9: Doppler Effect and Sonic Booms
2.3.10: Oscillations and Waves (Exercise)

Chapter 3: Classical Physics - Thermodynamics, Electricity and Magnetism, and Light

3.1: Electricity

3.1.1: Introduction to Electricity
3.1.2: Static Electricity and Charge- Conservation of Charge
3.1.3: Coulomb's Law
3.1.4: Electric Field- Concept of a Field Revisited
3.1.5: Electric Field Lines
3.1.6: Electric Potential and Potential Energy
3.1.7: Conductors and Applications of Electrostatics
3.1.8: Current
3.1.9: Ohm's Law- Resistance and Simple Circuits
3.1.10: Electric Power and Energy
3.1.11: Resistors in Series and Parallel
3.1.12: Electric Hazards and the Human Body
3.1.13: Electricity (Exercise)

3.2: Magnetism

3.2.1: Introduction to Magnetism
3.2.2: Magnets
3.2.3: Ferromagnets and Electromagnets
3.2.4: Magnetic Fields and Magnetic Field Lines
3.2.5: Magnetic Field Strength- Force on a Moving Charge in a Magnetic Field
3.2.6: Magnetic Force on a Current-Carrying Conductor

3.2.7: Motors and Meters
3.2.8: Magnetic Fields Produced by Currents- Ampere's Law
3.2.9: Induced Voltage and Magnetic Flux
3.2.10: Faraday's Law of Induction- Lenz's Law
3.2.11: Transformers
3.2.12: Alternating Current versus Direct Current
3.2.13: Magnetism (Exercise)

3.3: Light

3.3.1: Introduction to Light
3.3.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed
3.3.3: Production and Properties of Electromagnetic Waves
3.3.4: The Electromagnetic Spectrum- an Overview
3.3.5: The Electromagnetic Spectrum- Application Notes
3.3.6: Reflection
3.3.7: Refraction
3.3.8: Dispersion- The Rainbow and Prisms
3.3.9: Polarization
3.3.10: Image Formation by Lenses
3.3.11: Image Formation by Mirrors
3.3.12: Light (Exercises)

3.4: Thermal Physics

3.4.1: Introduction to Thermal Physics
3.4.2: Temperature
3.4.3: The Ideal Gas Law
3.4.4: Heat
3.4.5: Heat Transfer Methods
3.4.6: Temperature Change and Heat Capacity
3.4.7: Phase Change and Latent Heat
3.4.8: The First Law of Thermodynamics
3.4.9: The First Law of Thermodynamics and Heat Engine Processes
3.4.10: Introduction to the Second Law of Thermodynamics- Heat Engines and Their Efficiency
3.4.11: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated
3.4.12: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy
3.4.13: Statistical Interpretation of Entropy and the Second Law of Thermodynamics- The Underlying Explanation
3.4.14: Applications of Thermodynamics- Heat Pumps and Refrigerators
3.4.15: Thermal Physics (Exercises)

This page titled 1: Physics is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 1: MECHANICS I - MOTION AND FORCES

1.1: Kinematics

1.1.1: Introduction to One-Dimensional Kinematics

1.1.2: Displacement

1.1.3: Vectors, Scalars, and Coordinate Systems

1.1.4: Time, Velocity, and Speed

1.1.5: Acceleration

1.1.6: Motion Equations for Constant Acceleration in One Dimension

1.1.7: Falling Objects

1.1.8: Projectile Motion

1.1.9: Centripetal Acceleration

1.1.10: Kinematics (Exercise)

1.2: Dynamics

1.2.1: Introduction to Dynamics- Newton’s Laws of Motion

1.2.2: Development of Force Concept

1.2.3: Newton’s First Law of Motion- Inertia

1.2.4: Newton's Second Law of Motion- Force and Acceleration

1.2.5: Newton’s Third Law of Motion- Symmetry in Forces

1.2.6: Normal Force and Tension

1.2.7: Spring Force- Hooke’s Law

1.2.8: Friction

1.2.9: Newton’s Universal Law of Gravitation

1.2.10: Centripetal Force

1.2.11: Dynamics (Exercise)

This page titled Chapter 1: Mechanics I - Motion and Forces is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

- Unit 1: Mechanics I - Motion and Forces by Andrew Park is licensed CC BY 4.0.

1.1: KINEMATICS

1.1.1: Introduction to One-Dimensional Kinematics
1.1.2: Displacement
1.1.3: Vectors, Scalars, and Coordinate Systems
1.1.4: Time, Velocity, and Speed
1.1.5: Acceleration
1.1.6: Motion Equations for Constant Acceleration in One Dimension
1.1.7: Falling Objects
1.1.8: Projectile Motion
1.1.9: Centripetal Acceleration
1.1.10: Kinematics (Exercise)

This page titled 1.1: Kinematics is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

1.1.1: Introduction to One-Dimensional Kinematics



Figure 1.1.1.1: The motion of an American kestrel through the air can be described by the bird's displacement, speed, velocity, and acceleration. When it flies in a straight line without any change in direction, its motion is said to be one dimensional. (credit: Vince Maidens, Wikimedia Commons)

Objects are in motion everywhere we look. Everything from a tennis game to a space-probe flyby of the planet Neptune involves motion. When you are resting, your heart moves blood through your veins. And even in inanimate objects, there is continuous motion in the vibrations of atoms and molecules. Questions about motion are interesting in and of themselves: *How long will it take for a space probe to get to Mars?* But an understanding of motion is also key to understanding other concepts in physics. An understanding of acceleration, for example, is crucial to the study of force.

Our formal study of physics begins with **kinematics** which is defined as the *study of motion without considering its causes*. The word "kinematics" comes from a Greek term meaning motion and is related to other English words such as "cinema" (movies) and "kinesiology" (the study of human motion). In one-dimensional kinematics we will study only the *motion* of a football, for example, without worrying about what forces cause or change its motion. Such considerations come in other chapters. In this chapter, we examine the simplest type of motion—namely, motion along a straight line, or one-dimensional motion.

This page titled 1.1.1: Introduction to One-Dimensional Kinematics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.1: Introduction to One-Dimensional Kinematics by Andrew Park is licensed CC BY 4.0.

1.1.2: Displacement

Learning Objectives

- Define position, displacement, distance, and distance traveled.
- Explain the relationship between position and displacement.
- Distinguish between displacement and distance traveled.
- Calculate displacement and distance given initial position, final position, and the path between the two.



Figure 1.1.2.1: These cyclists in Vietnam can be described by their position relative to buildings and a canal. Their motion can be described by their change in position, or displacement, in the frame of reference. (credit: Suzan Black, Fotopedia)

Position

In order to describe the motion of an object, you must first be able to describe its **position**—where it is at any particular time. More precisely, you need to specify its position relative to a convenient reference frame. Earth is often used as a reference frame, and we often describe the position of an object as it relates to stationary objects in that reference frame. For example, a rocket launch would be described in terms of the position of the rocket with respect to the Earth as a whole, while a professor's position could be described in terms of where she is in relation to the nearby white board. (See Figure 1.1.2.2.) In other cases, we use reference frames that are not stationary but are in motion relative to the Earth. To describe the position of a person in an airplane, for example, we use the airplane, not the Earth, as the reference frame. (See Figure 1.1.2.3.)

Displacement

If an object moves relative to a reference frame (for example, if a professor moves to the right relative to a white board or a passenger moves toward the rear of an airplane), then the object's position changes. This change in position is known as **displacement**. The word "displacement" implies that an object has moved, or has been displaced.

Definition: DISPLACEMENT

Displacement is the *change in position* of an object:

$$\Delta x = x_{\mathrm{f}} - x_0,$$

where Δx is displacement, x_{f} is the final position, and x_0 is the initial position.

In this text the upper case Greek letter Δ always means "change in" whatever quantity follows it; thus, Δx means *change in*

position. Always solve for displacement by subtracting initial position x_0 from final position x_f.

Note that the SI unit for displacement is the meter (m), but sometimes kilometers, miles, feet, and other units of length are used. Keep in mind that when units other than the meter are used in a problem, you may need to convert them into meters to complete the calculation.

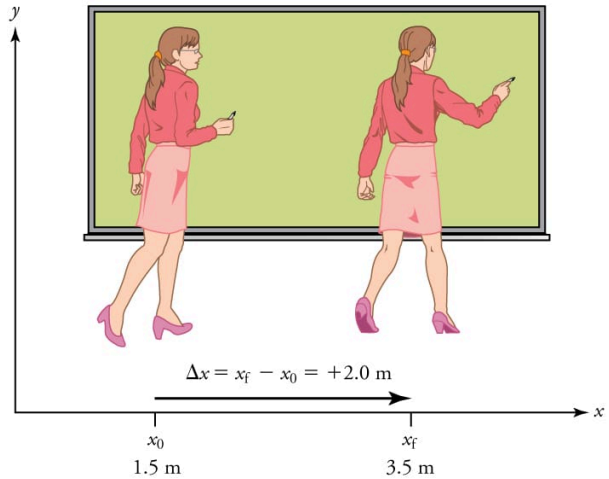


Figure 1.1.2.2: A professor paces left and right while lecturing. Her position relative to Earth is given by x. The **+2.0 m** displacement of the professor relative to Earth is represented by an arrow pointing to the right.

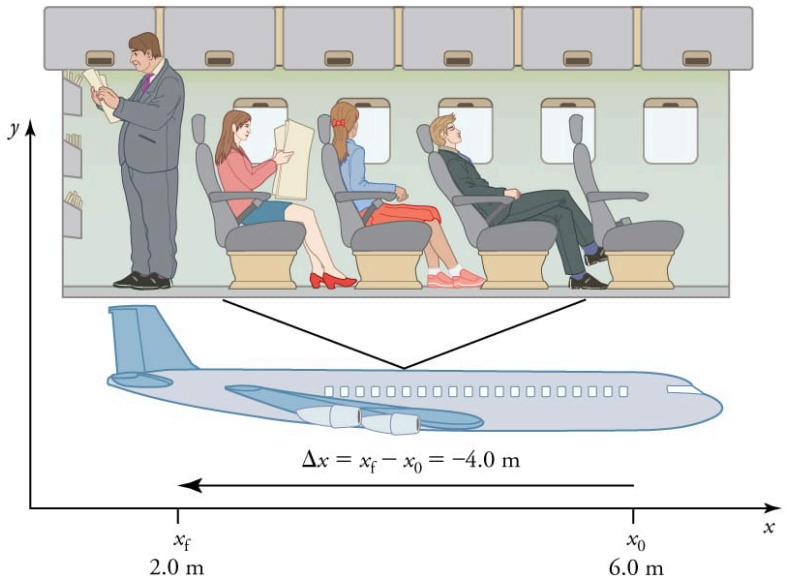


Figure 1.1.2.3: A passenger moves from his seat to the back of the plane. His location relative to the airplane is given by x. The **–4.0-m** displacement of the passenger relative to the plane is represented by an arrow toward the rear of the plane. Notice that the arrow representing his displacement is twice as long as the arrow representing the displacement of the professor (he moves twice as far) in Figure 1.1.2.2.

Note that displacement has a direction as well as a magnitude. The professor's displacement is 2.0 m to the right, and the airline passenger's displacement is 4.0 m toward the rear. In one-dimensional motion, direction can be specified with a plus or minus sign. When you begin a problem, you should select which direction is positive (usually that will be to the right or up, but you are free to select positive as being any direction). The professor's initial position is $x_0 = 1.5$ m and her final position is $x_f = 3.5$ m. Thus her displacement is

$$\Delta x = x_f - x_0 = 3.5 \text{ m} - 1.5 \text{ m} = +2.0 \text{ m}.$$

In this coordinate system, motion to the right is positive, whereas motion to the left is negative. Similarly, the airplane passenger's initial position is $x_0 = 6.0$ m and his final position is $x_f = 2.0$ m, so his displacement is

$$\Delta x = x_f - x_0 = 2.0 \text{ m} - 6.0 \text{ m} = -4.0 \text{ m}.$$

His displacement is negative because his motion is toward the rear of the plane, or in the negative x direction in our coordinate system.

Distance

Although displacement is described in terms of direction, distance is not. **Distance** is defined to be *the magnitude or size of displacement between two positions*. Note that the distance between two positions is not the same as the distance traveled between them. **Distance traveled** is *the total length of the path traveled between two positions*. Distance has no direction and, thus, no sign. For example, the distance the professor walks is 2.0 m. The distance the airplane passenger walks is 4.0 m.

MISCONCEPTION ALERT: DISTANCE TRAVELED VS. MAGNITUDE OF DISPLACEMENT

It is important to note that the *distance traveled*, however, can be greater than the magnitude of the displacement (by magnitude, we mean just the size of the displacement without regard to its direction; that is, just a number with a unit). For example, the professor could pace back and forth many times, perhaps walking a distance of 150 m during a lecture, yet still end up only 2.0 m to the right of her starting point. In this case her displacement would be +2.0 m, the magnitude of her displacement would be 2.0 m, but the distance she traveled would be 150 m. In kinematics we nearly always deal with displacement and magnitude of displacement, and almost never with distance traveled. One way to think about this is to assume you marked the start of the motion and the end of the motion. The displacement is simply the difference in the position of the two marks and is independent of the path taken in traveling between the two marks. The distance traveled, however, is the total length of the path taken between the two marks.

Exercise 1.1.2.1

A cyclist rides 3 km west and then turns around and rides 2 km east. (a) What is her displacement? (b) What distance does she ride? (c) What is the magnitude of her displacement?

Answer available at the end of this section.

Section Summary

- Kinematics is the study of motion without considering its causes. In this chapter, it is limited to motion along a straight line, called one-dimensional motion.
- Displacement is the change in position of an object.
- In symbols, displacement Δx is defined to be

$$\Delta x = x_{\mathrm{f}} - x_0,$$

  where x_0 is the initial position and x_f is the final position. In this text, the Greek letter Δ (delta) always means “change in” whatever quantity follows it. The SI unit for displacement is the meter (m). Displacement has a direction as well as a magnitude.
- When you start a problem, assign which direction will be positive.
- Distance is the magnitude of displacement between two positions.
- Distance traveled is the total length of the path traveled between two positions.

Glossary

kinematics
: the study of motion without considering its causes

position
: the location of an object at a particular time

displacement
: the change in position of an object

distance
: the magnitude of displacement between two positions

distance traveled
: the total length of the path traveled between two positions

Answers

Exercise 1.1.2.1

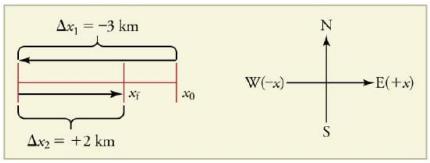


Figure
1.1.2.4

(a) The rider's displacement is $\Delta x = x_f - x_0 = -1$ km. (The displacement is negative because we take east to be positive and west to be negative.)
(b) The distance traveled is **3 km + 2 km = 5 km**.
(c) The magnitude of the displacement is **1 km**.

This page titled 1.1.2: Displacement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.2: Displacement by Andrew Park is licensed CC BY 4.0.

1.1.3: Vectors, Scalars, and Coordinate Systems

Learning Objectives

- Define and distinguish between scalar and vector quantities.
- Assign a coordinate system for a scenario involving one-dimensional motion.



Figure 1.1.3.1: The motion of this Eclipse Concept jet can be described in terms of the distance it has traveled (a scalar quantity) or its displacement in a specific direction (a vector quantity). In order to specify the direction of motion, its displacement must be described based on a coordinate system. In this case, it may be convenient to choose motion toward the left as positive motion (it is the forward direction for the plane), although in many cases, the x-coordinate runs from left to right, with motion to the right as positive and motion to the left as negative. (credit: Armchair Aviator, Flickr)

What is the difference between distance and displacement? Whereas displacement is defined by both direction and magnitude, distance is defined only by magnitude. Displacement is an example of a vector quantity. Distance is an example of a scalar quantity. A **vector** is any quantity with both *magnitude and direction*. Other examples of vectors include a velocity of 90 km/h east and a force of 500 newtons straight down.

The direction of a vector in one-dimensional motion is given simply by a plus (+) or minus (–) sign. Vectors are represented graphically by arrows. An arrow used to represent a vector has a length proportional to the vector's magnitude (e.g., the larger the magnitude, the longer the length of the vector) and points in the same direction as the vector.

Some physical quantities, like distance, either have no direction or none is specified. A **scalar** is any quantity that has a magnitude, but no direction. For example, a 20°C temperature, the 250 kilocalories (250 Calories) of energy in a candy bar, a 90 km/h speed limit, a person's 1.8 m height, and a distance of 2.0 m are all scalars—quantities with no specified direction. Note, however, that a scalar can be negative, such as a –20°C temperature. In this case, the minus sign indicates a point on a scale rather than a direction. Scalars are never represented by arrows.

Coordinate Systems for One-Dimensional Motion

In order to describe the direction of a vector quantity, you must designate a coordinate system within the reference frame. For one-dimensional motion, this is a simple coordinate system consisting of a one-dimensional coordinate line. In general, when describing horizontal motion, motion to the right is usually considered positive, and motion to the left is considered negative. With vertical motion, motion up is usually positive and motion down is negative. In some cases, however, as with the jet in Figure 1.1.3.1, it can be more convenient to switch the positive and negative directions. For example, if you are analyzing the motion of falling objects, it can be useful to define downwards as the positive direction. If people in a race are running to the left, it is useful to define left as the positive direction. It does not matter as long as the system is clear and consistent. Once you assign a positive direction and start solving a problem, you cannot change it.

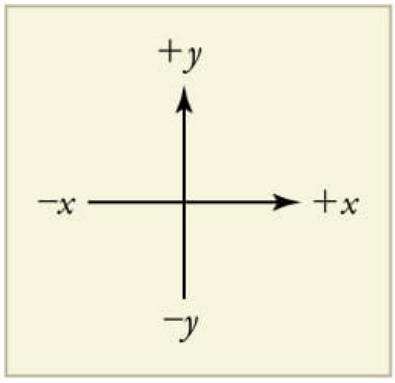


Figure 1.1.3.2: It is usually convenient to consider motion upward or to the right as positive (+) and motion downward or to the left as negative (–).

Exercise 1.1.3.1

A person's speed can stay the same as he or she rounds a corner and changes direction. Given this information, is speed a scalar or a vector quantity? Explain.

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- A vector is any quantity that has magnitude and direction.
- A scalar is any quantity that has magnitude but no direction.
- Displacement and velocity are vectors, whereas distance and speed are scalars.
- In one-dimensional motion, direction is specified by a plus or minus sign to signify left or right, up or down, and the like.

Glossary

scalar
: a quantity that is described by magnitude, but not direction

vector
: a quantity that is described by both magnitude and direction

Answers

Exercise 1.1.3.1

Exercise 1.1.3.1

Speed is a scalar quantity. It does not change at all with direction changes; therefore, it has magnitude only. If it were a vector quantity, it would change as direction changes (even if its magnitude remained constant).

This page titled 1.1.3: Vectors, Scalars, and Coordinate Systems is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.3: Vectors, Scalars, and Coordinate Systems by Andrew Park is licensed CC BY 4.0.

1.1.4: Time, Velocity, and Speed

Learning Objectives

- Explain the relationships between instantaneous velocity, average velocity, instantaneous speed, average speed, displacement, and time.
- Calculate velocity and speed given initial position, initial time, final position, and final time.
- Interpret graphs of position vs. time and velocity vs. time.



Figure 1.1.4.1: The motion of these racing snails can be described by their speeds and their velocities. (credit: tobitasflickr, Flickr)

There is more to motion than distance and displacement. Questions such as, “How long does a foot race take?” and “What was the runner’s speed?” cannot be answered without an understanding of other concepts. In this section we add definitions of time, velocity, and speed to expand our description of motion.

Time

The most fundamental physical quantities are defined by how they are measured. This is the case with time. Every measurement of time involves measuring a change in some physical quantity. It may be a number on a digital clock, a heartbeat, or the position of the Sun in the sky. In physics, the definition of time is simple—**time** is *change*, or the interval over which change occurs. It is impossible to know that time has passed unless something changes.

The amount of time or change is calibrated by comparison with a standard. The SI unit for time is the second, abbreviated s. We might, for example, observe that a certain pendulum makes one full swing every 0.75 s. We could then use the pendulum to measure time by counting its swings or, of course, by connecting the pendulum to a clock mechanism that registers time on a dial. This allows us to not only measure the amount of time, but also to determine a sequence of events.

How does time relate to motion? We are usually interested in elapsed time for a particular motion, such as how long it takes an airplane passenger to get from his seat to the back of the plane. To find elapsed time, we note the time at the beginning and end of the motion and subtract the two. For example, a lecture may start at 11:00 A.M. and end at 11:50 A.M., so that the elapsed time would be 50 min. **Elapsed time Δt** is the difference between the ending time and beginning time,

$$\Delta t = t_{\mathrm{f}} - t_0,$$

where Δt is the change in time or elapsed time, t_{f} is the time at the end of the motion, and t_0 is the time at the beginning of the motion. (As usual, the delta symbol, Δ, means the change in the quantity that follows it.)

Velocity

Your notion of velocity is probably the same as its scientific definition. You know that if you have a large displacement in a small

amount of time you have a large velocity, and that velocity has units of distance divided by time, such as miles per hour or kilometers per hour.

Definition: AVERAGE VELOCITY

Average velocity is *displacement (change in position) divided by the time of travel*,

$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{x_{\mathrm{f}} - x_0}{t_{\mathrm{f}} - t_0},$$

where $\bar{v}$ is the *average* (indicated by the bar over the v) velocity,Δx is the change in position (or displacement), and x_{f} and x_0 are the final and beginning positions at times t_{f} and t_0, respectively.

Notice that this definition indicates that *velocity is a vector because displacement is a vector*. It has both magnitude and direction. The SI unit for velocity is meters per second or m/s, but many other units, such as km/h, mi/h (also written as mph), and cm/s, are in common use. Suppose, for example, an airplane passenger took 5 seconds to move –4 m (the negative sign indicates that displacement is toward the back of the plane). His average velocity would be

$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{-4\ \mathrm{m}}{5\ \mathrm{s}} = -0.8\ \mathrm{m/s}.$$

The minus sign indicates the average velocity is also toward the rear of the plane.

The average velocity of an object does not tell us anything about what happens to it between the starting point and ending point, however. For example, we cannot tell from average velocity whether the airplane passenger stops momentarily or backs up before he goes to the back of the plane. To get more details, we must consider smaller segments of the trip over smaller time intervals.

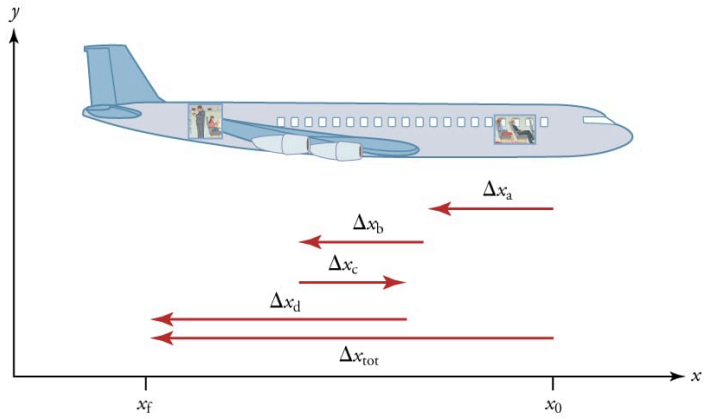


Figure 1.1.4.2: A more detailed record of an airplane passenger heading toward the back of the plane, showing smaller segments of his trip.

The smaller the time intervals considered in a motion, the more detailed the information. When we carry this process to its logical conclusion, we are left with an infinitesimally small interval. Over such an interval, the average velocity becomes the *instantaneous velocity* or the *velocity at a specific instant*. A car's speedometer, for example, shows the magnitude (but not the direction) of the instantaneous velocity of the car. (Police give tickets based on instantaneous velocity, but when calculating how long it will take to get from one place to another on a road trip, you need to use average velocity.) **Instantaneous velocity** v is the average velocity at a specific instant in time (or over an infinitesimally small time interval).

Mathematically, finding instantaneous velocity, v, at a precise instant t can involve taking a limit, a calculus operation beyond the scope of this text. However, under many circumstances, we can find precise values for instantaneous velocity without calculus.

Speed

In everyday language, most people use the terms "speed" and "velocity" interchangeably. In physics, however, they do not have the same meaning and they are distinct concepts. One major difference is that speed has no direction. Thus *speed is a scalar*. Just as we need to distinguish between instantaneous velocity and average velocity, we also need to distinguish between

instantaneous speed and average speed.

Instantaneous speed is the magnitude of instantaneous velocity. For example, suppose the airplane passenger at one instant had an instantaneous velocity of −3.0 m/s (the minus meaning toward the rear of the plane). At that same time his instantaneous speed was 3.0 m/s. Or suppose that at one time during a shopping trip your instantaneous velocity is 40 km/h due north. Your instantaneous speed at that instant would be 40 km/h—the same magnitude but without a direction. Average speed, however, is very different from average velocity. **Average speed** is the distance traveled divided by elapsed time.

We have noted that distance traveled can be greater than displacement. So average speed can be greater than average velocity, which is displacement divided by time. For example, if you drive to a store and return home in half an hour, and your car's odometer shows the total distance traveled was 6 km, then your average speed was 12 km/h. Your average velocity, however, was zero, because your displacement for the round trip is zero. (Displacement is change in position and, thus, is zero for a round trip.) Thus average speed is *not* simply the magnitude of average velocity.

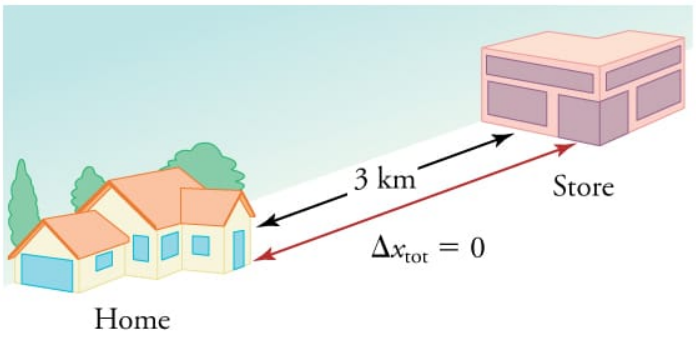


Figure 1.1.4.3: During a 30-minute round trip to the store, the total distance traveled is 6 km. The average speed is 12 km/h. The displacement for the round trip is zero, since there was no net change in position. Thus the average velocity is zero.

Another way of visualizing the motion of an object is to use a graph. A plot of position or of velocity as a function of time can be very useful. For example, for this trip to the store, the position, velocity, and speed-vs.-time graphs are displayed in Figure 1.1.4.4. (Note that these graphs depict a very simplified **model** of the trip. We are assuming that speed is constant during the trip, which is unrealistic given that we'll probably stop at the store. But for simplicity's sake, we will model it with no stops or changes in speed. We are also assuming that the route between the store and the house is a perfectly straight line.)

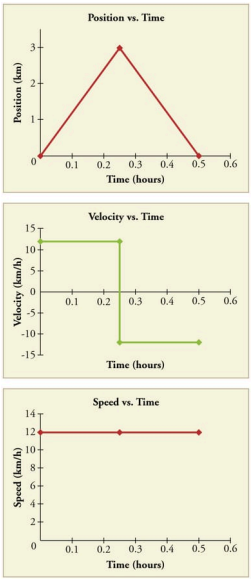


Figure 1.1.4.4: Position vs. time, velocity vs. time, and speed vs. time on a trip. Note that the velocity for the return trip is negative.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—GETTING A SENSE OF SPEED

If you have spent much time driving, you probably have a good sense of speeds between about 10 and 70 miles per hour. But what are these in meters per second? What do we mean when we say that something is moving at 10 m/s? To get a better sense of what these values really mean, do some observations and calculations on your own:

- calculate typical car speeds in meters per second
- estimate jogging and walking speed by timing yourself; convert the measurements into both m/s and mi/h
- determine the speed of an ant, snail, or falling leaf

Exercise 1.1.4.1

A commuter train travels from Baltimore to Washington, DC, and back in 1 hour and 45 minutes. The distance between the two stations is approximately 40 miles. What is (a) the average velocity of the train, and (b) the average speed of the train in m/s?

Answer available at the end of this section.

Section Summary

- Time is measured in terms of change, and its SI unit is the second (s). Elapsed time for an event is

$$\Delta t = t_{\mathrm{f}} - t_0,$$

  where t_{f} is the final time and t_0 is the initial time.
- Average velocity $\bar{v}$ is defined as displacement divided by the travel time. In symbols, average velocity is

$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{x_{\mathrm{f}} - x_0}{t_{\mathrm{f}} - t_0}.$$

- The SI unit for velocity is m/s.
- Velocity is a vector and thus has a direction.
- Instantaneous velocity v is the velocity at a specific instant or the average velocity for an infinitesimal interval.
- Instantaneous speed is the magnitude of the instantaneous velocity.
- Instantaneous speed is a scalar quantity, as it has no direction specified.
- Average speed is the total distance traveled divided by the elapsed time. (Average speed is *not* the magnitude of the average velocity.) Speed is a scalar quantity; it has no direction associated with it.

Glossary

average speed
: distance traveled divided by time during which motion occurs

average velocity
: displacement divided by time over which displacement occurs

instantaneous velocity
: velocity at a specific instant, or the average velocity over an infinitesimal time interval

instantaneous speed
: magnitude of the instantaneous velocity

time
: change, or the interval over which change occurs

model
simplified description that contains only those elements necessary to describe the physics of a physical situation

elapsed time
the difference between the ending time and beginning time

Answers

Exercise 1.1.4.1

(a) The average velocity of the train is zero because $x_f = x_0$; the train ends up at the same place it starts.

(b) The average speed of the train is calculated below. Note that the train travels 40 miles one way and 40 miles back, for a total distance of 80 miles.

$$\frac{\text{distance}}{\text{time}} = \frac{80 \text{ miles}}{105 \text{ minutes}} \tag{1.1.4.1}$$

$$\frac{80 \text{ miles}}{105 \text{ minutes}} \times \frac{5280 \text{ feet}}{1\text{mile}} \times \frac{1 \text{ meter}}{3.28 \text{ feet}} \times \frac{1 \text{ minute}}{60 \text{ seconds}} = 20 \text{ m/s} \tag{1.1.4.2}$$

This page titled 1.1.4: Time, Velocity, and Speed is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.4: Time, Velocity, and Speed by Andrew Park is licensed CC BY 4.0.

1.1.5: Acceleration

Learning Objectives

- Define and distinguish between velocity and acceleration, and between instantaneous and average acceleration.
- Calculate acceleration given initial time, initial velocity, final time, and final velocity.



Figure 1.1.5.1: A plane slows down as it comes in for landing in St. Maarten. It is accelerating in a direction opposite to its velocity. (credit: Steve Conry, Flickr)

In everyday conversation, to accelerate means to speed up. The accelerator in a car can in fact cause it to speed up. The greater the **acceleration**, the greater the change in velocity over a given time. The formal definition of acceleration is consistent with these notions, but more inclusive.

Definition: AVERAGE ACCELERATION

Average Acceleration is *the rate at which velocity changes*,

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{v_{\mathrm{f}} - v_0}{t_{\mathrm{f}} - t_0},$$

where $\bar{a}$ is average acceleration, v is velocity, and t is time. (The bar over the a means *average* acceleration.)

Because acceleration is velocity in m/s divided by time in s, the SI units for acceleration are $\mathrm{m/s^2}$, meters per second squared or meters per second per second, which literally means by how many meters per second the velocity changes every second.

Recall that velocity is a vector—it has both magnitude and direction. This means that a change in velocity can be a change in magnitude (or speed), but it can also be a change in *direction*. For example, if a car turns a corner at constant speed, it is accelerating because its direction is changing. The quicker you turn, the greater the acceleration. So there is an acceleration when velocity changes either in magnitude (an increase or decrease in speed) or in direction, or both.

ACCELERATION AS A VECTOR

Acceleration is a vector in the same direction as the *change* in velocity, Δv. Since velocity is a vector, it can change either in magnitude or in direction. Acceleration is therefore a change in either speed or direction, or both.

Keep in mind that although acceleration is in the direction of the *change* in velocity, it is not always in the direction of *motion*. If acceleration is in a direction opposite to the direction of motion, the object slows down.



Figure 1.1.5.2: A subway train in Sao Paulo, Brazil, slows down as it comes into a station. It is accelerating in a direction opposite to its direction of motion. (credit: Yusuke Kawasaki, Flickr)

Example 1.1.5.1: Calculating Acceleration: A Racehorse Leaves the Gate

A racehorse coming out of the gate accelerates from rest to a velocity of 15.0 m/s due west in 1.80 s. What is its average acceleration?

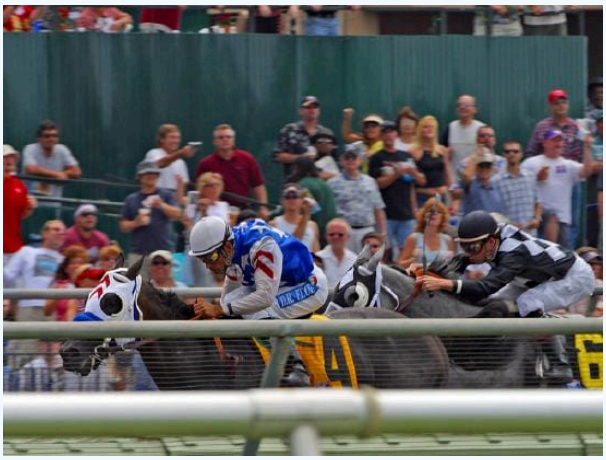


Figure
1.1.5.3

Strategy

First we draw a sketch and assign a coordinate system to the problem. This is a simple problem, but it always helps to visualize it. Notice that we assign east as positive and west as negative. Thus, in this case, we have negative velocity.

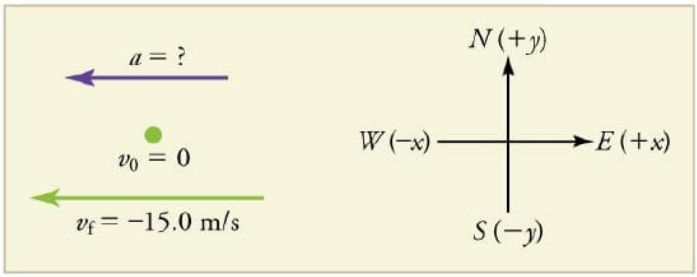


Figure
1.1.5.4

We can solve this problem by identifying Δv and Δt from the given information and then calculating the average acceleration directly from the equation $\bar{a} = \frac{\Delta v}{\Delta t} = \frac{v_f - v_0}{t_f - t_0}$.

Solution

1. Identify the knowns. $v_0 = 0$, $v_f = -15.0\ \text{m/s}$ (the negative sign indicates direction toward the west), $\Delta t = 1.80\ \text{s}$.

2. Find the change in velocity. Since the horse is going from zero to **–15.0 m/s**, its change in velocity equals its final velocity: $\Delta v = v_f = -15.0\ \text{m/s}$.

3. Plug in the known values (Δv and Δt) and solve for the unknown $\bar{a}$.

$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{-15.0\ \text{m/s}}{1.80\ \text{s}} = -8.33\ \text{m/s}^2$.

Discussion

The negative sign for acceleration indicates that acceleration is toward the west. An acceleration of $8.33\ \text{m/s}^2$ due west means that the horse increases its velocity by 8.33 m/s due west each second, that is, 8.33 meters per second per second, which we write as $8.33\ \text{m/s}^2$. This is truly an average acceleration, because the ride is not smooth. We shall see later that an acceleration of this magnitude would require the rider to hang on with a force nearly equal to his weight.

Instantaneous Acceleration

Instantaneous acceleration a, or the *acceleration at a specific instant in time*, is obtained by the same process as discussed for instantaneous velocity—that is, by considering an infinitesimally small interval of time. How do we find instantaneous acceleration using only algebra? The answer is that we choose an average acceleration that is representative of the motion. Figure 1.1.5.5 shows graphs of instantaneous acceleration versus time for two very different motions. In Figure 1.1.5.5 (a), the acceleration varies slightly and the average over the entire interval is nearly the same as the instantaneous acceleration at any time. In this case, we should treat this motion as if it had a constant acceleration equal to the average (in this case about $1.8\ \text{m/s}^2$). In Figure 1.1.5.5 (b), the acceleration varies drastically over time. In such situations it is best to consider smaller time intervals and choose an average acceleration for each. For example, we could consider motion over the time intervals from 0 to 1.0 s and from 1.0 to 3.0 s as separate motions with accelerations of $+3.0\ \text{m/s}^2$ and $-2.0\ \text{m/s}^2$, respectively.

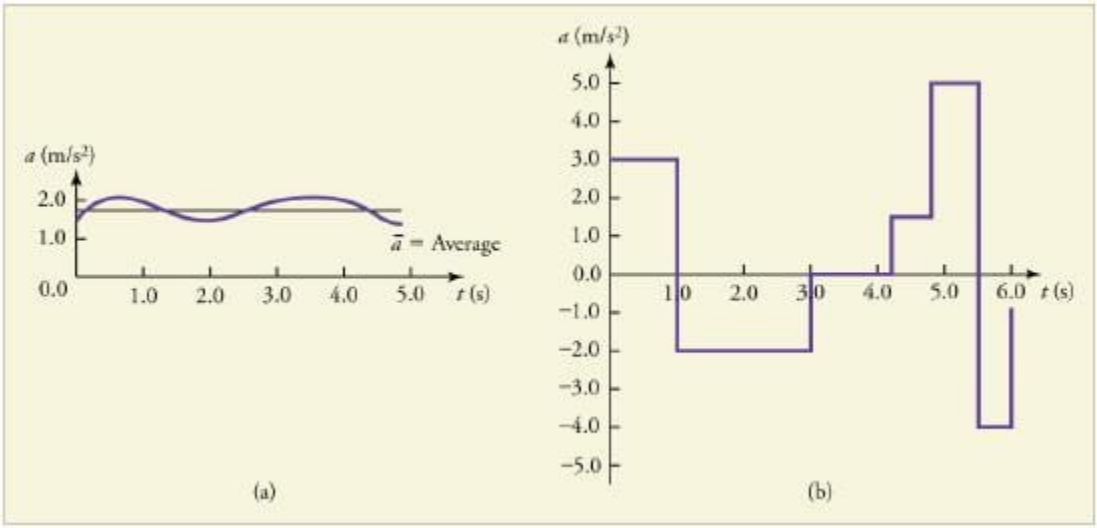


Figure 1.1.5.5: Graphs of instantaneous acceleration versus time for two different one-dimensional motions. (a) Here acceleration varies only slightly and is always in the same direction, since it is positive. The average over the interval is nearly the same as the acceleration at any given time. (b) Here the acceleration varies greatly, perhaps representing a package on a post office conveyor belt that is accelerated forward and backward as it bumps along. It is necessary to consider small time intervals (such as from 0 to 1.0 s) with constant or nearly constant acceleration in such a situation.

The next several examples consider the motion of the subway train shown in Figure 1.1.5.6. In (a) the shuttle moves to the right,

and in (b) it moves to the left. The examples are designed to further illustrate aspects of motion and to illustrate some of the reasoning that goes into solving problems.

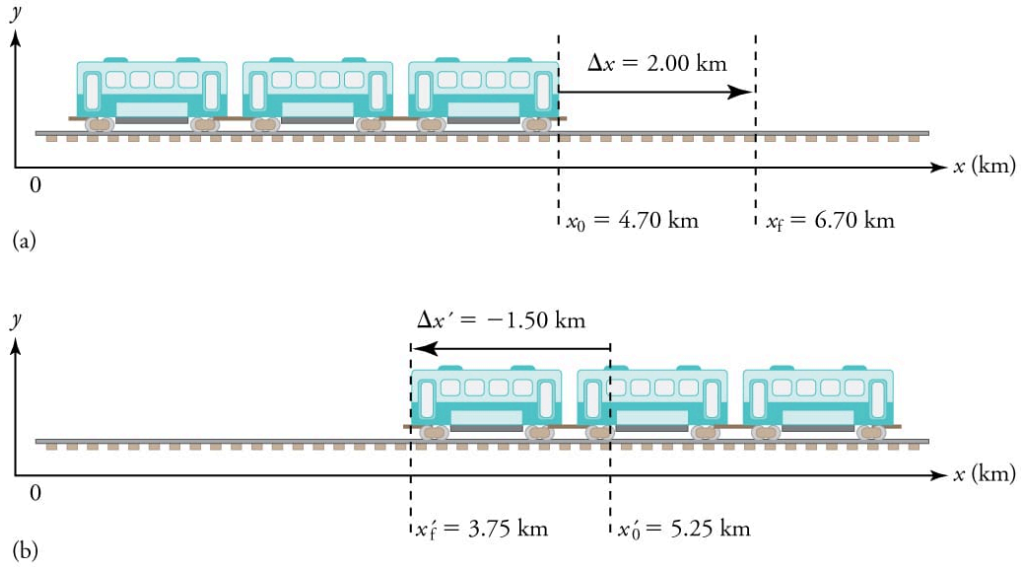


Figure 1.1.5.6: One-dimensional motion of a subway train considered in Example 1.1.5.2, Example 1.1.5.3, Example 1.1.5.4, Example 1.1.5.5 , Example 1.1.5.6, and Example 1.1.5.7. Here we have chosen the x-axis so that + means to the right and – means to the left for displacements, velocities, and accelerations. (a) The subway train moves to the right from x_0 to x_f. Its displacement Δx is +2.00 km. (b) The train moves to the left from x_0' to x_f'. Its displacement $\Delta x'$ is –1.50 km. (Note that the prime symbol (') is used simply to distinguish between displacement in the two different situations. The distances of travel and the size of the cars are on different scales to fit everything into the diagram.)

Example 1.1.5.2: Calculating Displacement: A Subway Train

What are the magnitude and sign of displacements for the motions of the subway train shown in parts (a) and (b) of Figure 1.1.5.6?

Strategy

A drawing with a coordinate system is already provided, so we don't need to make a sketch, but we should analyze it to make sure we understand what it is showing. Pay particular attention to the coordinate system. To find displacement, we use the equation $\Delta x = x_f - x_0$. This is straightforward since the initial and final positions are given.

Solution

1. Identify the knowns. In the figure we see that $x_f = 6.70$ km and $x_0 = 4.70$ km for part (a), and $x_f' = 3.75$ km and $x_0' = 5.25$ km for part (b).

2. Solve for displacement in part (a).

$$\Delta x = x_f - x_0 = 6.70 \text{ km} - 4.70 \text{ km} = +2.00 \text{ km}$$

3. Solve for displacement in part (b).

$$\Delta x' = x_f' - x_0' = 3.75 \text{ km} - 5.25 \text{ km} = -1.50 \text{ km}$$

Discussion

The direction of the motion in (a) is to the right and therefore its displacement has a positive sign, whereas motion in (b) is to the left and thus has a negative sign.

Example 1.1.5.3: Comparing Distance Traveled with Displacement: A Subway Train

What are the distances traveled for the motions shown in parts (a) and (b) of the subway train in Figure 1.1.5.6?

Strategy

To answer this question, think about the definitions of distance and distance traveled, and how they are related to displacement. Distance between two positions is defined to be the magnitude of displacement, which was found in Example 1.1.5.2. Distance traveled is the total length of the path traveled between the two positions. In the case of the subway train shown in Figure 1.1.5.6, the distance traveled is the same as the distance between the initial and final positions of the train.

Solution

1. The displacement for part (a) was +2.00 km. Therefore, the distance between the initial and final positions was 2.00 km, and the distance traveled was 2.00 km.

2. The displacement for part (b) was **–1.5 km**. Therefore, the distance between the initial and final positions was 1.50 km, and the distance traveled was 1.50 km.

Discussion

Distance is a scalar. It has magnitude but no sign to indicate direction.

Example 1.1.5.4: Calculating Acceleration: A Subway Train Speeding Up

Suppose the train in Figure 1.1.5.6 (a) accelerates from rest to 30.0 km/h in the first 20.0 s of its motion. What is its average acceleration during that time interval?

Strategy

It is worth it at this point to make a simple sketch:

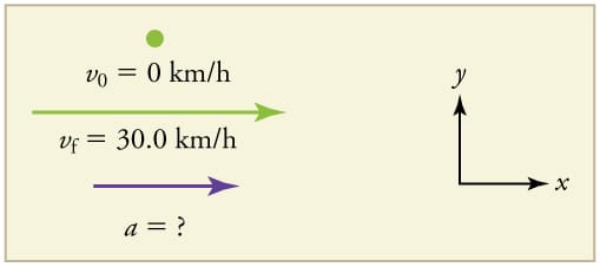


Figure
1.1.5.7

This problem involves three steps. First we must determine the change in velocity, then we must determine the change in time, and finally we use these values to calculate the acceleration.

Solution

1. Identify the knowns. $v_0 = 0$ (the trains starts at rest), $v_f = 30.0$ km/h, and $\Delta t = 20.0$ s.

2. Calculate Δv. Since the train starts from rest, its change in velocity is $\Delta v = +30.0$ km/h, where the plus sign means velocity to the right.

3. Plug in known values and solve for the unknown, $\bar{a}$.

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{+30.0 \text{ km/h}}{20.0 \text{ s}}$$

4. Since the units are mixed (we have both hours and seconds for time), we need to convert everything into SI units of meters and seconds.

$$\bar{a} = \left(\frac{+30 \text{ km/h}}{20.0 \text{ s}}\right)\left(\frac{10^3 \text{ m}}{1 \text{ km}}\right)\left(\frac{1 \text{ h}}{3600 \text{ s}}\right) = 0.417 \text{ m/s}^2$$

Discussion

The plus sign means that acceleration is to the right. This is reasonable because the train starts from rest and ends up with a velocity to the right (also positive). So acceleration is in the same direction as the *change* in velocity, as is always the case.

Example 1.1.5.5: Calculate Acceleration: A Subway Train Slowing Down

Now suppose that at the end of its trip, the train in Figure 1.1.5.6 (a) slows to a stop from a speed of 30.0 km/h in 8.00 s. What is its average acceleration while stopping?

Strategy

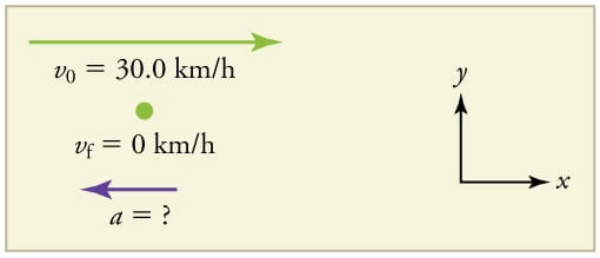


Figure
1.1.5.8

In this case, the train is slowing down and its acceleration is negative because it is toward the left. As in the previous example, we must find the change in velocity and the change in time and then solve for acceleration.

Solution

1. Identify the knowns. $v_0 = 30.0$ km/h, $v_f = 0$ km/h (the train is stopped, so its velocity is 0), and $\Delta t = 8.00$ s.

2. Solve for the change in velocity, Δv.

$$\Delta v = v_f - v_0 = 0 - 30.0 \text{ km/h} = -30.0 \text{ km/h}$$

3. Plug in the knowns, Δv and Δt, and solve for $\bar{a}$.

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{-30.0 \text{ km/h}}{8.00 \text{ s}}$$

4. Convert the units to meters and seconds.

$$\bar{a} = \frac{\Delta v}{\Delta t} = \left(\frac{-30.0 \text{ km/h}}{8.00 \text{ s}}\right)\left(\frac{10^3 \text{ m}}{1 \text{ km}}\right)\left(\frac{1 \text{ h}}{3600 \text{ s}}\right) = -1.04 \text{ m/s}^2.$$

Discussion

The minus sign indicates that acceleration is to the left. This sign is reasonable because the train initially has a positive velocity in this problem, and a negative acceleration would oppose the motion.

The graphs of position, velocity, and acceleration vs. time for the trains in Example 1.1.5.4 and Example 1.1.5.5 are displayed in Figure 1.1.5.9. (We have taken the velocity to remain constant from 20 to 40 s, after which the train slows down.)

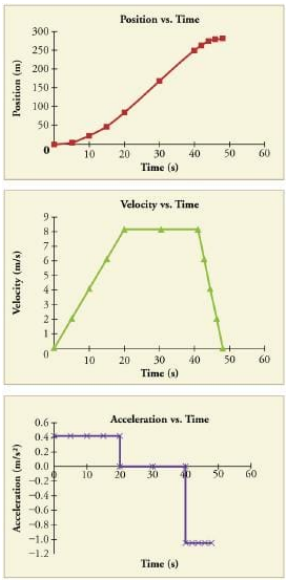


Figure 1.1.5.9: (a) Position of the train over time. Notice that the train's position changes slowly at the beginning of the journey, then more and more quickly as it picks up speed. Its position then changes more slowly as it slows down at the end of the journey. In the middle of the journey, while the velocity remains constant, the position changes at a constant rate. (b) Velocity of the train over time. The train's velocity increases as it accelerates at the beginning of the journey. It remains the same in the middle of the journey (where there is no acceleration). It decreases as the train slows down at the end of the journey. (c) The acceleration of the train over time. The train has positive acceleration as it speeds up at the beginning of the journey. It has no acceleration as it travels at constant velocity in the middle of the journey. Its acceleration is negative as it slows down at the end of the journey.

Example 1.1.5.6: Calculating Average Velocity: The Subway Train

What is the average velocity of the train in part b of Example 1.1.5.2, and shown again below, if it takes 5.00 min to make its trip?

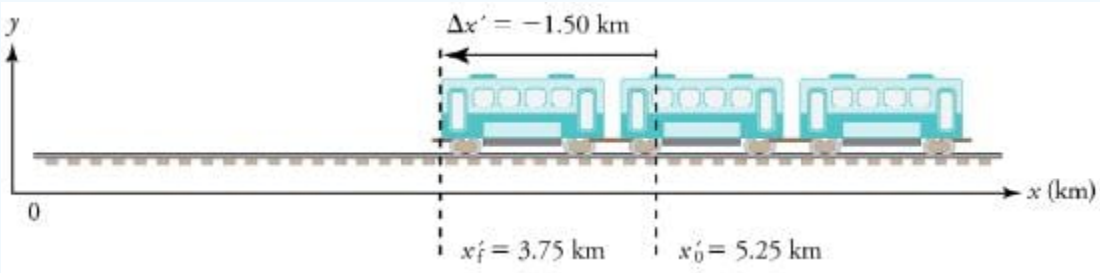


Figure
1.1.5.10

Strategy

Average velocity is displacement divided by time. It will be negative here, since the train moves to the left and has a negative displacement.

Solution

1. Identify the knowns. $x'_{\mathrm{f}} = 3.75\ \mathrm{km},\ x'_0 = 5.25\ \mathrm{km},\ \Delta t = 5.00\ \mathrm{min}.$

2. Determine displacement, $\Delta x'$. We found $\Delta x'$ to be **–1.5 km** in Example 1.1.5.2.

3. Solve for average velocity.

$$\bar{v} = \frac{\Delta x'}{\Delta t} = \frac{-1.50\ \mathrm{km}}{5.00\ \mathrm{min}}$$

4. Convert units.

$$\bar{v} = \frac{\Delta x'}{\Delta t} = \left(\frac{-1.50\ \mathrm{km}}{5.00\ \mathrm{min}}\right)\left(\frac{60\ \mathrm{min}}{1\ \mathrm{h}}\right) = -18.0\ \mathrm{km/h}$$

Discussion

The negative velocity indicates motion to the left.

Example 1.1.5.7: Calculating Deceleration: The Subway Train

Finally, suppose the train in Figure 1.1.5.10 slows to a stop from a velocity of 20.0 km/h in 10.0 s. What is its average acceleration?

Strategy

Once again, let's draw a sketch:

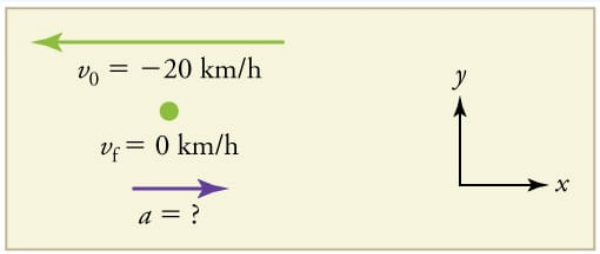


Figure
1.1.5.11

As before, we must find the change in velocity and the change in time to calculate average acceleration.

Solution

1. Identify the knowns. $v_0 = -20\ \mathrm{km/h},\ v_{\mathrm{f}} = 0\ \mathrm{km/h},\ \Delta t = 10.0\ \mathrm{s}.$

2. Calculate Δv. The change in velocity here is actually positive, since

$$\Delta v = v_{\mathrm{f}} - v_0 = 0 - (-20\ \mathrm{km/h}) = +20\ \mathrm{km/h}$$

3. Solve for $\bar{a}$.

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{+20.0\ \mathrm{km/h}}{10.0\ \mathrm{s}}$$

4. Convert units.

$$\bar{a} = \left(\frac{+20.0 \text{ km/h}}{10.0 \text{ s}}\right)\left(\frac{10^3 \text{ m}}{1 \text{ km}}\right)\left(\frac{1 \text{ h}}{3600 \text{ s}}\right) = +0.556 \text{ m/s}^2$$

Discussion

The plus sign means that acceleration is to the right. This is reasonable because the train initially has a negative velocity (to the left) in this problem and a positive acceleration opposes the motion (and so it is to the right).

Sign and Direction

Perhaps the most important thing to note about these examples is the signs of the answers. In our chosen coordinate system, plus means the quantity is to the right and minus means it is to the left. This is easy to imagine for displacement and velocity. But it is a little less obvious for acceleration. Most people interpret negative acceleration as the slowing of an object. This was not the case in Example 1.1.5.7, where a positive acceleration slowed a negative velocity. The crucial distinction was that the acceleration was in the opposite direction from the velocity. In fact, a negative acceleration will *increase* a negative velocity. For example, the train moving to the left in Figure 1.1.5.10 is sped up by an acceleration to the left. In that case, both v and a are negative. The plus and minus signs give the directions of the accelerations. If acceleration has the same sign as the change in velocity, the object is speeding up. If acceleration has the opposite sign of the change in velocity, the object is slowing down.

Exercise 1.1.5.1

An airplane lands on a runway traveling east. Describe its acceleration.

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- Acceleration is the rate at which velocity changes. In symbols, **average acceleration** $\bar{a}$ is

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{v_\text{f} - v_0}{t_\text{f} - t_0}$$

- The SI unit for acceleration is m/s^2.

- Acceleration is a vector, and thus has a both a magnitude and direction.
- Acceleration can be caused by either a change in the magnitude or the direction of the velocity.
- Instantaneous acceleration a is the acceleration at a specific instant in time.
- When an acceleration is in a direction opposite to that of the velocity of an object, the object slows down.

Glossary

acceleration
: the rate of change in velocity; the change in velocity over time

average acceleration
: the change in velocity divided by the time over which it changes

instantaneous acceleration
: acceleration at a specific point in time

Answers

Exercise 1.1.5.1

Exercise 1.1.5.1

If we take east to be positive, then the airplane has negative acceleration, as it is accelerating toward the west. It is also slowing down: its acceleration is opposite in direction to its velocity.

This page titled 1.1.5: Acceleration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.5: Acceleration by Andrew Park is licensed CC BY 4.0.

1.1.6: Motion Equations for Constant Acceleration in One Dimension

Learning Objectives

- Calculate displacement of an object that is not accelerating, given initial position and velocity.
- Calculate final velocity of an accelerating object, given initial velocity, acceleration, and time.
- Calculate displacement and final position of an accelerating object, given initial position, initial velocity, time, and acceleration.



Figure 1.1.6.1: Kinematic equations can help us describe and predict the motion of moving objects such as these kayaks racing in Newbury, England. (credit: Barry Skeates, Flickr)

We might know that the greater the acceleration of, say, a car moving away from a stop sign, the greater the displacement in a given time. But we have not developed a specific equation that relates acceleration and displacement. In this section, we develop some convenient equations for kinematic relationships, starting from the definitions of displacement, velocity, and acceleration already covered.

Notation: *t*, *x*, *v*, *a*

First, let us make some simplifications in notation. Taking the initial time to be zero, as if time is measured with a stopwatch, is a great simplification. Since elapsed time is $\Delta t = t_{\mathrm{f}} - t_0$, taking $t_0 = 0$ means that $\Delta t = t_{\mathrm{f}}$, the final time on the stopwatch. When initial time is taken to be zero, we use the subscript 0 to denote initial values of position and velocity. That is, x_0 *is the initial position* and v_0 *is the initial velocity*. We put no subscripts on the final values. That is, t *is the final time*, x *is the final position*, and v *is the final velocity*. This gives a simpler expression for elapsed time—now, $\Delta t = t$. It also simplifies the expression for displacement, which is now $\Delta x = x - x_0$. Also, it simplifies the expression for change in velocity, which is now $\Delta v = v - v_0$. To summarize, using the simplified notation, with the initial time taken to be zero,

$$\left.\begin{array}{l} \Delta t = t \\ \Delta x = x - x_0 \\ \Delta v = v - v_0 \end{array}\right\}$$

where *the subscript 0 denotes an initial value and the absence of a subscript denotes a final value* in whatever motion is under consideration.

We now make the important assumption that *acceleration is constant*. This assumption allows us to avoid using calculus to find instantaneous acceleration. Since acceleration is constant, the average and instantaneous accelerations are equal. That is,

$$\bar{a} = a = \text{ constant}$$

so we use the symbol a for acceleration at all times. Assuming acceleration to be constant does not seriously limit the situations we can study nor degrade the accuracy of our treatment. For one thing, acceleration *is* constant in a great number of situations. Furthermore, in many other situations we can accurately describe motion by assuming a constant acceleration equal to the average acceleration for that motion. Finally, in motions where acceleration changes drastically, such as a car accelerating to top speed and then braking to a stop, the motion can be considered in separate parts, each of which has its own constant acceleration.

SOLVING FOR DISPLACEMENT (Δx) AND FINAL POSITION (x) FROM AVERAGE VELOCITY WHEN ACCELERATION (a) IS CONSTANT

To get our first two new equations, we start with the definition of average velocity:

$$\bar{v} = \frac{\Delta x}{\Delta t}.$$

Substituting the simplified notation for Δx and Δt yields

$$\bar{v} = \frac{x - x_0}{t}.$$

Solving for x yields

$$x = x_0 + \bar{v}t,$$

where the average velocity is

$$\bar{v} = \frac{v_0 + v}{2} (\text{constant } a).$$

The equation $\bar{v} = \frac{v_0+v}{2}$ reflects the fact that, when acceleration is constant, v is just the simple average of the initial and final velocities. For example, if you steadily increase your velocity (that is, with constant acceleration) from 30 to 60 km/h, then your average velocity during this steady increase is 45 km/h. Using the equation $\bar{v} = \frac{v_0+v}{2}$ to check this, we see that

$$\bar{v} = \frac{v_0 + v}{2} = \frac{30 \text{ km/h} + 60 \text{ km/h}}{2} = 45 \text{ km/h},$$

which seems reasonable.

Example 1.1.6.1: Calculating Displacement: How Far does the Jogger Run?

A jogger runs down a straight stretch of road with an average velocity of 4.00 m/s for 2.00 min. What is his final position, taking his initial position to be zero?

Strategy

Draw a sketch.

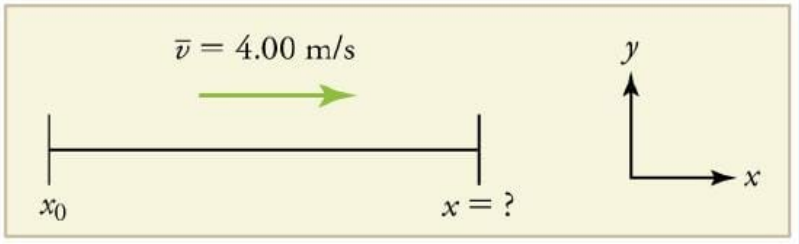


Figure
1.1.6.2

The final position x is given by the equation

$$x = x_0 + \bar{v}t.$$

To find x, we identify the values of x_0, $\bar{v}$, and t from the statement of the problem and substitute them into the equation.

Solution

1. Identify the knowns. $\bar{v} = 4.00\ \text{m/s},\ \Delta t = 2.00\ \text{min}$, and $x_0 = 0\ \text{m}$.

2. Enter the known values into the equation.

$$x = x_0 + \bar{v}t = 0 + (4.00\ \text{m/s})(120\ \text{s}) = 480\ \text{m}$$

Discussion

Velocity and final displacement are both positive, which means they are in the same direction.

The equation $x = x_0 + \bar{v}t$ gives insight into the relationship between displacement, average velocity, and time. It shows, for example, that displacement is a linear function of average velocity. (By linear function, we mean that displacement depends on $\bar{v}$ rather than on $\bar{v}$ raised to some other power, such as $\bar{v}^2$. When graphed, linear functions look like straight lines with a constant slope.) On a car trip, for example, we will get twice as far in a given time if we average 90 km/h than if we average 45 km/h.

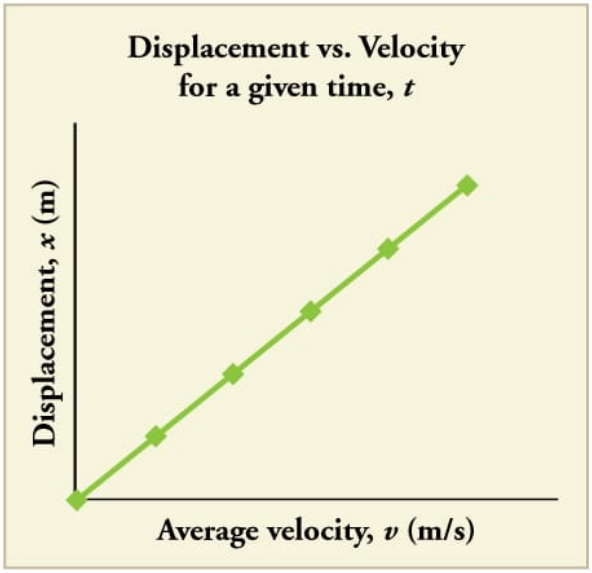


Figure 1.1.6.3: There is a linear relationship between displacement and average velocity. For a given time t, an object moving twice as fast as another object will move twice as far as the other object.

SOLVING FOR FINAL VELOCITY

We can derive another useful equation by manipulating the definition of acceleration.

$$a = \frac{\Delta v}{\Delta t}$$

Substituting the simplified notation for Δv and Δt gives us

$$a = \frac{v - v_0}{t} (\text{ constant } a).$$

Solving for v yields

$$v = v_0 + at(\text{ constant } a).$$

Example 1.1.6.2: Calculating Final Velocity: An Airplane Slowing Down after Landing

An airplane lands with an initial velocity of 70.0 m/s and then slows down at $1.50\ \text{m/s}^2$ for 40.0 s. What is its final velocity?

Strategy

Draw a sketch. We draw the acceleration vector in the direction opposite the velocity vector because the plane is slowing down.

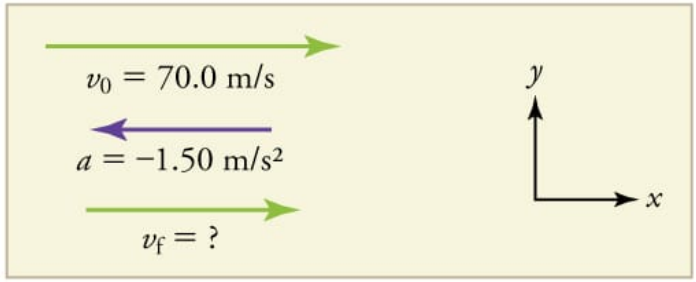


Figure
1.1.6.4

Solution

1. Identify the knowns. $v_0 = 70.0\ \text{m/s},\ a = -1.50\ \text{m/s}^2,\ t = 40.0\ \text{s}$.

2. Identify the unknown. In this case, it is final velocity, v_f.

3. Determine which equation to use. We can calculate the final velocity using the equation $v = v_0 + at$.

4. Plug in the known values and solve.

$$v = v_0 + at = 70.0\ \text{m/s} + \left(-1.50\ \text{m/s}^2\right)(40.0\ \text{s}) = 10.0\ \text{m/s}$$

Discussion

The final velocity is much less than the initial velocity, as desired when slowing down, but still positive. With jet engines, reverse thrust could be maintained long enough to stop the plane and start moving it backward. That would be indicated by a negative final velocity, which is not the case here.



Figure 1.1.6.5: The airplane lands with an initial velocity of 70.0 m/s and slows to a final velocity of 10.0 m/s before heading for the terminal. Note that the acceleration is negative because its direction is opposite to its velocity, which is positive.

In addition to being useful in problem solving, the equation $v = v_0 + at$ gives us insight into the relationships among velocity, acceleration, and time. From it we can see, for example, that

- final velocity depends on how large the acceleration is and how long it lasts

- if the acceleration is zero, then the final velocity equals the initial velocity $(v = v_0)$, as expected (i.e., velocity is constant)
- if a is negative, then the final velocity is less than the initial velocity

All of these observations fit our intuition, and it is always useful to examine basic equations in light of our intuition and experiences to check that they do indeed describe nature accurately.

MAKING CONNECTIONS: REAL-WORLD CONNECTION



Figure 1.1.6.6: The Space Shuttle *Endeavor* blasts off from the Kennedy Space Center in February 2010. (credit: Matthew Simantov, Flickr)

An intercontinental ballistic missile (ICBM) has a larger average acceleration than the Space Shuttle and achieves a greater velocity in the first minute or two of flight (actual ICBM burn times are classified—short-burn-time missiles are more difficult for an enemy to destroy). But the Space Shuttle obtains a greater final velocity, so that it can orbit the earth rather than come directly back down as an ICBM does. The Space Shuttle does this by accelerating for a longer time.

SOLVING FOR FINAL POSITION WHEN VELOCITY IS NOT CONSTANT (a≠0)

We can combine the equations above to find a third equation that allows us to calculate the final position of an object experiencing constant acceleration. We start with

$$v = v_0 + at.$$

Adding v_0 to each side of this equation and dividing by 2 gives

$$\frac{v_0 + v}{2} = v_0 + \frac{1}{2}at.$$

Since $\frac{v_0+v}{2} = \bar{v}$ for constant acceleration, then

$$\bar{v} = v_0 + \frac{1}{2}at.$$

Now we substitute this expression for $\bar{v}$ into the equation for displacement, $x = x_0 + \bar{v}t$, yielding

$$x = x_0 + v_0t + \frac{1}{2}at^2 (\text{ constant } a).$$

Example 1.1.6.3: Calculating Displacement of an Accelerating Object: Dragsters

Dragsters can achieve average accelerations of $26.0\ \text{m/s}^2$. Suppose such a dragster accelerates from rest at this rate for 5.56 s. How far does it travel in this time?



Figure 1.1.6.7: U.S. Army Top Fuel pilot Tony “The Sarge” Schumacher begins a race with a controlled burnout. (credit: Lt. Col. William Thurmond. Photo Courtesy of U.S. Army.)

Strategy

Draw a sketch.

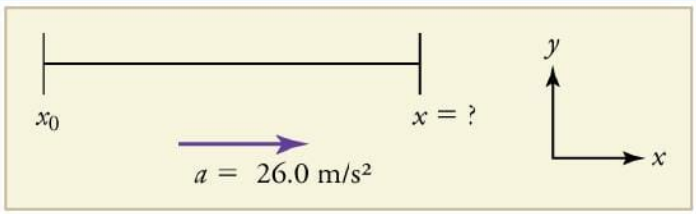


Figure
1.1.6.8

We are asked to find displacement, which is x if we take x_0 to be zero. (Think about it like the starting line of a race. It can be anywhere, but we call it 0 and measure all other positions relative to it.) We can use the equation $x = x_0 + v_0 t + \frac{1}{2}at^2$ once we identify v_0, a, and t from the statement of the problem.

Solution

1. Identify the knowns. Starting from rest means that $v_0 = 0$, a is given as $26.0\ \text{m/s}^2$ and t is given as 5.56 s.

2. Plug the known values into the equation to solve for the unknown x:

$$x = x_0 + v_0 t + \frac{1}{2}at^2.$$

Since the initial position and velocity are both zero, this simplifies to

$$x = \frac{1}{2}at^2.$$

Substituting the identified values of a and t gives

$$x = \frac{1}{2}\left(26.0\ \text{m/s}^2\right)(5.56\ \text{s})^2,$$

yielding

$$x = 402\ \text{m}.$$

Discussion

If we convert 402 m to miles, we find that the distance covered is very close to one quarter of a mile, the standard distance for drag racing. So the answer is reasonable. This is an impressive displacement in only 5.56 s, but top-notch dragsters can do a quarter mile in even less time than this.

What else can we learn by examining the equation $x = x_0 + v_0 t + \frac{1}{2}at^2$ We see that:

- displacement depends on the square of the elapsed time when acceleration is not zero. In Example $1.1.6.3$, the dragster covers only one fourth of the total distance in the first half of the elapsed time
- if acceleration is zero, then the initial velocity equals average velocity $(v_0 = \bar{v})$ and $x = x_0 + v_0 t + \frac{1}{2}at^2$ becomes $x = x_0 + v_0 t$

SOLVING FOR FINAL VELOCITY WHEN VELOCITY IS NOT CONSTANT (a≠0)

A fourth useful equation can be obtained from another algebraic manipulation of previous equations.

If we solve $v = v_0 + at$ for t, we get

$$t = \frac{v - v_0}{a}.$$

Substituting this and $\bar{v} = \frac{v_0 + v}{2}$ into $x = x_0 + \bar{v}t$, we get

$$v^2 = v_0^2 + 2a\,(x - x_0)\ (\text{ constant } a).$$

Example $1.1.6.4$: Calculating Final Velocity: Dragsters

Calculate the final velocity of the dragster in Example $1.1.6.3$ without using information about time.

Strategy

Draw a sketch.

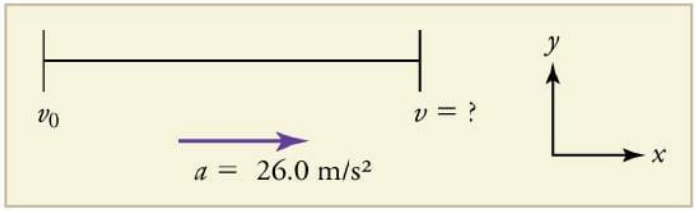


Figure
1.1.6.9

The equation $v^2 = v_0^2 + 2a\,(x - x_0)$ is ideally suited to this task because it relates velocities, acceleration, and displacement, and no time information is required.

Solution

1. Identify the known values. We know that $v_0 = 0$, since the dragster starts from rest. Then we note that $x - x_0 = 402\ \text{m}$ (this was the answer in Example $1.1.6.3$). Finally, the average acceleration was given to be $a = 26.0\ \text{m/s}^2$.

2. Plug the knowns into the equation $v^2 = v_0^2 + 2a\,(x - x_0)$ and solve for v.

$$v^2 = 0 + 2\left(26.0\ \text{m/s}^2\right)(402\ \text{m}).$$

Thus

$$v^2 = 2.09 \times 10^4 \text{ m}^2/\text{s}^2.$$

To get v, we take the square root:

$$v = \sqrt{2.09 \times 10^4 \text{ m}^2/\text{s}^2} = 145 \text{ m/s}.$$

Discussion

145 m/s is about 522 km/h or about 324 mi/h, but even this breakneck speed is short of the record for the quarter mile. Also, note that a square root has two values; we took the positive value to indicate a velocity in the same direction as the acceleration.

An examination of the equation $v^2 = v_0^2 + 2a(x - x_0)$ can produce further insights into the general relationships among physical quantities:

- The final velocity depends on how large the acceleration is and the distance over which it acts
- For a fixed magnitude of acceleration, a car that is going twice as fast doesn't simply stop in twice the distance—it takes much further to stop. (This is why we have reduced speed zones near schools.)

Putting Equations Together

In the following examples, we further explore one-dimensional motion, but in situations requiring slightly more algebraic manipulation. The examples also give insight into problem-solving techniques. The box below provides easy reference to the equations needed.

SUMMARY OF KINEMATIC EQUATIONS (CONSTANT a)

$$x = x_0 + \bar{v}t$$
$$\bar{v} = \frac{v_0 + v}{2}$$
$$v = v_0 + at$$
$$x = x_0 + v_0 t + \frac{1}{2}at^2$$
$$v^2 = v_0^2 + 2a(x - x_0)$$

Example 1.1.6.5: Calculating Displacement: How Far Does a Car Go When Coming to a Halt?

On dry concrete, suppose a car can slow down at a rate of 7.00 m/s^2, whereas on wet concrete the maximum magnitude of acceleration is only 5.00 m/s^2. Find the distances necessary to stop a car moving at 30.0 m/s (about 110 km/h) (a) on dry concrete and (b) on wet concrete. (c) Repeat both calculations, finding the displacement from the point where the driver sees a traffic light turn red, taking into account his reaction time of 0.500 s to get his foot on the brake.

Strategy

Draw a sketch.

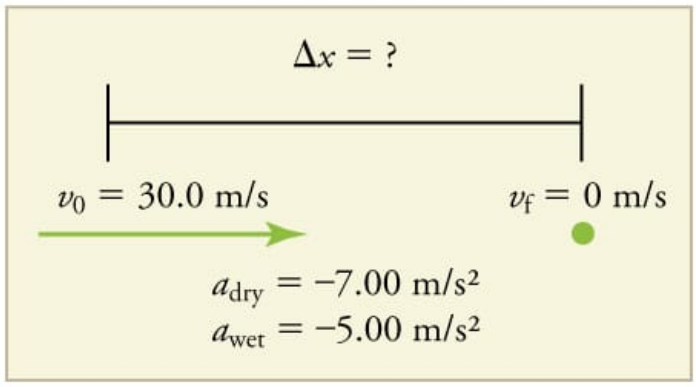


Figure
1.1.6.10

In order to determine which equations are best to use, we need to list all of the known values and identify exactly what we need to solve for. We shall do this explicitly in the next several examples, using tables to set them off.

Solution for (a)

1. Identify the knowns and what we want to solve for. We know that $v_0 = 30.0\ \mathrm{m/s};\ v = 0;\ a = -7.00\ \mathrm{m/s^2}$ (a is negative because it is in a direction opposite to velocity). We take x_0 to be 0. We are looking for displacement Δx, or $x - x_0$.

2. Identify the equation that will help up solve the problem. The best equation to use is

$$v^2 = v_0^2 + 2a\,(x - x_0)\,.$$

This equation is best because it includes only one unknown, x. We know the values of all the other variables in this equation. (There are other equations that would allow us to solve for x, but they require us to know the stopping time, t, which we do not know. We could use them but it would entail additional calculations.)

3. Rearrange the equation to solve for x.

$$x - x_0 = \frac{v^2 - v_0^2}{2a}$$

4. Enter known values.

$$x - 0 = \frac{0^2 - (30.0\ \mathrm{m/s})^2}{2\,(-7.00\ \mathrm{m/s^2})}$$

Thus,

$$x = 64.3\ \mathrm{m\ on\ dry\ concrete.}$$

Solution for (b)

This part can be solved in exactly the same manner as Part A. The only difference is that the acceleration is $-5.00\ \mathrm{m/s^2}$. The result is

$$x_{\mathrm{wet}}\ = 90.0\ \mathrm{m\ on\ wet\ concrete.}\ .$$

Solution for (c)

Once the driver reacts, the stopping distance is the same as it is in Parts A and B for dry and wet concrete. So to answer this question, we need to calculate how far the car travels during the reaction time, and then add that to the stopping time. It is reasonable to assume that the velocity remains constant during the driver's reaction time.

1. Identify the knowns and what we want to solve for. We know that $\bar{v} = 30.0 \text{ m/s}$; $t_{\text{reaction}} = 0.500 \text{ s}$; $a_{\text{reaction}} = 0$. We take $x_{0\text{-reaction}}$ to be 0. We are looking for $x_{reaction}$.

2. Identify the best equation to use.

$x = x_0 + \bar{v}t$ works well because the only unknown value is x, which is what we want to solve for.

3. Plug in the knowns to solve the equation.

$$x = 0 + (30.0 \text{ m/s})(0.500 \text{ s}) = 15.0 \text{ m}.$$

This means the car travels 15.0 m while the driver reacts, making the total displacements in the two cases of dry and wet concrete 15.0 m greater than if he reacted instantly.

4. Add the displacement during the reaction time to the displacement when braking.

$$x_{\text{braking}} + x_{\text{reaction}} = x_{\text{total}}$$

a. 64.3 m + 15.0 m = 79.3 m when dry
b. 90.0 m + 15.0 m = 105 m when wet

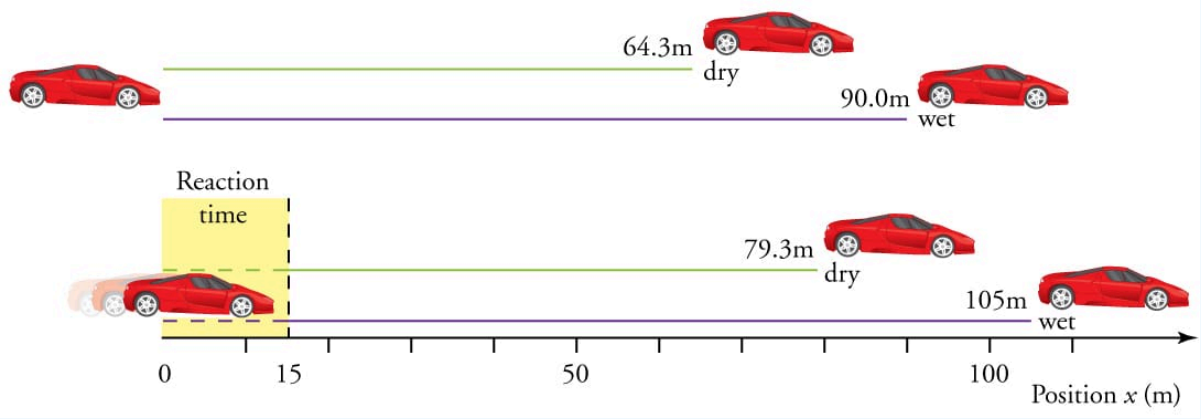


Figure 1.1.6.11: The distance necessary to stop a car varies greatly, depending on road conditions and driver reaction time. Shown here are the braking distances for dry and wet pavement, as calculated in this example, for a car initially traveling at 30.0 m/s. Also shown are the total distances traveled from the point where the driver first sees a light turn red, assuming a 0.500 s reaction time.

Discussion

The displacements found in this example seem reasonable for stopping a fast-moving car. It should take longer to stop a car on wet rather than dry pavement. It is interesting that reaction time adds significantly to the displacements. But more important is the general approach to solving problems. We identify the knowns and the quantities to be determined and then find an appropriate equation. There is often more than one way to solve a problem. The various parts of this example can in fact be solved by other methods, but the solutions presented above are the shortest.

With the basics of kinematics established, we can go on to many other interesting examples and applications. In the process of developing kinematics, we have also glimpsed a general approach to problem solving that produces both correct answers and insights into physical relationships.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—BREAKING NEWS

We have been using SI units of meters per second squared to describe some examples of acceleration of cars, runners, and trains. To achieve a better feel for these numbers, one can measure the braking acceleration of a car doing a slow (and safe) stop. Recall that, for average acceleration, $\bar{a} = \Delta v/\Delta t$. While traveling in a car, slowly apply the brakes as you come up to

a stop sign. Have a passenger note the initial speed in miles per hour and the time taken (in seconds) to stop. From this, calculate the acceleration in miles per hour per second. Convert this to meters per second squared and compare with other accelerations mentioned in this chapter. Calculate the distance traveled in braking.

Exercise 1.1.6.1

A manned rocket accelerates at a rate of $20\ \mathrm{m/s^2}$ during launch. How long does it take the rocket reach a velocity of 400 m/s?

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- To simplify calculations we take acceleration to be constant, so that $\bar{a} = a$ at all times.
- We also take initial time to be zero.
- Initial position and velocity are given a subscript 0; final values have no subscript. Thus,

$$\left.\begin{aligned}\Delta t &= t\\ \Delta x &= x - x_0\\ \Delta v &= v - v_0\end{aligned}\right\}$$

- The following kinematic equations for motion with constant a are useful:

$$\begin{aligned}x &= x_0 + \bar{v}t\\ \bar{v} &= \frac{v_0 + v}{2}\\ v &= v_0 + at\\ x &= x_0 + v_0 t + \frac{1}{2}at^2\\ v^2 &= v_0^2 + 2a\left(x - x_0\right)\end{aligned}$$

- In vertical motion, y is substituted for x.

Answers

Exercise 1.1.6.1

Exercise 1.1.6.1

To answer this, choose an equation that allows you to solve for time t, given only a, v_0, and v.

$$v = v_0 + at$$

Rearrange to solve for t.

$$t = \frac{v - v_0}{a} = \frac{400\ \mathrm{m/s} - 0\ \mathrm{m/s}}{20\ \mathrm{m/s^2}} = 20\ \mathrm{s}$$

This page titled 1.1.6: Motion Equations for Constant Acceleration in One Dimension is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.6: Motion Equations for Constant Acceleration in One Dimension by Andrew Park is licensed CC BY 4.0.

1.1.7: Falling Objects

Learning Objectives

- Describe the effects of gravity on objects in motion.
- Describe the motion of objects that are in free fall.
- Calculate the position and velocity of objects in free fall.

Falling objects form an interesting class of motion problems. For example, we can estimate the depth of a vertical mine shaft by dropping a rock into it and listening for the rock to hit the bottom. By applying the kinematics developed so far to falling objects, we can examine some interesting situations and learn much about gravity in the process.

Gravity

The most remarkable and unexpected fact about falling objects is that, if air resistance and friction are negligible, then in a given location all objects fall toward the center of Earth with the *same constant acceleration*, *independent of their mass*. This experimentally determined fact is unexpected, because we are so accustomed to the effects of air resistance and friction that we expect light objects to fall slower than heavy ones.

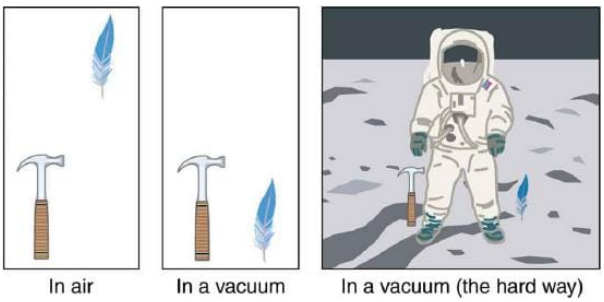


Figure 1.1.7.1: A hammer and a feather will fall with the same constant acceleration if air resistance is considered negligible. This is a general characteristic of gravity not unique to Earth, as astronaut David R. Scott demonstrated on the Moon in 1971, where the acceleration due to gravity is only $1.67\ \mathrm{m/s^2}$.

In the real world, air resistance can cause a lighter object to fall slower than a heavier object of the same size. A tennis ball will reach the ground after a hard baseball dropped at the same time. (It might be difficult to observe the difference if the height is not large.) Air resistance opposes the motion of an object through the air, while friction between objects—such as between clothes and a laundry chute or between a stone and a pool into which it is dropped—also opposes motion between them. For the ideal situations of these first few chapters, an object *falling without air resistance or friction* is defined to be in **free-fall**.

The force of gravity causes objects to fall toward the center of Earth. The acceleration of free-falling objects is therefore called the **acceleration due to gravity**. The acceleration due to gravity is *constant*, which means we can apply the kinematics equations to any falling object where air resistance and friction are negligible. This opens a broad class of interesting situations to us. The acceleration due to gravity is so important that its magnitude is given its own symbol, g. It is constant at any given location on Earth and has the average value

$$g = 9.80\ \mathrm{m/s^2}.$$

Although g varies from $9.78\ \mathrm{m/s^2}$ to $9.83\ \mathrm{m/s^2}$, depending on latitude, altitude, underlying geological formations, and local topography, the average value of $9.80\ \mathrm{m/s^2}$ will be used in this text unless otherwise specified. The direction of the acceleration due to gravity is *downward (towards the center of Earth)*. In fact, its direction *defines* what we call vertical. Note that whether the acceleration a in the kinematic equations has the value $+g$ or $-g$ depends on how we define our coordinate system. If we define the upward direction as positive, then $a = -g = -9.80\ \mathrm{m/s^2}$, and if we define the downward direction as positive, then $a = g = 9.80\ \mathrm{m/s^2}$. In certain examples and homework exercises, slightly inaccurate but more convenient value of $g \simeq 10\ \mathrm{m/s^2}$ will be used with appropriate specifications.

One-Dimensional Motion Involving Gravity

The best way to see the basic features of motion involving gravity is to start with the simplest situations and then progress toward more complex ones. So we start by considering straight up and down motion with no air resistance or friction. These assumptions mean that the velocity (if there is any) is vertical. If the object is dropped, we know the initial velocity is zero. Once the object has left contact with whatever held or threw it, the object is in free-fall. Under these circumstances, the motion is one-dimensional and has constant acceleration of magnitude g. We will also represent vertical displacement with the symbol y and use x for horizontal displacement.

KINEMATIC EQUATIONS FOR OBJECTS IN FREE-FALL

$$v = v_0 - gt$$

$$y = y_0 + v_0 t - \frac{1}{2} g t^2$$

$$v^2 = v_0^2 - 2g\,(y - y_0)$$

Example 1.1.7.1: Calculating Position and Velocity of a Falling Object: A Rock Thrown Upward

A person standing on the edge of a high cliff throws a rock straight up with an initial velocity of 13.0 m/s. The rock misses the edge of the cliff as it falls back to earth. Calculate the position and velocity of the rock 1.00 s, 2.00 s, and 3.00 s after it is thrown, neglecting the effects of air resistance.

Strategy

Draw a sketch.

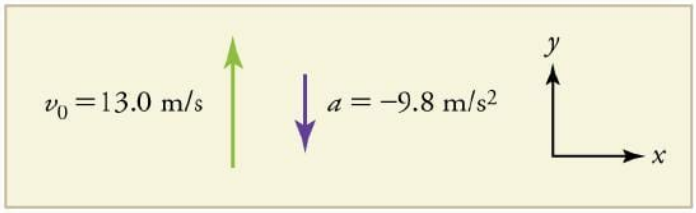


Figure
1.1.7.2

We are asked to determine the position y at various times. It is reasonable to take the initial position y_0 to be zero. This problem involves one-dimensional motion in the vertical direction. We use plus and minus signs to indicate direction, with up being positive and down negative. Since up is positive, and the rock is thrown upward, the initial velocity must be positive too. The acceleration due to gravity is downward, so a is negative. It is crucial that the initial velocity and the acceleration due to gravity have opposite signs. Opposite signs indicate that the acceleration due to gravity opposes the initial motion and will slow and eventually reverse it.

Since we are asked for values of position and velocity at three times, we will refer to these as y_1 and v_1; y_2 and v_2; and y_3 and v_3.

Solution for Position y_1

1. Identify the knowns. We know that $y_0 = 0;\ v_0 = 13.0\ \text{m/s};\ a = -g = -9.80\ \text{m/s}^2$; and $t = 1.00$ s.

2. Identify the best equation to use. We will use $y = y_0 + v_0 t + \frac{1}{2} a t^2$ because it includes only one unknown, y (or y_1, here), which is the value we want to find.

3. Plug in the known values and solve for y_1.

$$y_1 = 0 + (13.0 \text{ m/s})(1.00 \text{ s}) + \frac{1}{2}\left(-9.80 \text{ m/s}^2\right)(1.00 \text{ s})^2 = 8.10 \text{ m}$$

Discussion

The rock is 8.10 m above its starting point at $t = 1.00$ s, since $y_1 > y_0$. It could be *moving* up or down; the only way to tell is to calculate v_1 and find out if it is positive or negative.

Solution for Velocity v_1

1. Identify the knowns. We know that $y_0 = 0$; $v_0 = 13.0$ m/s; $a = -g = -9.80$ m/s^2; and $t = 1.00s$. We also know from the solution above that $y_1 = 8.10$ m.

2. Identify the best equation to use. The most straightforward is $v = v_0 - gt$ (from $v = v_0 + at$, where a = gravitational acceleration $= -g$).

3. Plug in the knowns and solve.

$$v_1 = v_0 - gt = 13.0 \text{ m/s} - \left(9.80 \text{ m/s}^2\right)(1.00 \text{ s}) = 3.20 \text{ m/s}$$

Discussion

The positive value for v_1 means that the rock is still heading upward at t**=1.00s**. However, it has slowed from its original 13.0 m/s, as expected.

Solution for Remaining Times

The procedures for calculating the position and velocity at t **= 2.00 s** and **3.00 s** are the same as those above. The results are summarized in Table 1.1.7.1 and illustrated in Figure 1.1.7.3.

Table 1.1.7.1: Results

Time, *t*	Position, *y*	Velocity, *v*	Acceleration, *a*
1.00 s	8.10 m	3.20 m	-9.80 m/s^2
2.00 s	6.40 m	–6.60 m	-9.80 m/s^2
3.00 s	–5.10 m	–16.4 m	-9.80 m/s^2

Graphing the data helps us understand it more clearly.

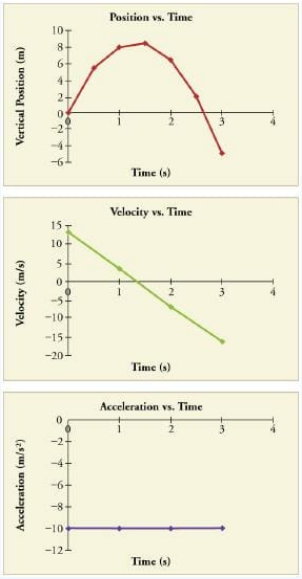


Figure 1.1.7.3: Vertical position, vertical velocity, and vertical acceleration vs. time for a rock thrown vertically up at the edge of a cliff. Notice that velocity changes linearly with time and that acceleration is constant. *Misconception Alert!* Notice that the position vs. time graph shows vertical position only. It is easy to get the impression that the graph shows some horizontal motion—the shape of the graph looks like the path of a projectile. But this is not the case; the horizontal axis is *time*, not space. The actual path of the rock in space is straight up, and straight down.

Discussion

The interpretation of these results is important. At 1.00 s the rock is above its starting point and heading upward, since y_1 and v_1 are both positive. At 2.00 s, the rock is still above its starting point, but the negative velocity means it is moving downward. At 3.00 s, both y_3 and v_3 are negative, meaning the rock is below its starting point and continuing to move downward. Notice that when the rock is at its highest point (at 1.5 s), its velocity is zero, but its acceleration is still $-9.80\ \mathrm{m/s^2}$. Its acceleration is $-9.80\ \mathrm{m/s^2}$ for the whole trip—while it is moving up and while it is moving down. Note that the values for y are the positions (or displacements) of the rock, not the total distances traveled. Finally, note that free-fall applies to upward motion as well as downward. Both have the same acceleration—the acceleration due to gravity, which remains constant the entire time. Astronauts training in the famous Vomit Comet, for example, experience free-fall while arcing up as well as down, as we will discuss in more detail later.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—REACTION TIME

A simple experiment can be done to determine your reaction time. Have a friend hold a ruler between your thumb and index finger, separated by about 1 cm. Note the mark on the ruler that is right between your fingers. Have your friend drop the ruler unexpectedly, and try to catch it between your two fingers. Note the new reading on the ruler. Assuming acceleration is that due to gravity, calculate your reaction time. How far would you travel in a car (moving at 30 m/s) if the time it took your foot to go from the gas pedal to the brake was twice this reaction time?

Example 1.1.7.2: Calculating Velocity of a Falling Object: A Rock Thrown Down

What happens if the person on the cliff throws the rock straight down, instead of straight up? To explore this question, calculate the velocity of the rock when it is 5.10 m below the starting point, and has been thrown downward with an initial speed of 13.0 m/s.

Strategy

Draw a sketch.

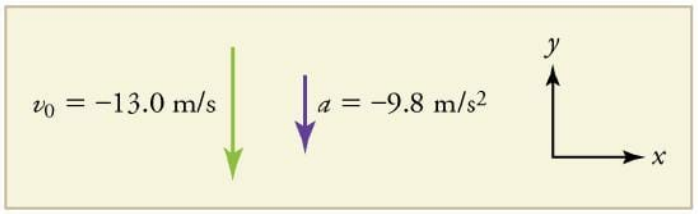


Figure
1.1.7.4

Since up is positive, the final position of the rock will be negative because it finishes below the starting point at $y_0 = 0$. Similarly, the initial velocity is downward and therefore negative, as is the acceleration due to gravity. We expect the final velocity to be negative since the rock will continue to move downward.

Solution

1. Identify the knowns. $y_0 = 0;\ y_1 = -5.10\ \text{m};\ v_0 = -13.0\ \text{m/s};\ a = -g = -9.80\ \text{m/s}^2$.

2. Choose the kinematic equation that makes it easiest to solve the problem. The equation $v^2 = v_0^2 + 2a\,(y - y_0)$ works well because the only unknown in it is v. (We will plug y_1 in for y.)

3. Enter the known values

$$v^2 = (-13.0\ \text{m/s})^2 + 2\left(-9.80\ \text{m/s}^2\right)(-5.10\ \text{m} - 0\ \text{m}) = 268.96\ \text{m}^2/\text{s}^2,$$

where we have retained extra significant figures because this is an intermediate result.

Taking the square root, and noting that a square root can be positive or negative, gives

$$v = \pm 16.4\ \text{m/s}.$$

The negative root is chosen to indicate that the rock is still heading down. Thus,

$$v = -16.4\ \text{m/s}.$$

Discussion

Note that *this is exactly the same velocity the rock had at this position when it was thrown straight upward with the same initial speed*. (See Example 1.1.7.1 and Figure 1.1.7.5 (a).) This is not a coincidental result. Because we only consider the acceleration due to gravity in this problem, the *speed* of a falling object depends only on its initial speed and its vertical position relative to the starting point. For example, if the velocity of the rock is calculated at a height of 8.10 m above the starting point (using the method from Example 1.1.7.1) when the initial velocity is 13.0 m/s straight up, a result of $\pm 3.20\ \text{m/s}$ is obtained. Here both signs are meaningful; the positive value occurs when the rock is at 8.10 m and heading up, and the negative value occurs when the rock is at 8.10 m and heading back down. It has the same *speed* but the opposite direction.

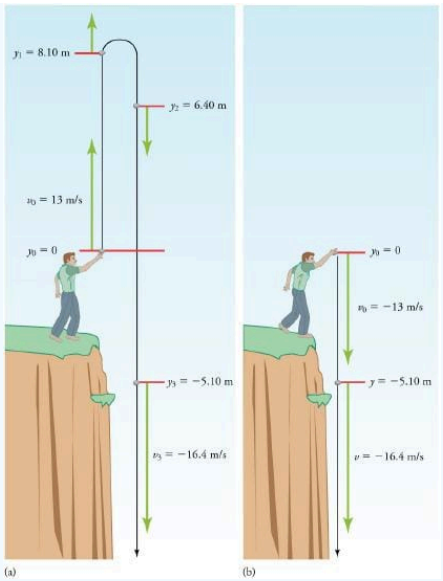


Figure 1.1.7.5: (a) A person throws a rock straight up, as explored in Example 1.1.7.1. The arrows are velocity vectors at 0, 1.00, 2.00, and 3.00 s. (b) A person throws a rock straight down from a cliff with the same initial speed as before, as in Example 1.1.7.2. Note that at the same distance below the point of release, the rock has the same velocity in both cases.

Another way to look at it is this: In Example 1.1.7.1, the rock is thrown up with an initial velocity of **13.0 m/s**. It rises and then falls back down. When its position is $y = 0$ on its way back down, its velocity is **–13.0 m/s**. That is, it has the same speed on its way down as on its way up. We would then expect its velocity at a position of $y = -5.10\text{m}$ to be the same whether we have thrown it upwards at **+13.0 m/s** or thrown it downwards at **–13.0 m/s**. The velocity of the rock on its way down from $y = 0$ is the same whether we have thrown it up or down to start with, as long as the speed with which it was initially thrown is the same.

Example 1.1.7.3: Find g from Data on a Falling Object

The acceleration due to gravity on Earth differs slightly from place to place, depending on topography (e.g., whether you are on a hill or in a valley) and subsurface geology (whether there is dense rock like iron ore as opposed to light rock like salt beneath you.) The precise acceleration due to gravity can be calculated from data taken in an introductory physics laboratory course. An object, usually a metal ball for which air resistance is negligible, is dropped and the time it takes to fall a known distance is measured. See, for example, Figure 1.1.7.6. Very precise results can be produced with this method if sufficient care is taken in measuring the distance fallen and the elapsed time.

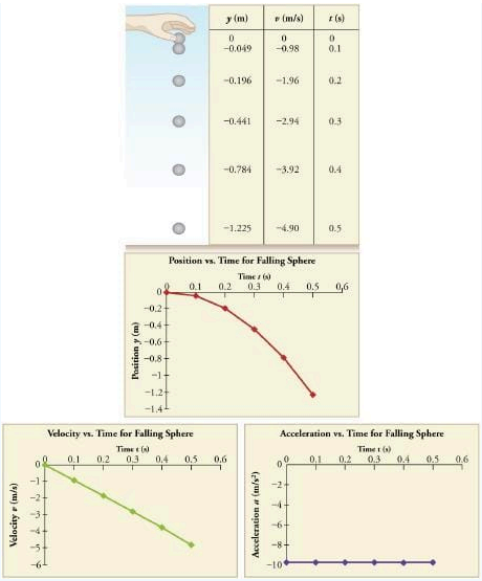


Figure 1.1.7.6: Positions and velocities of a metal ball released from rest when air resistance is negligible. Velocity is seen to increase linearly with time while displacement increases with time squared. Acceleration is a constant and is equal to gravitational acceleration.

Suppose the ball falls 1.0000 m in 0.45173 s. Assuming the ball is not affected by air resistance, what is the precise acceleration due to gravity at this location?

Strategy

Draw a sketch.

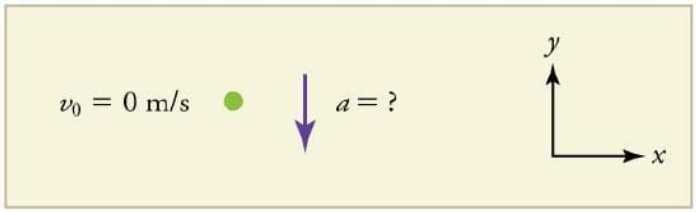


Figure
1.1.7.7

We need to solve for acceleration a. Note that in this case, displacement is downward and therefore negative, as is acceleration.

Solution

1. Identify the knowns. $y_0 = 0;\ y = -1.0000\ \text{m};\ t = 0.45173;\ v_0 = 0.$.

2. Choose the equation that allows you to solve for a using the known values.

$$y = y_0 + v_0 t + \frac{1}{2}at^2$$

3. Substitute 0 for v_0 and rearrange the equation to solve for a. Substituting 0 for v_0 yields

$$y = y_0 + \frac{1}{2}at^2.$$

Solving for a gives

$$a = \frac{2\,(y - y_0)}{t^2}.$$

4. Substitute known values yields

$$a = \frac{2(-1.0000\ \text{m} - 0)}{(0.45173\ \text{s})^2} = -9.8010\ \text{m/s}^2,$$

so, because $a = -g$ with the directions we have chosen,

$$g = 9.8010\ \text{m/s}^2.$$

Discussion

The negative value for a indicates that the gravitational acceleration is downward, as expected. We expect the value to be somewhere around the average value of $9.80\ \text{m/s}^2$, so $9.8010\ \text{m/s}^2$ makes sense. Since the data going into the calculation are relatively precise, this value for g is more precise than the average value of $9.80\ \text{m/s}^2$; it represents the local value for the acceleration due to gravity.

Exercise 1.1.7.1

A chunk of ice breaks off a glacier and falls 30.0 meters before it hits the water. Assuming it falls freely (there is no air resistance), how long does it take to hit the water?

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- An object in free-fall experiences constant acceleration if air resistance is negligible.
- On Earth, all free-falling objects have an acceleration due to gravity g, which averages

$$g = 9.80\ \text{m/s}^2.$$

- Whether the acceleration a should be taken as $+g$ or $-g$ is determined by your choice of coordinate system. If you choose the upward direction as positive,$a = -g = -9.80\ \text{m/s}^2$ is negative. In the opposite case, $a = +\text{g} = 9.80\ \text{m/s}^2$ is positive. Since acceleration is constant, the kinematic equations above can be applied with the appropriate $+g$ or $-g$ substituted for a.
- For objects in free-fall, up is normally taken as positive for displacement, velocity, and acceleration.

Glossary

free-fall
: the state of movement that results from gravitational force only

acceleration due to gravity
: acceleration an object undergoes while in free-fall

Answers

Exercise 1.1.7.1

Exercise 1.1.7.1

We know that initial position $y_0 = 0$, final position $y = -30.0\ \text{m}$, and $a = -g = -9.80\ \text{m/s}^2$. We can then use the equation $y = y_0 + v_0 t + \frac{1}{2}at^2$ to solve for t. Inserting $a = -g$, we obtain

$$y = 0 + 0 - \frac{1}{2}gt^2$$
$$t^2 = \frac{2y}{-g}$$
$$t \quad = \pm\sqrt{\frac{2y}{-g}} = \pm\sqrt{\frac{2(-30.0\text{ m})}{-9.80\text{ m/s}^2}} = \pm\sqrt{6.12\text{ s}^2} = 2.47\text{ s} \approx 2.5\text{ s}$$

where we take the positive value as the physically relevant answer. Thus, it takes about 2.5 seconds for the piece of ice to hit the water.

This page titled 1.1.7: Falling Objects is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.7: Falling Objects by Andrew Park is licensed CC BY 4.0.

1.1.8: Projectile Motion

Learning Objectives

- Identify and explain the properties of a projectile, such as acceleration due to gravity, range, and trajectory.
- Apply the principle of independence of motion to solve projectile motion problems.

Projectile motion is the **motion** of an object thrown or projected into the air, subject to only the acceleration of gravity. The object is called a **projectile**, and its path is called its **trajectory**. The motion of falling objects is a simple one-dimensional type of projectile motion in which there is no horizontal movement. In this section, we consider two-dimensional projectile motion, such as that of a football or other object for which **air resistance** *is negligible*.

The most important fact to know here is that *motions along perpendicular axes are independent* and thus can be considered separately. In particular, it is often convenient to consider vertical and horizontal motions separately, or independently. The key to analyzing two-dimensional projectile motion is to break it into two motions, one along the horizontal axis and the other along the vertical. And we will assume all forces except gravity are negligible (that is, we ignore forces due to air resistance, wind, friction, etc.). The vertical component of acceleration is, then, $a_y = -g = -9.80\ \mathrm{m/s^2}$, where we take upward direction as the positive y direction. And because there is no force in the horizontal direction, $a_x = 0$. Figure $1.1.8.1$ illustrates this approach to analyzing projectile motion.

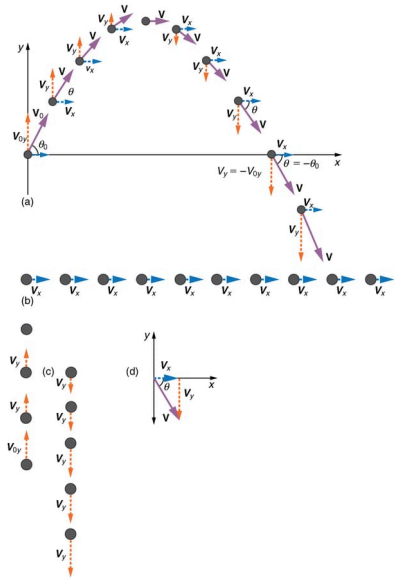


Figure $1.1.8.1$: (a) We analyze two-dimensional projectile motion by breaking it into two independent one-dimensional motions along the vertical and horizontal axes. (b) The horizontal motion is simple, because $a_x = 0$ and v_x is thus constant. (c) The velocity in the vertical direction begins to decrease as the object rises; at its highest point, the vertical velocity is zero. As the object falls towards the Earth again, the vertical velocity increases again in magnitude. This is exactly same as in the case of falling objects (or objects thrown directly upward). (d) The *x* - and *y* -motions can be recombined to give the total velocity at any given point on the trajectory.

One of the conceptual aspects of projectile motion we can discuss without a detailed analysis is the range. On level ground, we define **range** to be the horizontal distance R traveled by a projectile. The range of projectiles describes everyday phenomena such as how far a football can be thrown or kicked. Also, investigating the range of projectiles can shed light on other interesting phenomena, such as the orbits of satellites around the Earth. Let us consider projectile range further.

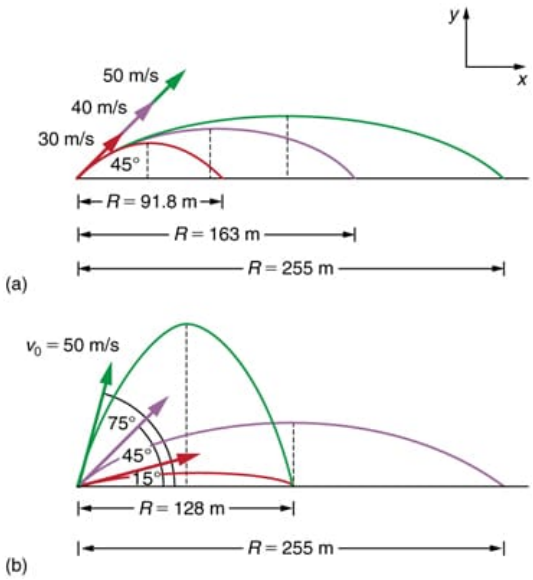


Figure 1.1.8.2: Trajectories of projectiles on level ground. (a) The greater the initial speed v_0, the greater the range for a given initial angle. (b) The effect of initial angle θ_0 on the range of a projectile with a given initial speed. Note that the range is the same for 15° and 75°, although the maximum heights of those paths are different.

How does the initial velocity of a projectile affect its range? It is intuitive to guess that, the greater the initial speed v_0, the greater the range, as shown in Figure 1.1.8.2 (a). The initial angle θ_0 also has a dramatic effect on the range, as illustrated in Figure 1.1.8.2 (b). For a fixed initial speed, such as might be produced by a cannon, the maximum range is obtained with $\theta_0 = 45^\circ$. Interestingly, for every initial angle except 45°, there are two angles that give the same range. The range also depends on the value of the acceleration of gravity g. The lunar astronaut Alan Shepherd was able to drive a golf ball a great distance on the Moon because gravity is weaker there. The range R of a projectile on *level ground* for which air resistance is negligible is given by

$$R = \frac{v_0^2 \sin 2\theta_0}{g},$$

where v_0 is the initial speed and θ_0 is the initial angle relative to the horizontal. This formula (which can be derived using algebra and trigonometry) fits the major features of projectile range as described.

When we speak of the range of a projectile on level ground, we assume that R is very small compared with the circumference of the Earth. If, however, the range is large, the Earth curves away below the projectile and acceleration of gravity changes direction along the path. The range is larger than predicted by the range equation given above because the projectile has farther to fall than it would on level ground. (See Figure 1.1.8.3.) If the initial speed is great enough, the projectile goes into orbit. This possibility was recognized centuries before it could be accomplished. When an object is in orbit, the Earth curves away from underneath the object at the same rate as it falls. The object thus falls continuously but never hits the surface.

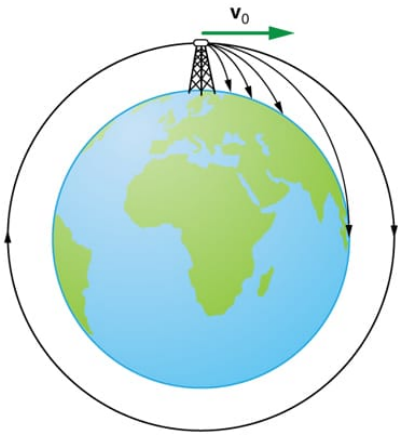


Figure 1.1.8.3: Projectile to satellite. In each case shown here, a projectile is launched from a very high tower to avoid air resistance. With increasing initial speed, the range increases and becomes longer than it would be on level ground because the Earth curves away underneath its path. With a large enough initial speed, orbit is achieved.

Summary

- Projectile motion is the motion of an object through the air that is subject only to the acceleration of gravity.
- The most important fact regarding projectile motion is that motions along vertical direction and the horizontal direction are independent.
- The maximum horizontal distance traveled by a projectile is called the **range**. The range R of a projectile on level ground launched at an angle θ_0 above the horizontal with initial speed v_0 is given by

$$R = \frac{v_0^2 \sin 2\theta_0}{g}.$$

Glossary

centripetal acceleration
: the acceleration of an object moving in a circle, directed toward the center

ultracentrifuge
: a centrifuge optimized for spinning a rotor at very high speeds

uniform circular motion
: an object moving in a circular path at constant speed

This page titled 1.1.8: Projectile Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.8: Projectile Motion by Andrew Park is licensed CC BY 4.0.

1.1.9: Centripetal Acceleration

Learning Objectives

- Explain what centripetal acceleration is.
- Use the formula for centripetal acceleration in simple situations.

We defined acceleration as a change in velocity, either in its magnitude or in its direction, or both. When an object moves along a circular path, the direction of its velocity changes constantly, so there is always an associated acceleration, even if the speed of the object is constant. You experience this acceleration yourself when you turn a corner in your car. What you notice is a sideways acceleration because you and the car are changing direction. The sharper the curve and the greater your speed, the more noticeable this acceleration will become. In this section we briefly examine the direction and magnitude of that acceleration.

Figure 1.1.9.1 shows an object moving in a circular path at constant speed, called **uniform circular motion**. The direction of the instantaneous velocity is shown at two points along the path. Acceleration is in the direction of the change in velocity, which points directly toward the center of rotation (the center of the circular path). This pointing is shown with the vector diagram in the figure. We call the acceleration of an object moving in uniform circular motion (resulting from a net external force) the **centripetal acceleration**(a_c); centripetal means "toward the center" or "center seeking."

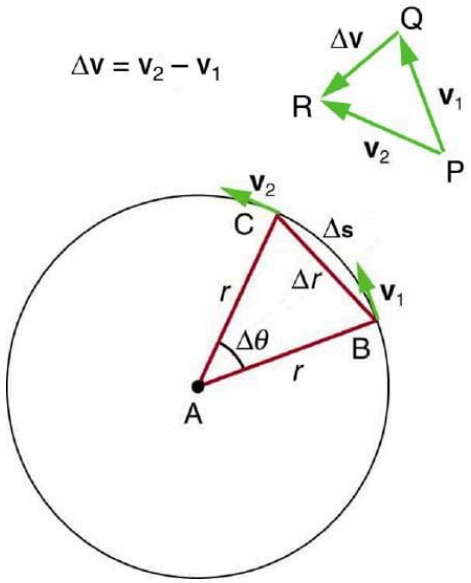


Figure 1.1.9.1: The directions of the velocity of an object at two different points are shown, and the change in velocity $\Delta \mathbf{v}$ is seen to point directly toward the center of curvature. (See small inset.) Because $\mathbf{a}_c = \Delta \mathbf{v}/\Delta t$, the acceleration is also toward the center; a_c is called centripetal acceleration. (For small time differences, $\Delta\theta$ is very small, and the arc length Δs is approximately equal to the chord length Δr.)

The direction of centripetal acceleration is toward the center of curvature, but what is its magnitude? If we use the geometry shown in Figure 1.1.9.1 along with some kinematics equations, we can obtain (detailed derivation skipped)

$$a_c = \frac{v^2}{r}, \tag{1.1.9.1}$$

which is the acceleration of an object in a circle of radius r at a speed v. Verify for yourself that a_c has unit of $\mathrm{m/s^2}$, as expected for acceleration.

We see in Equation (1.1.9.1) that centripetal acceleration is greater at high speeds and in sharp curves (smaller radius), as you have noticed when driving a car. But it is a bit surprising that a_c is proportional to speed squared, implying, for example, that it is four times as hard to take a curve at 100 km/h than at 50 km/h. A sharp corner has a small radius, so that a_c is greater for tighter turns, as you have probably noticed.

Section Summary

- Centripetal acceleration a_{c} is the acceleration experienced while in uniform circular motion. It always points toward the center of rotation. It is perpendicular to the linear velocity v and has the magnitude

$$a_{\mathrm{c}} = \frac{v^2}{r}.$$

- The unit of centripetal acceleration is $\mathrm{m/s^2}$.

This page titled 1.1.9: Centripetal Acceleration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.9: Centripetal Acceleration by Andrew Park is licensed CC BY 4.0.

1.1.10: Kinematics (Exercise)

Conceptual Questions

1.1: Displacement

1. Give an example in which there are clear distinctions among distance traveled, displacement, and magnitude of displacement. Specifically identify each quantity in your example.

2. Under what circumstances does distance traveled equal magnitude of displacement? What is the only case in which magnitude of displacement and displacement are exactly the same?

3. Bacteria move back and forth by using their flagella (structures that look like little tails). Speeds of up to $50\ \mu\text{m/s}\ \left(50 \times 10^{-6}\ \text{m/s}\right)$ have been observed. The total distance traveled by a bacterium is large for its size, while its displacement is small. Why is this?

1.2: Vectors, Scalars, and Coordinate Systems

4. A student writes, "*A bird that is diving for prey has a speed of* –10 m/s." What is wrong with the student's statement? What has the student actually described? Explain.

5. What is the speed of the bird in Exercise 2.2.4?

6. Acceleration is the change in velocity over time. Given this information, is acceleration a vector or a scalar quantity? Explain.

7. A weather forecast states that the temperature is predicted to be –5°C the following day. Is this temperature a vector or a scalar quantity? Explain.

1.3: Time, Velocity, and Speed

8. Give an example (but not one from the text) of a device used to measure time and identify what change in that device indicates a change in time.

9. There is a distinction between average speed and the magnitude of average velocity. Give an example that illustrates the difference between these two quantities.

10. Does a car's odometer measure position or displacement? Does its speedometer measure speed or velocity?

11. If you divide the total distance traveled on a car trip (as determined by the odometer) by the time for the trip, are you calculating the average speed or the magnitude of the average velocity? Under what circumstances are these two quantities the same?

12. How are instantaneous velocity and instantaneous speed related to one another? How do they differ?

1.4: Acceleration

13. Is it possible for speed to be constant while acceleration is not zero? Give an example of such a situation.

14. Is it possible for velocity to be constant while acceleration is not zero? Explain.

15. Give an example in which velocity is zero yet acceleration is not.

16. If a subway train is moving to the left (has a negative velocity) and then comes to a stop, what is the direction of its acceleration? Is the acceleration positive or negative?

17. Plus and minus signs are used in one-dimensional motion to indicate direction. What is the sign of an acceleration that reduces the magnitude of a negative velocity? Of a positive velocity?

1.5: Motion Equations for Constant Acceleration in One Dimension

18. In a constant-acceleration motion, how does the position change? How does the velocity change?

19. Describe the relationship between the position and the velocity in a constant-acceleration motion.

20. Describe the relationship between the velocity and the acceleration in a constant-acceleration motion.

21. In a constant-acceleration motion starting from rest, how does the distance traveled change with the duration of motion? If the motion continues for double the time (total), how much does the distance traveled increase (total)?

1.6: Falling Objects

22. What is the acceleration of a rock thrown straight upward on the way up? At the top of its flight? On the way down?

23. An object that is thrown straight up falls back to Earth. This is one-dimensional motion. (a) When is its velocity zero? (b) Does its velocity change direction? (c) Does the acceleration due to gravity have the same sign on the way up as on the way down?

24. Suppose you throw a rock nearly straight up at a coconut in a palm tree, and the rock misses on the way up but hits the coconut on the way down. Neglecting air resistance, how does the speed of the rock when it hits the coconut on the way down compare with what it would have been if it had hit the coconut on the way up? Is it more likely to dislodge the coconut on the way up or down? Explain.

25. If an object is thrown straight up and air resistance is negligible, then its speed when it returns to the starting point is the same as when it was released. If air resistance were not negligible, how would its speed upon return compare with its initial speed? How would the maximum height to which it rises be affected?

26. The severity of a fall depends on your speed when you strike the ground. All factors but the acceleration due to gravity being the same, how many times higher could a safe fall on the Moon be than on Earth (gravitational acceleration on the Moon is about 1/6 that of the Earth)?

27. How many times higher could an astronaut jump on the Moon than on Earth if his takeoff speed is the same in both locations (gravitational acceleration on the Moon is about 1/6 of g on Earth)?

1.7: Projectile Motion

28. Answer the following questions for projectile motion on level ground assuming negligible air resistance (the initial angle being neither 0° nor 90°): (a) Is the velocity ever zero? (b) When is the velocity a minimum? A maximum? (c) Can the velocity ever be the same as the initial velocity at a time other than at $t = 0$? (d) Can the speed ever be the same as the initial speed at a time other than at $t = 0$?

29. Answer the following questions for projectile motion on level ground assuming negligible air resistance (the initial angle being neither 0° nor 90°): (a) Is the acceleration ever zero? (b) Is the acceleration ever in the same direction as a component of velocity? (c) Is the acceleration ever opposite in direction to a component of velocity?

30. For a fixed initial speed, the range of a projectile is determined by the angle at which it is fired. For all but the maximum, there are two angles that give the same range. Considering factors that might affect the ability of an archer to hit a target, such as wind, explain why the smaller angle (closer to the horizontal) is preferable. When would it be necessary for the archer to use the larger angle? Why does the punter in a football game use the higher trajectory?

31. During a lecture demonstration, a professor places two coins on the edge of a table. She then flicks one of the coins

horizontally off the table, simultaneously nudging the other over the edge. Describe the subsequent motion of the two coins, in particular discussing whether they hit the floor at the same time.

1.8: Centripetal Acceleration

32. Give examples of centripetal acceleration from everyday experience.

33. A satellite in a circular orbit around the Earth is undergoing a centripetal acceleration. Explain how this description is consistent with the description that the satellite in free-fall.

34. Can centripetal acceleration change the speed of circular motion? Explain.

Problems & Exercises

1.1: Displacement

35.

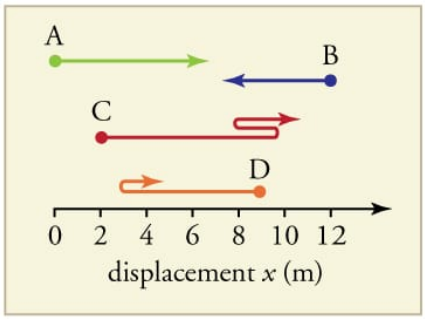


Figure 1.1.10.1

Find the following for path A in Figure 1.1.10.1: (a) The distance traveled. (b) The magnitude of the displacement from start to finish. (c) The displacement from start to finish.

Answer

(a) 7 m

(b) 7 m

(c) +7 m

36. Find the following for path B in Figure 1.1.10.1: (a) The distance traveled. (b) The magnitude of the displacement from start to finish. (c) The displacement from start to finish.

37. Find the following for path C in Figure 1.1.10.1: (a) The distance traveled. (b) The magnitude of the displacement from start to finish. (c) The displacement from start to finish.

Answer

(a) 13 m

(b) 9 m

(c) +9 m

38. Find the following for path D in Figure 1.1.10.1: (a) The distance traveled. (b) The magnitude of the displacement from start to finish. (c) The displacement from start to finish.

This page titled 1.1.10: Kinematics (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 2.10: Kinematics (Exercise) by Andrew Park is licensed CC BY 4.0.

1.2: DYNAMICS

1.2.1: Introduction to Dynamics- Newton’s Laws of Motion

1.2.2: Development of Force Concept

1.2.3: Newton’s First Law of Motion- Inertia

1.2.4: Newton's Second Law of Motion- Force and Acceleration

1.2.5: Newton’s Third Law of Motion- Symmetry in Forces

1.2.6: Normal Force and Tension

1.2.7: Spring Force- Hooke’s Law

1.2.8: Friction

1.2.9: Newton’s Universal Law of Gravitation

1.2.10: Centripetal Force

1.2.11: Dynamics (Exercise)

This page titled 1.2: Dynamics is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

- 3: Dynamics by Andrew Park is licensed CC BY 4.0.

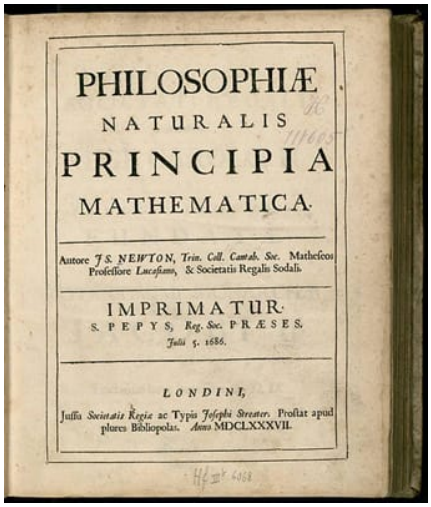
1.2.1: Introduction to Dynamics- Newton's Laws of Motion



Figure 1.2.1.1: Newton's laws of motion describe the motion of the dolphin's path. (credit: Jin Jang)

Motion draws our attention. Motion itself can be beautiful, causing us to marvel at the forces needed to achieve spectacular motion, such as that of a dolphin jumping out of the water, or a pole vaulter, or the flight of a bird, or the orbit of a satellite. The study of motion is kinematics, but kinematics only *describes* the way objects move—their velocity and their acceleration. **Dynamics** considers the forces that affect the motion of moving objects and systems. Newton's laws of motion are the foundation of dynamics. These laws provide an example of the breadth and simplicity of principles under which nature functions. They are also universal laws in that they apply to similar situations on Earth as well as in space.

Issac Newton's (1642–1727) laws of motion were just one part of the monumental work that has made him legendary. The development of Newton's laws marks the transition from the Renaissance into the modern era. This transition was characterized by a revolutionary change in the way people thought about the physical universe. For many centuries natural philosophers had debated the nature of the universe based largely on certain rules of logic with great weight given to the thoughts of earlier classical philosophers such as Aristotle (384–322 BC). Among the many great thinkers who contributed to this change were Newton and Galileo.


PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA

Autore *J.S. NEWTON, Trin. Coll. Cantab. Soc.* Matheseos Professore *Lucasiano*, & Societatis Regalis Sodali.

IMPRIMATUR.
S. PEPYS, *Reg. Soc.* PRÆSES.
Julii 5. 1686.

LONDINI,
Justu *Societatis Regiæ* ac Typis *Josephi Streater*. Prostat apud plures Bibliopolas. *Anno* MDCLXXXVII.

Figure 1.2.1.2: Issac Newton's monumental work, *Philosophiae Naturalis Principia Mathematica*, was published in 1687. It proposed scientific laws that are still used today to describe the motion of objects. (credit: Service commun de la documentation de l'Université de Strasbourg)

Galileo was instrumental in establishing *observation* as the absolute determinant of truth, rather than "logical" argument. Galileo's use of the telescope was his most notable achievement in demonstrating the importance of observation. He discovered moons orbiting Jupiter and made other observations that were inconsistent with certain ancient ideas and religious dogma. For this reason, and because of the manner in which he dealt with those in authority, Galileo was tried by the Inquisition and punished. He spent the final years of his life under a form of house arrest. Because others before Galileo had also made discoveries by *observing* the nature of the universe, and because repeated observations verified those of Galileo, his work could not be suppressed or denied. After his death, his work was verified by others, and his ideas were eventually accepted by the church and scientific communities.

Galileo also contributed to the formation of what is now called Newton's first law of motion. Newton made use of the work of his predecessors, which enabled him to develop laws of motion, discover the law of gravity, invent calculus, and make great contributions to the theories of light and color. It is amazing that many of these developments were made with Newton working alone, without the benefit of the usual interactions that take place among scientists today.

It was not until the advent of modern physics early in the 20th century that it was discovered that Newton's laws of motion produce a good approximation to motion only when the objects are moving at speeds much, much less than the speed of light and when those objects are larger than the size of most molecules (about 10^{-9} m in diameter). These constraints define the realm of classical mechanics. At the beginning of the 20th century, Albert Einstein (1879–1955) developed the theory of relativity and, along with many other scientists, developed quantum theory. This theory does not have the constraints present in classical physics. All of the situations we consider in this chapter are in the realm of classical physics.

MAKING CONNECTIONS: PAST AND PRESENT PHILOSOPHY

The importance of observation and the concept of *cause and effect* were not always so entrenched in human thinking. This realization was a part of the evolution of modern physics from natural philosophy. The achievements of Galileo, Newton, Einstein, and others were key milestones in the history of scientific thought. Most of the scientific theories that are described in this book descended from the work of these scientists.

This page titled 1.2.1: Introduction to Dynamics- Newton's Laws of Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.1: Introduction to Dynamics- Newton's Laws of Motion by Andrew Park is licensed CC BY 4.0.

1.2.2: Development of Force Concept

Learning Objectives

- Understand the definition of force.

Dynamics is the study of the forces that cause objects and systems to move. To understand this, we need a working definition of force. Our intuitive definition of **force**—that is, a push or a pull—is a good place to start. We know that a push or pull has both magnitude and direction (therefore, it is a vector quantity) and can vary considerably in each regard. For example, a cannon exerts a strong force on a cannonball that is launched into the air. In contrast, Earth exerts only a tiny downward pull on a flea.

A more quantitative definition of force can be based on some standard force, just as distance is measured in units relative to a standard distance. One possibility is to stretch a spring a certain fixed distance, as illustrated in Figure 1.2.2.1, and use the force it exerts to pull itself back to its relaxed shape—called a *restoring force*—as a standard. The magnitude of all other forces can be stated as multiples of this standard unit of force. Many other possibilities exist for standard forces. Some alternative definitions of force will be given later in this chapter.

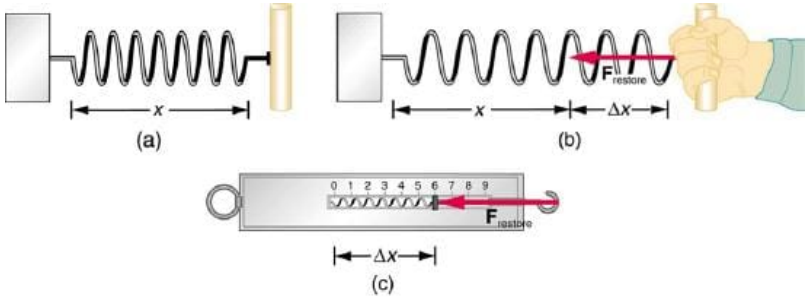


Figure 1.2.2.1: The force exerted by a stretched spring can be used as a standard unit of force. (a) This spring has a length x when undistorted. (b) When stretched a distance Δx, the spring exerts a restoring force, $\mathbf{F}_{\text{restore}}$, which is reproducible. (c) A spring scale is one device that uses a spring to measure force. The force $\mathbf{F}_{\text{restore}}$ is exerted on whatever is attached to the hook. Here $\mathbf{F}_{\text{restore}}$ has a magnitude of 6 units in the force standard being employed.

TAKE-HOME EXPERIMENT: FORCE STANDARDS

To investigate force standards and cause and effect, get two identical rubber bands. Hang one rubber band vertically on a hook. Find a small household item that could be attached to the rubber band using a paper clip, and use this item as a weight to investigate the stretch of the rubber band. Measure the amount of stretch produced in the rubber band with one, two, and four of these (identical) items suspended from the rubber band. What is the relationship between the number of items and the amount of stretch? How large a stretch would you expect for the same number of items suspended from two rubber bands? What happens to the amount of stretch of the rubber band (with the weights attached) if the weights are also pushed to the side with a pencil?

Section Summary

- **Dynamics** is the study of how forces affect the motion of objects.
- **Force** is a push or pull that can be defined in terms of various standards, and it is a vector having both magnitude and direction.

Glossary

dynamics
: the study of how forces affect the motion of objects and systems

force

a push or pull on an object with a specific magnitude and direction; can be represented by vectors; can be expressed as a multiple of a standard force

This page titled 1.2.2: Development of Force Concept is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.2: Development of Force Concept by Andrew Park is licensed CC BY 4.0.

1.2.3: Newton’s First Law of Motion- Inertia

Learning Objectives

- Define mass and inertia.
- Understand Newton's first law of motion.

Experience suggests that an object at rest will remain at rest if left alone, and that an object in motion tends to slow down and stop unless some effort is made to keep it moving. What **Newton’s first law of motion** states, however, is the following:

NEWTON’S FIRST LAW OF MOTION

A body at rest remains at rest, or, if in motion, remains in motion at a constant velocity unless acted on by a net external force.

Note the repeated use of the verb “remains.” We can think of this law as preserving the status quo of motion.

Rather than contradicting our experience, **Newton’s first law of motion** states that there must be a *cause* (which is a net external force) *for there to be any change in velocity (either a change in magnitude or direction)*. We will define *net external force* in the next section. An object sliding across a table or floor slows down due to the net force of friction acting on the object. If friction disappeared, would the object still slow down?

The idea of cause and effect is crucial in accurately describing what happens in various situations. For example, consider what happens to an object sliding along a rough horizontal surface. The object quickly grinds to a halt. If we spray the surface with talcum powder to make the surface smoother, the object slides farther. If we make the surface even smoother by rubbing lubricating oil on it, the object slides farther yet. Extrapolating to a frictionless surface, we can imagine the object sliding in a straight line indefinitely. Friction is thus the *cause* of the slowing (consistent with Newton’s first law). The object would not slow down at all if friction were completely eliminated. Consider an air hockey table. When the air is turned off, the puck slides only a short distance before friction slows it to a stop. However, when the air is turned on, it creates a nearly frictionless surface, and the puck glides long distances without slowing down. Additionally, if we know enough about the friction, we can accurately predict how quickly the object will slow down. Friction is an external force.

Newton’s first law is completely general and can be applied to anything from an object sliding on a table to a satellite in orbit to blood pumped from the heart. Experiments have thoroughly verified that any change in velocity (speed or direction) must be caused by an external force. The idea of *generally applicable or universal laws* is important not only here—it is a basic feature of all laws of physics. Identifying these laws is like recognizing patterns in nature from which further patterns can be discovered. The genius of Galileo, who first developed the idea for the first law, and Newton, who clarified it, was to ask the fundamental question, “What is the cause?” Thinking in terms of cause and effect is a worldview fundamentally different from the typical ancient Greek approach when questions such as “Why does a tiger have stripes?” would have been answered in Aristotelian fashion, “That is the nature of the beast.” True perhaps, but not a useful insight.

Mass

The property of a body to remain at rest or to remain in motion with constant velocity is called **inertia**. Newton’s first law is often called the **law of inertia**. As we know from experience, some objects have more inertia than others. It is obviously more difficult to change the motion of a large boulder than that of a basketball, for example. The inertia of an object is measured by its **mass**. Roughly speaking, mass is a measure of the amount of “stuff” (or matter) in something. The quantity or amount of matter in an object is determined by the numbers of atoms and molecules of various types it contains. Unlike weight, mass does not vary with location. The mass of an object is the same on Earth, in orbit, or on the surface of the Moon. In practice, it is very difficult to count and identify all of the atoms and molecules in an object, so masses are not often determined in this manner. Operationally, the masses of objects are determined by comparison with the standard kilogram.

Exercise 1.2.3.1

Which has more mass: a kilogram of cotton balls or a kilogram of gold?

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- **Newton's first law of motion** states that a body at rest remains at rest, or, if in motion, remains in motion at a constant velocity unless acted on by a net external force. This is also known as the **law of inertia**.
- **Inertia** is the tendency of an object to remain at rest or remain in motion. Inertia is related to an object's mass.
- **Mass** is the quantity of matter in a substance.

Answers

Exercise 1.2.3.1

Exercise 1.2.3.1

They are equal. A kilogram of one substance is equal in mass to a kilogram of another substance. The quantities that might differ between them are volume and density.

This page titled 1.2.3: Newton's First Law of Motion- Inertia is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.3: Newton's First Law of Motion- Inertia by Andrew Park is licensed CC BY 4.0.

1.2.4: Newton's Second Law of Motion- Force and Acceleration

Learning Objectives

- Define net force, external force, and system.
- Understand Newton's second law of motion.
- Apply Newton's second law to determine the weight of an object.

Newton's second law of motion is closely related to Newton's first law of motion. It mathematically states the cause and effect relationship between force and changes in motion. Newton's second law of motion is more quantitative and is used extensively to calculate what happens in situations involving a force. Before we can write down Newton's second law as a simple equation giving the exact relationship of force, mass, and acceleration, we need to sharpen some ideas that have already been mentioned.

First, what do we mean by a change in motion? The answer is that a change in motion is equivalent to a change in velocity. A change in velocity means, by definition, that there is an **acceleration**. Newton's first law says that a net external force causes a change in motion; thus, we see that a *net external force causes acceleration*.

Another question immediately arises. What do we mean by an external force? An intuitive notion of external is correct—an **external force** acts from outside the **system** of interest. For example, in Figure 1.2.4.1(a) the system of interest is the wagon plus the child in it. The two forces exerted by the other children are external forces. An internal force acts between elements of the system. Again looking at Figure 1.2.4.1(a), the force the child in the wagon exerts to hang onto the wagon is an internal force between elements of the system of interest. Only external forces affect the motion of a system, according to Newton's first law. (The internal forces actually cancel, as we shall see in the next section.) *You must define the boundaries of the system before you can determine which forces are external*. Sometimes the system is obvious, whereas other times identifying the boundaries of a system is more subtle. The concept of a system is fundamental to many areas of physics, as is the correct application of Newton's laws. This concept will be revisited many times on our journey through physics.

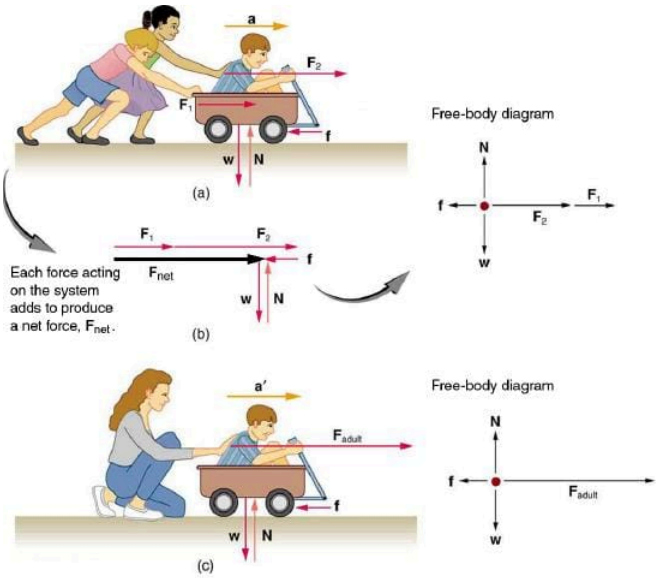


Figure 1.2.4.1: Different forces exerted on the same mass produce different accelerations. (a) Two children push a wagon with a child in it. Arrows representing all external forces are shown. The system of interest is the wagon and its rider. The weight w of the system and the support of the ground N are also shown for completeness and are assumed to cancel. The vector f represents the friction acting on the wagon, and it acts to the left, opposing the motion of the wagon. (b) All of the external forces acting on the system add together to produce a net force, F_{net} . The free-body diagram shows all of the forces acting on the system of interest. The dot represents the center of mass of the system. Each force vector extends from this dot. Because there are two forces acting to the right, we draw the vectors collinearly. (c) A larger net external force produces a larger acceleration ($|a'| > |a|$) when an adult pushes the child.

Figure 1.2.4.1 is our first example of a **free-body diagram**, which is a technique used to illustrate all the external forces acting on a body. The body is represented by a single isolated point (or free body), and only those forces acting on the body from the outside (external forces) are shown. Free-body diagrams are very useful in analyzing forces acting on a system and are employed extensively in the study and application of Newton's laws of motion.

Now, it seems reasonable that acceleration should be directly proportional to and in the same direction as the net (total) external force acting on a system. This assumption has been verified experimentally and is illustrated in Figure 1.2.4.1. In part (a), a smaller force causes a smaller acceleration than the larger force illustrated in part (c). For completeness, the vertical forces are also shown; they are assumed to cancel since there is no acceleration in the vertical direction. The vertical forces are the weight w and the support of the ground N, and the horizontal force f represents the force of friction. These will be discussed in more detail in later sections. For now, we will say that friction is a force that opposes the motion past each other of objects that are touching. Figure 1.2.4.1(b) shows how vectors representing the external forces add together to produce a net force, $\boldsymbol{F}_{\text{net}}$.

To obtain an equation for Newton's second law, we first write the relationship of acceleration and net external force as the proportionality

$$\boldsymbol{a} \propto \boldsymbol{F}_{\text{net}}$$

where the symbol $\propto$ means "proportional to," and $\boldsymbol{F}_{\text{net}}$ is the **net external force**, the vector sum of all external forces. This proportionality states what we have said in words—*acceleration is directly proportional to the net external force*. Once the system of interest is chosen, it is important to identify the external forces, so that the internal forces can be ignored (again, as we will see later, they exactly cancel out, allowing for a tremendous simplification).

Now, hopefully it also seems reasonable that acceleration should be inversely proportional to the mass of the system. That is, if the system has more "stuff", then for the same net external force applied, the acceleration is less. And indeed, as illustrated in Figure 1.2.4.2, the same net external force applied to a car produces a much smaller acceleration than when applied to a basketball. The proportionality is written as

$$a \propto \frac{1}{m}$$

where m is the mass of the system. In fact, this is how we will *define* mass. That is, for an object B to have double the mass of an object A, the same net external force applied on object B produces half the acceleration it does on object A.

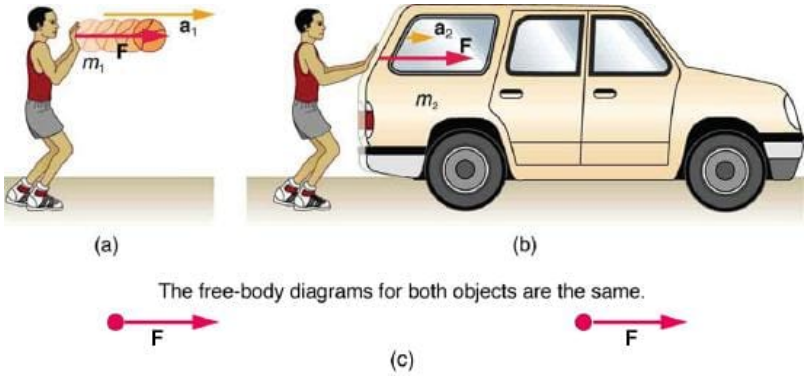


Figure 1.2.4.2: The same force exerted on systems of different masses produces different accelerations. (a) A basketball player pushes on a basketball to make a pass. (The effect of gravity on the ball is ignored.) (b) The same player exerts an identical force on a stalled SUV and produces a far smaller acceleration (even if friction is negligible). (c) The free-body diagrams are identical, permitting direct comparison of the two situations. A series of patterns for the free-body diagram will emerge as you do more problems.

As it turns out, the acceleration of an object depends only on the net external force and the mass of the object. Combining the two proportionalities just given yields Newton's second law of motion.

NEWTON'S SECOND LAW OF MOTION

The acceleration of a system is directly proportional to and in the same direction as the net external force acting on the system, and inversely proportional to its mass.

In equation form, Newton's second law of motion is

$$\boldsymbol{a} = \frac{\boldsymbol{F}_{\text{net}}}{m}$$

This is often written in the more familiar form

$$\boldsymbol{F}_{\text{net}} = m\boldsymbol{a},$$

with the vector notation indicating that the net external force is in the same direction as acceleration. When only the magnitude of force and acceleration are considered, this equation is simply (note the lack of vector notations)

$$F_{\text{net}} = ma.$$

Although these equations are really the same, the first gives more insight into what Newton's second law means. The law is a *cause and effect relationship* among three quantities. That is, acceleration is *caused* by a net external force, not vice versa, as the second equation might mistakenly imply.

Units of Force

$F_{\text{net}} = ma$ is used to define the units of force in terms of the three basic units for mass, length, and time. The SI unit of force is called the **newton** (abbreviated N) and 1 N is the force needed to accelerate a system of mass 1 kg at the rate of $1\ \text{m/s}^2$. Putting these together,

$$1\ \text{N} = 1\ \text{kg} \cdot \text{m/s}^2.$$

While almost the entire world uses the newton for the unit of force, in the United States the most familiar unit of force is the pound (lb), where 1 N = 0.225 lb.

Weight and the Gravitational Force

When an object is dropped, it accelerates toward the center of Earth. Newton's second law states that a net force on an object is responsible for its acceleration. If air resistance is negligible, the net force on a falling object is the gravitational force, commonly called its **weight**, w. Weight can be denoted as a vector w because it has a direction; *down* is, by definition, the direction of gravity, so w is directed downward. The magnitude of weight is denoted as w. Galileo was instrumental in showing that, in the absence of air resistance, all objects fall with the same acceleration g. Using Galileo's result and Newton's second law, we can derive an equation for the magnitude of weight.

Consider an object with mass m falling downward toward Earth. It experiences only the downward force of gravity, which has magnitude w. Newton's second law states that the magnitude of the net external force on an object is $F_{\text{net}} = ma$.

Since the object experiences only the downward force of gravity, $F_{\text{net}} = w$. We know that the acceleration of an object due to gravity is g, or $a = g$. Substituting these into Newton's second law gives

Definition: WEIGHT

This is the equation for *weight*—the gravitational force on a mass m:

$$w = mg.$$

Since $g = 9.80\ \text{m/s}^2$ on Earth, the weight of a 1.0 kg object on Earth is 9.8 N, as we see:

$$w = mg = (1.0\ \text{kg})\left(9.80\ \text{m/s}^2\right) = 9.8\ \text{N}.$$

When the net external force on an object is its weight, we say that it is in **free-fall**. That is, the only force acting on the object is the force of gravity. In the real world, when objects fall downward toward Earth, they are never truly in free-fall because there is

always some upward force from the air acting on the object.

The acceleration due to gravity g varies slightly over the surface of Earth, so that the weight of an object depends on location and is not an intrinsic property of the object. Weight varies dramatically if one leaves Earth. On the Moon, for example, the acceleration due to gravity is only $1.67\ \mathrm{m/s^2}$. A 1.0-kg mass thus has a weight of 9.8 N on Earth and only about 1.7 N on the Moon.

The broadest definition of weight in this sense is that *the weight of an object is the gravitational force on it from the nearest large body*, such as Earth, the Moon, the Sun, and so on. This is the most common and useful definition of weight in physics. It differs dramatically, however, from the definition of weight used by NASA and the popular media in relation to space travel and exploration. When they speak of "weightlessness" and "microgravity," they are really referring to the phenomenon we call "free-fall" in physics. We shall use the above definition of weight, and we will make careful distinctions between free-fall and actual weightlessness.

It is important to be aware that weight and mass are very different physical quantities, although they are closely related. Mass is the quantity of matter (how much "stuff") and does not vary, whereas weight is the gravitational force and does vary depending on gravity. It is tempting to equate the two, since most of our examples take place on Earth, where the weight of an object only varies a little with the location of the object. Furthermore, the terms *mass* and *weight* are used interchangeably in everyday language; for example, our medical records often show our "weight" in kilograms, but never in the correct units of newtons.

COMMON MISCONCEPTIONS: MASS VS. WEIGHT

Mass and weight are often used interchangeably in everyday language. However, in science, these terms are distinctly different from one another. Mass is a measure of how much matter is in an object. The typical measure of mass is the kilogram (or the "slug" in English units). Weight, on the other hand, is a measure of the force of gravity acting on an object. Weight is equal to the mass of an object (m) multiplied by the acceleration due to gravity (g). Like any other force, weight is measured in terms of newtons (or pounds in English units). This interchangeable use is why you might have heard an expression such as, "1 kilogram is 2.2 pounds." The *correct* statement is "1 kilogram mass weighs 2.2 pounds on Earth."

Assuming the mass of an object is kept intact, it will remain the same, regardless of its location. However, because weight depends on the acceleration due to gravity, the weight of an object *can change* when the object enters into a region with stronger or weaker gravity. For example, the acceleration due to gravity on the Moon is $1.67\ \mathrm{m/s^2}$ (which is much less than the acceleration due to gravity on Earth, $9.80\ \mathrm{m/s^2}$). If you measured your weight on Earth and then measured your weight on the Moon, you would find that you "weigh" much less, even though you do not look any skinnier. This is because the force of gravity is weaker on the Moon. In fact, when people say that they are "losing weight," they really mean that they are losing "mass" (which in turn causes them to weigh less).

TAKE-HOME EXPERIMENT: MASS AND WEIGHT

What do bathroom scales measure? When you stand on a bathroom scale, what happens to the scale? It depresses slightly. The scale contains springs that compress in proportion to your weight—similar to rubber bands expanding when pulled. The springs provide a measure of your weight (for an object which is not accelerating). This is a force in newtons (or pounds). In most countries, the measurement in newtons is divided by 9.80 to give a reading in mass units of kilograms. The scale measures weight but is calibrated to provide information about mass. While standing on a bathroom scale, push down on a table next to you. What happens to the reading? Why? Would your scale measure the same "mass" on Earth as on the Moon?

Example 1.2.4.1: What Acceleration Can a Person Produce when Pushing a Lawn Mower?

Suppose that the net external force (push minus friction) exerted on a lawn mower is 51 N (about 11 lb) parallel to the ground. The mass of the mower is 24 kg. What is its acceleration?



Figure 1.2.4.3: The net force on a lawn mower is 51 N to the right. At what rate does the lawn mower accelerate to the right?

Strategy

Since F_{net} and m are given, the acceleration can be calculated directly from Newton's second law as stated in $F_{\text{net}} = ma$.

Solution

The magnitude of the acceleration a is $a = \frac{F_{\text{net}}}{m}$. Entering known values gives

$$a = \frac{51 \text{ N}}{24 \text{ kg}}$$

Substituting the units $\text{kg} \cdot \text{m/s}^2$ for N yields

$$a = \frac{51 \text{ kg} \cdot \text{m/s}^2}{24 \text{ kg}} = 2.1 \text{ m/s}^2.$$

Discussion

The direction of the acceleration is the same direction as that of the net force, which is parallel to the ground. There is no information given in this example about the individual external forces acting on the system, but we can say something about their relative magnitudes. For example, the force exerted by the person pushing the mower forward must be greater than the friction opposing the motion (directed backward) to result in a forward net force, and the vertical forces must cancel if there is to be no acceleration in the vertical direction. The acceleration found is small enough to be reasonable for a person pushing a mower. When the person reaches the top speed, the acceleration will be zero (no change of velocity) and the force exerted by the person pushing the mower will be equal to the friction opposing the motion.

Example 1.2.4.2: What Rocket Thrust Accelerates This Sled?

Prior to manned space flights, rocket sleds were used to test aircraft, missile equipment, and physiological effects on human subjects at high speeds. They consisted of a platform that was mounted on one or two rails and propelled by several rockets. Calculate the magnitude of force exerted by each rocket, called its thrust T, for the four-rocket propulsion system shown in Figure 1.2.4.4. The sled's initial acceleration is 49 m/s^2 the mass of the system is 2100 kg, and the force of friction opposing the motion is known to be 650 N.

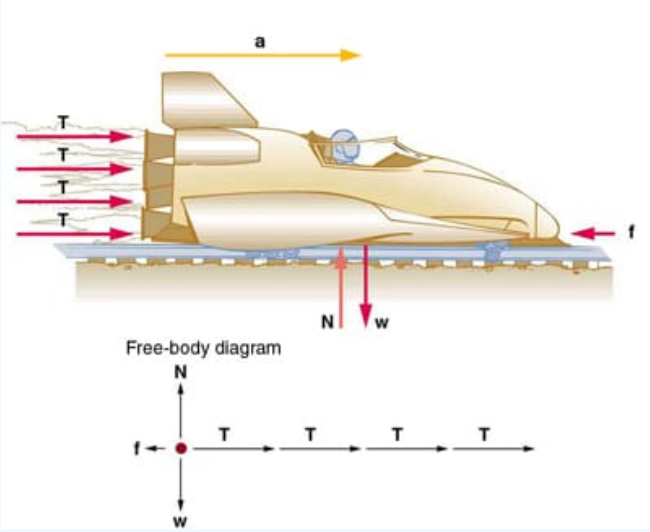


Figure 1.2.4.4: A sled experiences a rocket thrust that accelerates it to the right. Each rocket creates an identical thrust T. As in other situations where there is only horizontal acceleration, the vertical forces cancel. The ground exerts an upward force N on the system which balances out the weight force, w. The system here is the sled, its rockets, and rider, so none of the internal forces *between* these objects are considered. f represents the friction force (figure not drawn to scale).

Strategy

Although there are forces acting vertically and horizontally, we assume the vertical forces cancel since there is no vertical acceleration. This leaves us with only horizontal forces and a simpler one-dimensional problem. Directions are indicated with plus or minus signs, with right taken as the positive direction. See the free-body diagram in the figure.

Solution

Since acceleration, mass, and the force of friction are given, we start with Newton's second law and look for ways to find the thrust of the engines. Since we have defined the direction of the force and acceleration as acting "to the right," we need to consider only the magnitudes of these quantities in the calculations. Hence we begin with

$$F_{\text{net}} = ma,$$

where F_{net} is the net force along the horizontal direction. We can see from Figure 1.2.4.4 that the engine thrusts add, while friction opposes the thrust. In equation form, the net external force is

$$F_{\text{net}} = 4T - f.$$

Substituting this into Newton's second law gives

$$F_{\text{net}} = ma = 4T - f.$$

Using a little algebra, we solve for the total thrust 4*T*:

$$4T = ma + f.$$

Substituting known values yields

$$4T = ma + f = (2100 \text{ kg})\left(49 \text{ m/s}^2\right) + 650 \text{ N}.$$

So the total thrust is

$$4T = 1.04 \times 10^5 \text{ N},$$

and the individual thrusts are

$$T = \frac{1.04 \times 10^5 \text{ N}}{4} = 2.6 \times 10^4 \text{ N}.$$

Discussion

The numbers are quite large, so to put the result in perspective, note that acceleration of 49 m/s^2 is approximately 5 times the gravitational acceleration on Earth. It's a large acceleration, which requires a large net external force to produce.

Experiments such as this were performed in the early 1960s to test the limits of human endurance and the setup designed to protect human subjects in jet fighter emergency ejections. Speeds of 1000 km/h were obtained, with accelerations of 45 g's. (Recall that g, the acceleration due to gravity, is 9.80 m/s^2. When we say that an acceleration is g's, it is $45 \times 9.80 \text{ m/s}^2$, which is approximately 440 m/s^2. While living subjects are not used any more, land speeds of 10,000 km/h have been obtained with rocket sleds. In this example, as in the preceding one, the system of interest is obvious. We will see in later examples that choosing the system of interest is crucial—and the choice is not always obvious.

Newton's second law of motion gives a relationship among acceleration, force, and mass. It can help us make predictions. Each of those physical quantities can be defined independently, so the second law tells us something basic and universal about nature. The next section introduces the third and final law of motion.

Section Summary

- Acceleration, a, is defined as a rate of change in velocity, resulting from a change in the magnitude and/or the direction of the velocity.
- An external force is one acting on a system from outside the system, as opposed to internal forces, which act between components within the system.
- Newton's second law of motion states that the acceleration of a system is directly proportional to and in the same direction as the net external force acting on the system, and inversely proportional to its mass.
- In equation form, Newton's second law of motion is $\boldsymbol{a} = \frac{\boldsymbol{F}_{\text{net}}}{m}$, often written in the more familiar form: $\boldsymbol{F}_{\text{net}} = m\boldsymbol{a}$.
- The weight w of an object is defined as the force of gravity acting on an object of mass mm. Given acceleration due to gravity g, the magnitude of weight is:

$$w = mg$$

Glossary

acceleration
: the rate at which an object's velocity changes over a period of time

free-fall
: a situation in which the only force acting on an object is the force due to gravity

external force
: a force acting on an object or system that originates outside of the object or system

net external force
: the vector sum of all external forces acting on an object or system; causes a mass to accelerate

free-body diagram
: a sketch showing all of the external forces acting on an object or system; the system is represented by a dot, and the forces are represented by vectors extending outward from the dot

Newton's second law of motion
: The acceleration of a system is directly proportional to and in the same direction as the net external force acting on the system, and inversely proportional to its mass.

system
: the object or the group of objects under consideration

weight
: the force due to gravity; $w = mg$ for objects on Earth

This page titled 1.2.4: Newton's Second Law of Motion- Force and Acceleration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.4: Newton's Second Law of Motion- Force and Acceleration by Andrew Park is licensed CC BY 4.0.

1.2.5: Newton's Third Law of Motion- Symmetry in Forces

Learning Objectives

- Understand Newton's third law of motion.
- Apply Newton's third law to define systems and solve problems of motion.

Whenever one body exerts a force on another—the first also experiences a force (equal in magnitude and opposite in direction). Numerous common experiences, such as stubbing a toe or throwing a ball, confirm this. It is precisely stated in **Newton's third law of motion**.

NEWTON'S THIRD LAW OF MOTION

Whenever one body exerts a force on a second body, the first body experiences a force by the second body that is equal in magnitude and opposite in direction to the force that it itself exerts.

This law represents a certain *symmetry in nature*: Forces always occur in pairs, and one body cannot exert a force on another without experiencing a force itself. We sometimes refer to this law loosely as "action-reaction," where the force exerted is the action and the force experienced as a consequence is the reaction. Newton's third law has practical uses in analyzing the origin of forces and understanding which forces are external to a system.

We can readily see Newton's third law at work by taking a look at how people move about. Consider a swimmer pushing off from the side of a pool, as illustrated in Figure 1.2.5.1. She pushes against the pool wall with her feet and accelerates in the direction *opposite* to that of her push. The wall has exerted an equal and opposite force back on the swimmer. You might think that two equal and opposite forces would cancel, but they do not *because they act on different bodies*. In this case, there are two bodies to consider: the swimmer and the wall. If we define the swimmer as the system of interest, as in the figure, then $\boldsymbol{F}_{\text{wall on feet}}$ is an external force on this system and affects its motion. The swimmer moves in the direction of $\boldsymbol{F}_{\text{wall on feet}}$. In contrast, the force $\boldsymbol{F}_{\text{feet on wall}}$ acts on the wall and not on our system of interest. Thus $\boldsymbol{F}_{\text{feet on wall}}$ does not affect the motion of the system and does not counteract $\boldsymbol{F}_{\text{wall on feet}}$. Note that the swimmer pushes the wall in the direction opposite to that in which she wishes to move. The reaction force by the wall on her pushes her in the desired direction.

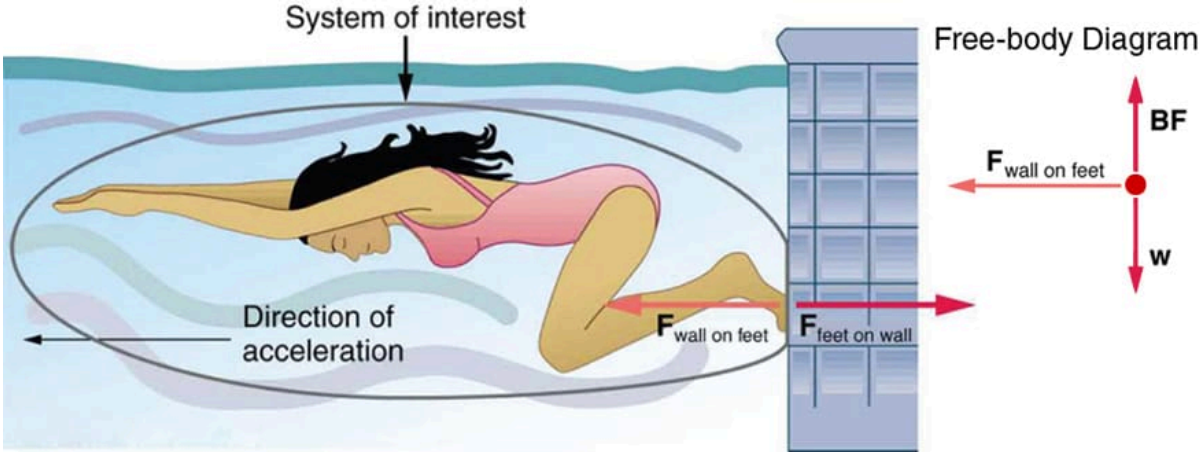


Figure 1.2.5.1: When the swimmer exerts a force $\boldsymbol{F}_{\text{feet on wall}}$ on the wall, she accelerates in the direction opposite to that of her push. This means the net external force on her is in the direction opposite to $\boldsymbol{F}_{\text{feet on wall}}$. This opposition occurs because, in accordance with Newton's third law of motion, the wall exerts a force $\boldsymbol{F}_{\text{wall on feet}}$ on her, equal in magnitude but in the direction opposite to the one she exerts on it. The line around the swimmer indicates the system of interest. Note that $\boldsymbol{F}_{\text{feet on wall}}$ does not act on this system (the swimmer) and, thus, does not cancel $\boldsymbol{F}_{\text{wall on feet}}$. Thus the free-body diagram shows only $\boldsymbol{F}_{\text{wall on feet}}$, w, the gravitational force, and **BF**, the buoyant force of the water supporting the swimmer's weight. The vertical forces w and **BF** cancel since there is no vertical motion.

Other examples of Newton's third law are easy to find. As a professor paces in front of a whiteboard, she exerts a force backward

on the floor. The floor exerts a reaction force forward on the professor that causes her to accelerate forward. Similarly, a car accelerates because the ground pushes forward on the drive wheels in reaction to the drive wheels pushing backward on the ground. You can see evidence of the wheels pushing backward when tires spin on a gravel road and throw rocks backward. In another example, rockets move forward by expelling gas backward at high velocity. This means the rocket exerts a large backward force on the gas in the rocket combustion chamber, and the gas therefore exerts a large reaction force forward on the rocket. This reaction force is called **thrust**. It is a common misconception that rockets propel themselves by pushing on the ground or on the air behind them. The rockets propel themselves by pushing their own exhaust gases, and that is how they work even in a vacuum. Helicopters create lift by pushing air down, thereby experiencing an upward reaction force. Birds and airplanes also fly by exerting force on air in a direction opposite to that of whatever force they need. For example, the wings of a bird force air downward and backward in order to get lift and move forward. An octopus propels itself in the water by ejecting water through a funnel from its body, similar to a jet ski. In a situation similar to Sancho's, professional cage fighters experience reaction forces when they punch, sometimes breaking their hand by hitting an opponent's body.

Example 1.2.5.1: Getting Up To Speed: Choosing the Correct System

A physics professor pushes a cart of demonstration equipment to a lecture hall, as seen in Figure 1.2.5.2. Her mass is 65.0 kg, the cart's is 12.0 kg, and the equipment's is 7.0 kg. Calculate the acceleration produced when the professor exerts a backward force of 150 N on the floor. All forces opposing the motion, such as friction on the cart's wheels and air resistance, total 24.0 N.

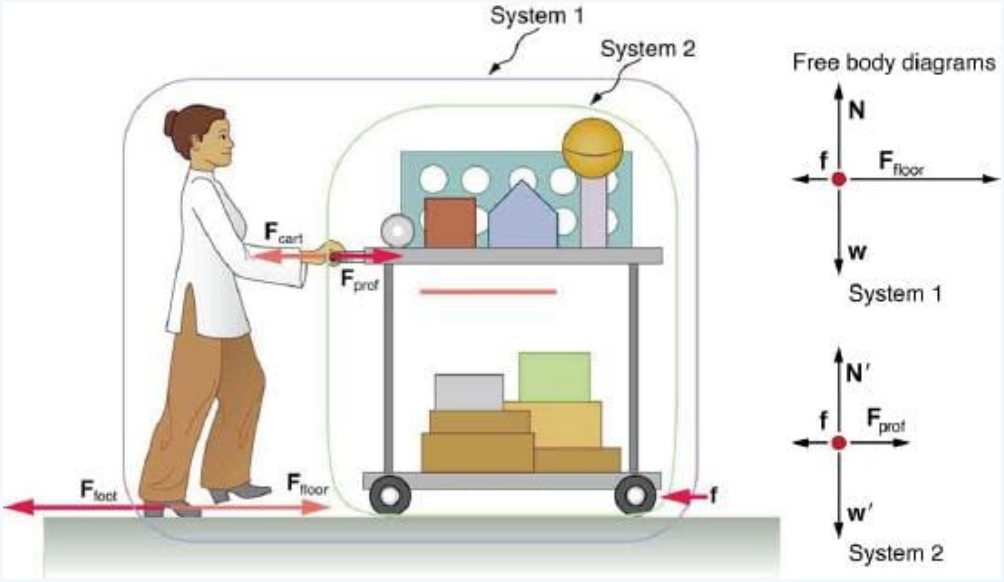


Figure 1.2.5.2: A professor pushes a cart of demonstration equipment. The lengths of the arrows are proportional to the magnitudes of the forces (except for **f**, since it is too small to draw to scale). Different questions are asked in each example; thus, the system of interest must be defined differently for each. System 1 is appropriate for this example, since it asks for the acceleration of the entire group of bodies. Only F_{floor} and f are external forces acting on System 1 along the line of motion. All other forces either cancel or act on the outside world. System 2 is appropriate for Example 1.2.5.2 so that F_{prof} will be an external force and enter into Newton's second law.

Strategy

Since they accelerate as a unit, we define the system to be the professor, cart, and equipment. This is System 1 in Figure 1.2.5.2. The professor pushes backward with a force F_{foot} of 150 N. According to Newton's third law, the floor exerts a forward reaction force F_{floor} of 150 N on System 1. Because all motion is horizontal, we can assume there is no net force in the vertical direction, and treat this as a one-dimensional problem in the horizontal direction. As noted, f opposes the motion and is thus in the opposite direction of F_{floor}. Note that we do not include the forces F_{prof} or F_{cart} because these are internal forces, and we do not include F_{foot} because it acts on the floor, not on the system. There are no other significant forces acting on System 1. If the net external force can be found from all this information, we can use Newton's second law to find the acceleration as requested. See the free-body diagram in the figure.

Solution

Newton's second law is given by

$$a = \frac{F_{\text{net}}}{m}.$$

The net external force on System 1 is deduced from Figure 1.2.5.2 and the discussion above to be

$$F_{\text{net}} = F_{\text{floor}} - f = 150 \text{ N} - 24.0 \text{ N} = 126 \text{ N}.$$

The mass of System 1 is

$$m = (65.0 + 12.0 + 7.0)\text{kg} = 84 \text{ kg}.$$

These values of F_{net} and m produce an acceleration of

$$a = \frac{F_{\text{net}}}{m},$$
$$a = \frac{126 \text{ N}}{84 \text{ kg}} = 1.5 \text{ m/s}^2$$

Discussion

None of the forces between components of System 1, such as between the professor's hands and the cart, contribute to the net external force because they are internal to System 1. Another way to look at this is to note that forces between components of a system cancel because they are equal in magnitude and opposite in direction. For example, the force exerted by the professor on the cart results in an equal and opposite force back on her. In this case both forces act on the same system and, therefore, cancel. Thus internal forces (between components of a system) cancel. Choosing System 1 was crucial to solving this problem.

Example 1.2.5.2: Force on the Cart—Choosing a New System

Calculate the force the professor exerts on the cart in Figure 1.2.5.2 using data from the previous example if needed.

Strategy

If we now define the system of interest to be the cart plus equipment (System 2 in Figure 1.2.5.2), then the net external force on System 2 is the force the professor exerts on the cart minus friction. The force she exerts on the cart, $\mathbf{F}_{\text{prof}}$, is an external force acting on System 2. $\mathbf{F}_{\text{prof}}$ was internal to System 1, but it is external to System 2 and will enter Newton's second law for System 2.

Solution

Newton's second law can be used to find $\mathbf{F}_{\text{prof}}$. Starting with

$$a = \frac{F_{\text{net}}}{m}$$

and noting that the magnitude of the net external force on System 2 is

$$F_{\text{net}} = F_{\text{prof}} - f,$$

we solve for $\mathbf{F}_{\text{prof}}$, the desired quantity:

$$F_{\text{prof}} = F_{\text{net}} + f.$$

The value of f is given, so we must calculate net F_{net}. That can be done since both the acceleration and mass of System 2 are known. Using Newton's second law we see that

$$F_{\text{net}} = ma,$$

where the mass of System 2 is 19.0 kg (m= 12.0 kg + 7.0 kg) and its acceleration was found to be $a = 1.5\ \mathrm{m/s^2}$ in the previous example. Thus,

$$F_{\mathrm{net}} = ma,$$
$$F_{\mathrm{net}} = (19.0\ \mathrm{kg})\left(1.5\ \mathrm{m/s^2}\right) = 29\ \mathrm{N}.$$

Now we can find the desired force:

$$F_{\mathrm{prof}} = F_{\mathrm{net}} + f,$$
$$F_{\mathrm{prof}} = 29\ \mathrm{N} + 24.0\ \mathrm{N} = 53\ \mathrm{N}.$$

Discussion

It is interesting that this force is significantly less than the 150-N force the professor exerted backward on the floor. Not all of that 150-N force is transmitted to the cart; some of it accelerates the professor.

The choice of a system is an important analytical step both in solving problems and in thoroughly understanding the physics of the situation (which is not necessarily the same thing).

COMMON MISCONCEPTION: REACTION VS. RESULT

The phrases "action" and "reaction" in reference to the pair of forces described by Newton's third law often leads to unfortunate misunderstandings regarding the nature of the reaction force. Because of the common English usage of the word "reaction" (root word: react)—for example, in film, "reaction shot" means a shot of actors' faces as they react to an event that just happened—many students intuitively guess that there is a cause-and-effect relationship between "action force" and "reaction force."

There is no such relationship. Newton's third law describes *symmetry* in forces. Action force and reaction force are symmetric—they occur simultaneously; they are two sides of one and the same interaction. Action force does not cause reaction force, no more than the head of a coin causes the tail of the coin (they merely occur together, always). Here's a quick check to see if you understood the pairing of action force and reaction force correctly: try swapping the labels. That is, if you originally called A "action force" and B "reaction force", try calling B "action force" and A "reaction force." Does this re-labeling make some sense? If not, chances are, you have mis-identified the pair of action force and reaction force. (Here, a little weirdness is O.K., but if it feels like complete nonsense to swap the labels—like reversal of a cause-and-effect relationship—that tells you you used "reaction" in the common English usage, not how the word is used in Newton's third law.)

Section Summary

- **Newton's third law of motion** represents a basic symmetry in nature. It states: Whenever one body exerts a force on a second body, the first body experiences a force that is equal in magnitude and opposite in direction to the force that the first body exerts.
- A **thrust** is a reaction force that pushes a body forward in response to a backward force. Rockets, airplanes, and cars are pushed forward by a thrust reaction force.

Glossary

Newton's third law of motion
: Whenever one body exerts a force on a second body, the first body experiences a force by the second body that is equal in magnitude and opposite in direction to the force that it itself exerts.

thrust
: a reaction force that pushes a body forward; rockets, airplanes, and cars are pushed forward by a thrust, a reaction force to propellants pushed backward

This page titled 1.2.5: Newton's Third Law of Motion- Symmetry in Forces is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.5: Newton's Third Law of Motion- Symmetry in Forces by Andrew Park is licensed CC BY 4.0.

1.2.6: Normal Force and Tension

Learning Objectives

- Define normal force.
- Explain the relationship between normal force and apparent weight.
- Describe properties of tension force.
- Apply Newton's laws of motion to analyze forces.

Forces are given many names, such as push, pull, thrust, lift, weight, friction, and tension. Traditionally, forces have been grouped into several categories and given names relating to their source, how they are transmitted, or their effects. The most important of these categories are discussed in this section, together with some interesting applications. Further examples of forces are discussed later in this text.

Normal Force

Weight (also called force of gravity) is a pervasive force that acts at all times and must be counteracted to keep an object from falling. You definitely notice that you must support the weight of a heavy object by pushing up on it when you hold it stationary, as illustrated in Figure 1.2.6.1(a). But how do inanimate objects like a table support the weight of a mass placed on them, such as shown in Figure 1.2.6.1(b)? When the bag of dog food is placed on the table, the table actually sags slightly under the load. This would be noticeable if the load were placed on a card table, but even rigid objects deform when a force is applied to them. Unless the object is deformed beyond its limit, it will exert a restoring force much like a deformed spring (or trampoline or diving board). The greater the deformation, the greater the restoring force. So when the load is placed on the table, the table sags until the restoring force becomes as large as the weight of the load. At this point the net external force on the load is zero. That is the situation when the load is stationary on the table. The table sags quickly, and the sag is slight so we do not notice it. But it is similar to the sagging of a trampoline when you climb onto it.

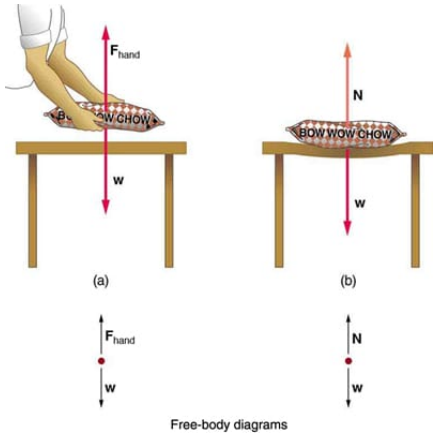


Figure 1.2.6.1: (a) The person holding the bag of dog food must supply an upward force F_{hand} equal in magnitude and opposite in direction to the weight of the food w. (b) The card table sags when the dog food is placed on it, much like a stiff trampoline. Elastic restoring forces in the table grow as it sags until they supply a force **N** equal in magnitude and opposite in direction to the weight of the load.

We must conclude that whatever supports a load, be it animate or not, must supply an upward force equal to the weight of the load, as we assumed in a few of the previous examples. The force supporting a load is perpendicular to the surface of contact between the load and its support, and this force is called a **normal force**, often indicated with symbol **N** (Please do not confuse this with the abbreviation for newton, the unit of force, N). The word *normal* means perpendicular to a surface. The normal force is not always equal to the object's weight, if there are other forces acting on the object, or if the object is accelerating, so that the net force is not zero.

Normal force also goes by another name, **apparent weight**. This is because all the forces you feel, especially your sensation of weight, are directly attributed to contact forces, such as the normal force. When NASA astronauts are weightlessly floating in the

International Space Station, their actual weight (gravitational force on them by the Earth) is not zero, but their apparent weight (sensation of weight due to contact forces) is zero.

Tension

A **tension** is a force along the length of a medium, especially a force carried by a flexible medium, such as a rope or cable. The word "tension" comes from a Latin word meaning "to stretch." Not coincidentally, the flexible cords that carry muscle forces to other parts of the body are called *tendons*. Any flexible connector, such as a string, rope, chain, wire, or cable, can exert pulls only parallel to its length; thus, a force carried by a flexible connector is a tension with direction parallel to the connector. It is important to understand that tension is a pull in a connector. In contrast, consider the phrase: "You can't push a rope." The tension force pulls outward along the two ends of a rope.

Consider a person holding a mass on a rope as shown in Figure 1.2.6.2.

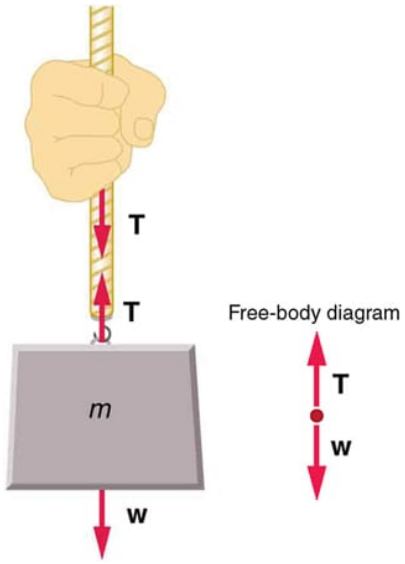


Figure 1.2.6.2: When a perfectly flexible connector (one requiring no force to bend it) such as this rope transmits a force ***T***, that force must be parallel to the length of the rope, as shown. The pull such a flexible connector exerts is a tension. Note that the rope pulls with equal force but in opposite directions on the hand and the supported mass (neglecting the weight of the rope). This is an example of Newton's third law. The rope is the medium that carries the equal and opposite forces between the two objects. The tension anywhere in the rope between the hand and the mass is equal. Once you have determined the tension in one location, you have determined the tension at all locations along the rope.

Tension in the rope must equal the weight of the supported mass, as we can prove using Newton's second law. If the 5.00-kg mass in the figure is stationary, then its acceleration is zero, and thus $\boldsymbol{F}_{\text{net}} = 0$. The only external forces acting on the mass are its weight w and the tension ***T*** supplied by the rope. Thus,

$$F_{\text{net}} = T - w = 0,$$

where ***T*** and w are the magnitudes of the tension and weight and their signs indicate direction, with up being positive here. Thus, just as you would expect, the tension equals the weight of the supported mass:

$$T = w = mg.$$

For a 5.00-kg mass, then (neglecting the mass of the rope) we see that

$$T = mg = (5.00\ \text{kg})\left(9.80\ \text{m/s}^2\right) = 49.0\ \text{N}.$$

If we cut the rope and insert a spring, the spring would extend a length corresponding to a force of 49.0 N, providing a direct observation and measure of the tension force in the rope.

Flexible connectors are often used to transmit forces around corners, such as in a hospital traction system, a finger joint, or a bicycle brake cable. If there is no friction, the tension is transmitted undiminished. Only its direction changes, and it is always parallel to the flexible connector. This is illustrated in Figure 1.2.6.3 (a) and (b).

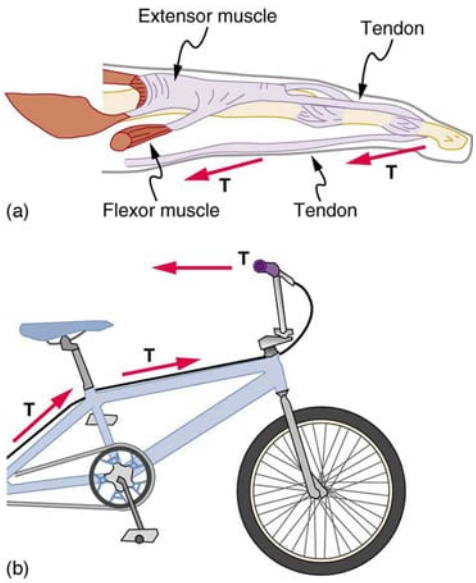


Figure 1.2.6.3: (a) Tendons in the finger carry force ***T*** from the muscles to other parts of the finger, usually changing the force's direction, but not its magnitude (the tendons are relatively friction free). (b) The brake cable on a bicycle carries the tension ***T*** from the handlebars to the brake mechanism. Again, the direction but not the magnitude of ***T*** is changed.



Figure 1.2.6.4: Unless an infinite tension is exerted, any flexible connector—such as the chain at the bottom of the picture—will sag under its own weight, giving a characteristic curve when the weight is evenly distributed along the length, known as a catenary curve. Suspension bridges—such as the Golden Gate Bridge shown in this image—are essentially very heavy flexible connectors. The weight of the bridge is evenly distributed along the length of flexible connectors, usually cables, which take on the characteristic shape. (credit: Leaflet, Wikimedia Commons)

Normal force and tension are examples of forces that are determined not by a specific formula but by enforcing a constraint on motion state of a body. In the example of the bag on a table above, the normal force is equal to weight, because that is the value needed to ensure that the acceleration of the bag is zero. What happens if the body experiences a non-zero acceleration? We will consider example below of a person standing on a scale (which measures his apparent weight) while riding in an elevator.

Example 1.2.6.1: What Does the Bathroom Scale Read in an Elevator?

Figure 1.2.6.5 shows a 75.0-kg man (weight of about 165 lb.) standing on a bathroom scale in an elevator. Calculate the scale reading: (a) if the elevator accelerates upward at a rate of $1.20\ \mathrm{m/s^2}$, and (b) if the elevator moves upward at a constant speed of 1 m/s.

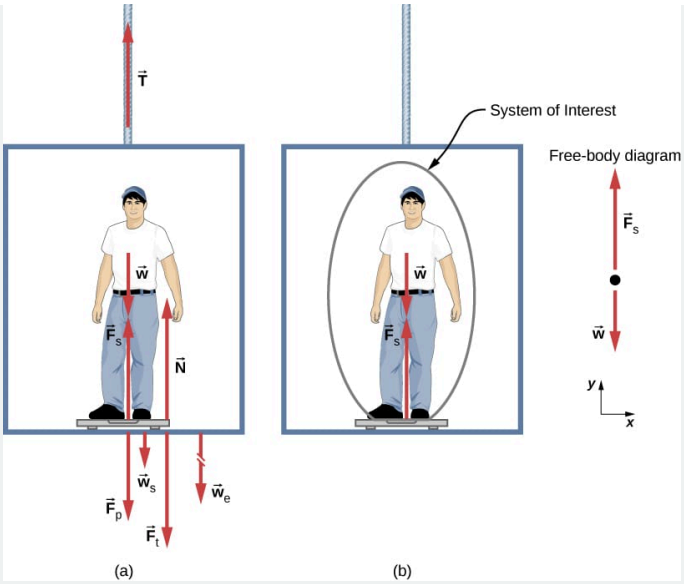


Figure 1.2.6.5: (a) The various forces acting when a person stands on a bathroom scale in an elevator. The arrows are approximately correct for when the elevator is accelerating upward—broken arrows represent forces too large to be drawn to scale. $\vec{\mathbf{T}}$ is the tension in the supporting cable, $\vec{\mathbf{w}}$ is the weight of the person, $\vec{\mathbf{w}}_{\mathrm{s}}$ is the weight of the scale, $\vec{\mathbf{w}}_{\mathrm{e}}$ is the weight of the elevator, $\vec{\mathbf{F}}_{\mathrm{s}}$ is the force of the scale on the person, $\vec{\mathbf{F}}_{\mathrm{P}}$ is the force of the person on the scale, $\vec{\mathbf{F}}_{\mathrm{t}}$ is the force of the scale on the floor of the elevator, and $\vec{\mathbf{N}}$ is the force of the floor upward on the scale. (b) The free-body diagram shows only the external forces acting *on* the designated system of interest—the person—and is the diagram we use for the solution of the problem.

Strategy

If the scale at rest is accurate, its reading equals $\vec{\mathbf{F}}_{\mathrm{P}}$, the magnitude of the force the person exerts downward on it. Figure 1.2.6.5(a) shows the numerous forces acting on the elevator, scale, and person. It makes this one-dimensional problem look much more formidable than if the person is chosen to be the system of interest and a free-body diagram is drawn, as in Figure 1.2.6.5(b). Analysis of the free-body diagram using Newton's laws can produce answers to both Figure 1.2.6.5(a) and (b) of this example, as well as some other questions that might arise. The only forces acting on the person are his weight $\vec{\mathbf{w}}$ and the upward force of the scale $\vec{\mathbf{F}}_{\mathrm{s}}$. According to Newton's third law, $\vec{\mathbf{F}}_{\mathrm{P}}$ and $\vec{\mathbf{F}}_{\mathrm{s}}$ are equal in magnitude and opposite in direction, so that we need to find F_{S} in order to find what the scale reads. We can do this, as usual, by applying Newton's second law,

$$\vec{\mathbf{F}}_{\mathrm{net}} = m\vec{\mathbf{a}}.$$

From the free-body diagram, we see that the net force $\vec{\mathbf{F}}_{\mathrm{net}}$ is sum of the scale force $\vec{\mathbf{F}}_{\mathrm{s}}$ and the weight $\vec{\mathbf{w}}$. Since the scale force and the weight are pointing in opposite directions, in terms of the magnitudes of the vectors,

$$F_{\mathrm{net}} = F_s - w = ma.$$

Solving for F_{S} gives us an equation with only one unknown:

$$F_{\mathrm{s}} = ma + w,$$

or, because $w = mg$, simply

$$F_{\mathrm{s}} = ma + mg.$$

No assumptions were made about the acceleration, so this solution should be valid for a variety of accelerations in addition to those in this situation. (*Note:* We are considering the case when the elevator is accelerating upward. If the elevator is accelerating downward, Newton's second law becomes $F_s - w = -ma$.)

Solution

a. We have $a = 1.20\ \text{m/s}^2$, so that

$$F_\text{s} = (75.0\ \text{kg})\left(9.80\ \text{m/s}^2\right) + (75.0\ \text{kg})\left(1.20\ \text{m/s}^2\right)$$

yielding

$$F_\text{s} = 825\ \text{N}.$$

b. Now, what happens when the elevator reaches a constant upward velocity? Will the scale still read more than his weight? For any constant velocity—up, down, or stationary—acceleration is zero because $a = \frac{\Delta v}{\Delta t}$ and $\Delta v = 0$. Thus,

$$F_\text{s} = ma + mg = 0 + mg$$

or

$$F_\text{s} = (75.0\ \text{kg})\left(9.80\ \text{m/s}^2\right),$$

which gives

$$F_\text{s} = 735\ \text{N}.$$

Significance

The scale reading in Figure 1.2.6.5(a) is about 185 lb. What would the scale have read if he were stationary? Since his acceleration would be zero, the force of the scale would be equal to his weight:

$$\begin{aligned} F_\text{net} &= ma = 0 = F_\text{s} - w \\ F_\text{s} &= w = mg \\ F_\text{s} &= (75.0\ \text{kg})\left(9.80\ \text{m/s}^2\right) = 735\ \text{N}. \end{aligned}$$

Thus, the scale reading in the elevator is greater than his 735-N (165-lb.) weight. This means that the scale is pushing up on the person with a force greater than his weight, as it must in order to accelerate him upward. Clearly, the greater the acceleration of the elevator, the greater the scale reading, consistent with what you feel in rapidly accelerating versus slowly accelerating elevators. In Figure 1.2.6.5(b), the scale reading is 735 N, which equals the person's weight. This is the case whenever the elevator has a constant velocity—moving up, moving down, or stationary.

Section Summary

- When objects rest on a surface, the surface applies a force to the object that supports the weight of the object. This supporting force acts perpendicular to and away from the surface. It is called a normal force, and it also goes by the name "apparent weight."
- *When objects rest on a non-accelerating horizontal surface*, the magnitude of the normal force is equal to the weight of the object:

$$N = mg.$$

- The pulling force that acts along a stretched flexible connector, such as a rope or cable, is called tension, ***T***. When a rope supports the weight of an object that is at rest, the tension in the rope is equal to the weight of the object:

$$T = mg.$$

Glossary

normal force
: the force that a surface applies to an object to support the weight of the object; acts perpendicular to the surface on which the object rests

apparent weight
: sensation of weight due to contact forces; usually the same as normal force

tension
: the pulling force that acts along a medium, especially a stretched flexible connector, such as a rope or cable; when a rope supports the weight of an object, the force on the object due to the rope is called a tension force

This page titled 1.2.6: Normal Force and Tension is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.6: Normal Force and Tension by Andrew Park is licensed CC BY 4.0.

1.2.7: Spring Force- Hooke's Law

Learning Objectives

- Describe the restoring force and displacement.
- Explain oscillatory motion under a spring force.

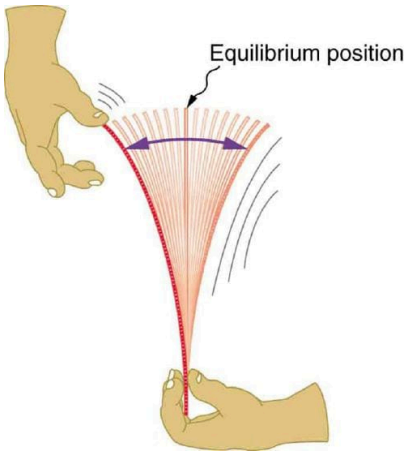


Figure 1.2.7.1: When displaced from its vertical equilibrium position, this plastic ruler oscillates back and forth because of the restoring force opposing displacement. When the ruler is on the left, there is a force to the right, and vice versa.

Newton's first law implies that an object oscillating back and forth is experiencing forces. Without force, the object would move in a straight line at a constant speed rather than oscillate. Consider, for example, plucking a plastic ruler to the left as shown in Figure 1.2.7.1. The deformation of the ruler creates a force in the opposite direction, known as a **restoring force**. Once released, the restoring force causes the ruler to move back toward its stable equilibrium position, where the net force on it is zero. However, by the time the ruler gets there, it gains momentum and continues to move to the right, producing the opposite deformation. It is then forced to the left, back through equilibrium, and the process is repeated until dissipative forces dampen the motion. These forces remove mechanical energy from the system, gradually reducing the motion until the ruler comes to rest.

The simplest oscillations occur when the restoring force is directly proportional to displacement. This is called Hooke's law force, or spring force:

$$F = -kx.$$

Here, F is the restoring force, x is the displacement from equilibrium or **deformation**, and k is a constant related to the difficulty in deforming the system. The minus sign indicates the restoring force is in the direction opposite to the displacement.

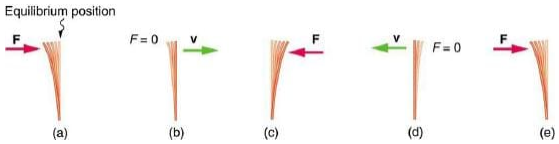


Figure 1.2.7.2: (a) The plastic ruler has been released, and the restoring force is returning the ruler to its equilibrium position. (b) The net force is zero at the equilibrium position, but the ruler has momentum and continues to move to the right. (c) The restoring force is in the opposite direction. It stops the ruler and moves it back toward equilibrium again. (d) Now the ruler has momentum to the left. (e) In the absence of damping (caused by frictional forces), the ruler reaches its original position. From there, the motion will repeat itself.

The **spring constant** k is related to the rigidity (or stiffness) of a system—the larger the spring constant, the greater the restoring force, and the stiffer the system. The units of k are newtons per meter (N/m). Figure 1.2.7.3 shows a graph of the absolute value of the restoring force versus the displacement for a system that can be described by Hooke's law—a simple spring in this case. The slope of the graph equals the spring constant k in newtons per meter.

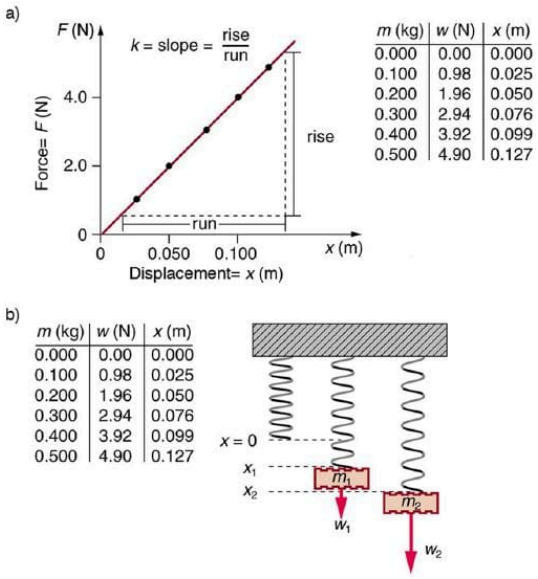


Figure 1.2.7.3: (a) A graph of absolute value of the restoring force versus displacement is displayed. The fact that the graph is a straight line means that the system obeys Hooke's law. The slope of the graph is the spring constant k. (b) The data in the graph were generated by measuring the displacement of a spring from equilibrium while supporting various weights. The restoring force equals the weight supported, if the mass is stationary.

Example 1.2.7.1: How Stiff Are Car Springs?



Figure 1.2.7.4: The mass of a car increases due to the introduction of a passenger. This affects the displacement of the car on its suspension system. (credit: exfordy on Flickr)

What is the spring constant for the suspension system of a car that settles 1.20 cm when an 80.0-kg person gets in?

Strategy

Consider the car to be in its equilibrium position $x = 0$ before the person gets in. The car then settles down 1.20 cm, which means it is displaced to a position $x = -1.20 \times 10^{-2}$ m. At that point, the springs supply a restoring force F equal to the person's weight $w = mg = (80.0 \text{ kg})(9.80 \text{ m/s}^2) = 784$ N. We take this force to be F in Hooke's law. Knowing F and x, we can then solve the spring constant k.

Solution

1. Solve Hooke's law, $F = -kx$, for k:

$$k = -\frac{F}{x}.$$

Substitute known values and solve k:

$$\begin{aligned} k &= -\frac{784\ \text{N}}{-1.20 \times 10^{-2}\ \text{m}} \\ &= 6.53 \times 10^{4}\ \text{N/m}. \end{aligned}$$

Discussion

Note that F and x have opposite signs because they are in opposite directions—the restoring force is up, and the displacement is down. Also, note that the car would oscillate up and down when the person got in if it were not for damping (due to frictional forces) provided by shock absorbers. Bouncing cars are a sure sign of bad shock absorbers.

Section Summary

- An oscillation is a back and forth motion of an object between two points of deformation.
- An oscillation may create a wave, which is a disturbance that propagates from where it was created.
- The simplest type of oscillations and waves are related to systems that can be described by Hooke's law:

$$F = -kx,$$

where F is the restoring force, x is the displacement from equilibrium or deformation, and k is the spring constant of the system.

Glossary

deformation
: displacement from equilibrium

spring constant
: a constant related to the rigidity of a system: the larger the spring constant, the more rigid the system; the spring constant is represented by *k*

restoring force
: force acting in opposition to the force caused by a deformation

This page titled 1.2.7: Spring Force- Hooke's Law is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.7: Spring Force- Hooke's Law by Andrew Park is licensed CC BY 4.0.

1.2.8: Friction

Learning Objectives

- Discuss the general characteristics of friction.
- Describe the various types of friction.
- Explain effect of friction on motion.

Friction is a force that is around us all the time that opposes relative motion between systems in contact but also allows us to move (which you have discovered if you have ever tried to walk on ice). While a common force, the behavior of friction is actually very complicated and is still not completely understood. We have to rely heavily on observations for whatever understandings we can gain. However, we can still deal with its more elementary general characteristics and understand the circumstances in which it behaves.

Definition: FRICTION

Friction is a force that opposes relative motion between systems in contact.

One of the simpler characteristics of friction is that it is parallel to the contact surface between systems and always in a direction that opposes motion or attempted motion of the systems relative to each other. If two systems are in contact and moving relative to one another, then the friction between them is called **kinetic friction**. For example, friction slows a hockey puck sliding on ice. But when objects are stationary, **static friction** can act between them; the static friction is usually greater than the kinetic friction between the objects.

Definition: KINETIC FRICTION

If two systems are in contact and moving relative to one another, then the friction between them is called kinetic friction.

Imagine, for example, trying to slide a heavy crate across a concrete floor—you may push harder and harder on the crate and not move it at all. This means that the static friction responds to what you do—it increases to be equal to and in the opposite direction of your push. But if you finally push hard enough, the crate seems to slip suddenly and starts to move. Once in motion it is easier to keep it in motion than it was to get it started, indicating that the kinetic friction force is less than the static friction force. If you add mass to the crate, say by placing a box on top of it, you need to push even harder to get it started and also to keep it moving. Furthermore, if you oiled the concrete you would find it to be easier to get the crate started and keep it going (as you might expect).

Figure 1.2.8.1 is a crude pictorial representation of how friction occurs at the interface between two objects. Close-up inspection of these surfaces shows them to be rough. So when you push to get an object moving (in this case, a crate), you must raise the object until it can skip along with just the tips of the surface hitting, break off the points, or do both. A considerable force can be resisted by friction with no apparent motion. The harder the surfaces are pushed together (such as if another box is placed on the crate), the more force is needed to move them. Part of the friction is due to adhesive forces between the surface molecules of the two objects, which explain the dependence of friction on the nature of the substances. Adhesion varies with substances in contact and is a complicated aspect of surface physics. Once an object is moving, there are fewer points of contact (fewer molecules adhering), so less force is required to keep the object moving. At small but nonzero speeds, friction is nearly independent of speed.

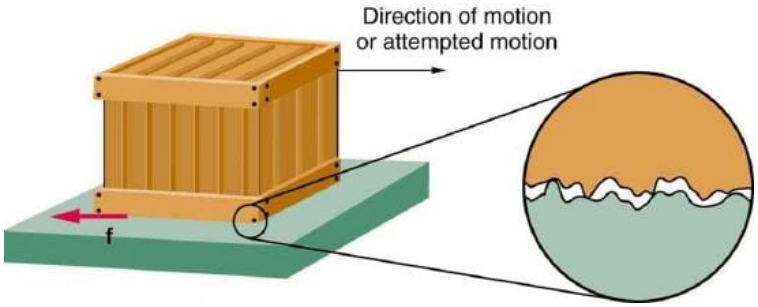


Figure 1.2.8.1: Frictional forces, such as f, always oppose motion or attempted motion between objects in contact. Friction arises in part because of the roughness of the surfaces in contact, as seen in the expanded view. In order for the object to move, it must rise to where the peaks can skip along the bottom surface. Thus a force is required just to set the object in motion. Some of the peaks will be broken off, also requiring a force to maintain motion. Much of the friction is actually due to attractive forces between molecules making up the two objects, so that even perfectly smooth surfaces are not friction-free. Such adhesive forces also depend on the substances the surfaces are made of, explaining, for example, why rubber-soled shoes slip less than those with leather soles.

The magnitude of the frictional force has two forms: one for static situations (static friction), the other for when there is motion (kinetic friction).

When there is no motion between the objects, the **magnitude of static friction $\mathbf{f_s}$** is

$$f_\mathrm{s} \leq \mu_\mathrm{s} N,$$

where μ_s is the coefficient of static friction and ***N*** is the magnitude of the normal force (the force perpendicular to the surface).

MAGNITUDE OF STATIC FRICTION

Magnitude of static friction f_s is

$$f_\mathrm{s} \leq \mu_\mathrm{s} N,$$

where μ_s is the coefficient of static friction and N is the magnitude of the normal force.

The symbol ≤ means *less than or equal to*, implying that static friction can have a minimum and a maximum value of $\mu_\mathrm{s} N$. Static friction is a responsive force that increases to be equal and opposite to whatever force is exerted, up to its maximum limit. Once the applied force exceeds $f_{\mathrm{s(max)}}$, the object will move. Thus

$$f_{\mathrm{s(max)}} = \mu_\mathrm{s} N.$$

Once an object is moving, the **magnitude of kinetic friction** f_k is given by

$$\mathrm{f_k} = \mu_\mathrm{k} N,$$

where μ_k is the coefficient of kinetic friction. A system in which $\mathrm{f_k} = \mu_\mathrm{k} N$ is described as a system in which *friction behaves simply*.

MAGNITUDE OF KINETIC FRICTION

The magnitude of kinetic friction f_k is given by

$$f_\mathrm{k} = \mu_\mathrm{k} N,$$

where μ_k is the coefficient of kinetic friction.

As seen in Table 1.2.8.1, the coefficients of kinetic friction are less than their static counterparts. That values of μ in Table 1.2.8.1 are stated to only one or, at most, two digits is an indication of the approximate description of friction given by the above two equations.

Table 1.2.8.1: Coefficients of Static and Kinetic Friction

System	Static friction μ_s	Kinetic friction μ_k
Rubber on dry concrete	1.00	0.70
Rubber on wet concrete	0.70	0.50
Wood on wood	0.50	0.30
Waxed wood on wet snow	0.14	0.10
Metal on wood	0.50	0.30
Steel on steel (dry)	0.60	0.30
Steel on steel (oiled)	0.05	0.03
Teflon on steel	0.04	0.04
Bone lubricated by synovial fluid	0.016	0.015
Shoes on wood	0.90	0.70
Shoes on ice	0.10	0.05
Ice on ice	0.10	0.03
Steel on ice	0.40	0.02

The equations given earlier include the dependence of friction on materials and the normal force. The direction of friction is always opposite that of motion, parallel to the surface between objects, and perpendicular to the normal force. For example, if the crate you try to push (with a force parallel to the floor) has a mass of 100 kg, then the normal force would be equal to its weight, $W = mg = (100\ \text{kg})\,(9.80\ \text{m/s}^2) = 980\ \text{N}$, perpendicular to the floor. If the coefficient of static friction is 0.45, you would have to exert a force parallel to the floor greater than $f_{s(\max)} = \mu_s N = (0.45)(980\ \text{N}) = 440\ \text{N}$ to move the crate. Once there is motion, friction is less and the coefficient of kinetic friction might be 0.30, so that a force of only 290 N ($f_k = \mu_k N = (0.30)(980\ \text{N}) = 290\ \text{N}$) would keep it moving at a constant speed. If the floor is lubricated, both coefficients are considerably less than they would be without lubrication. Coefficient of friction is a unit less quantity with a magnitude usually between 0 and 1.0. The coefficient of the friction depends on the two surfaces that are in contact.

TAKE-HOME EXPERIMENT

Find a small plastic object (such as a food container) and slide it on a kitchen table by giving it a gentle tap. Now spray water on the table, simulating a light shower of rain. What happens now when you give the object the same-sized tap? Now add a few drops of (vegetable or olive) oil on the surface of the water and give the same tap. What happens now? This latter situation is particularly important for drivers to note, especially after a light rain shower. Why?

Many people have experienced the slipperiness of walking on ice. However, many parts of the body, especially the joints, have much smaller coefficients of friction—often three or four times less than ice. A joint is formed by the ends of two bones, which are connected by thick tissues. The knee joint is formed by the lower leg bone (the tibia) and the thighbone (the femur). The hip is a ball (at the end of the femur) and socket (part of the pelvis) joint. The ends of the bones in the joint are covered by cartilage, which provides a smooth, almost glassy surface. The joints also produce a fluid (synovial fluid) that reduces friction and wear. A damaged or arthritic joint can be replaced by an artificial joint (Figure 1.2.8.2). These replacements can be made of metals (stainless steel or titanium) or plastic (polyethylene), also with very small coefficients of friction.



Figure 1.2.8.2: Artificial knee replacement is a procedure that has been performed for more than 20 years. In this figure, we see the post-op X-rays of the right knee joint replacement. (credit: Mike Baird, Flickr)

Other natural lubricants include saliva produced in our mouths to aid in the swallowing process, and the slippery mucus found between organs in the body, allowing them to move freely past each other during heartbeats, during breathing, and when a person moves. Artificial lubricants are also common in hospitals and doctor's clinics. For example, when ultrasonic imaging is carried out, the gel that couples the transducer to the skin also serves to lubricate the surface between the transducer and the skin—thereby reducing the coefficient of friction between the two surfaces. This allows the transducer to move freely over the skin.

We have discussed that when an object rests on a horizontal surface, there is a normal force supporting it equal in magnitude to its weight. Furthermore, simple friction is always proportional to the normal force.

MAKING CONNECTIONS: SUBMICROSCOPIC EXPLANATIONS OF FRICTION

The simpler aspects of friction dealt with so far are its macroscopic (large-scale) characteristics. Great strides have been made in the atomic-scale explanation of friction during the past several decades. Researchers are finding that the atomic nature of friction seems to have several fundamental characteristics. These characteristics not only explain some of the simpler aspects of friction—they also hold the potential for the development of nearly friction-free environments that could save hundreds of billions of dollars in energy which is currently being converted (unnecessarily) to heat.

Figure 1.2.8.3 illustrates one macroscopic characteristic of friction that is explained by microscopic (small-scale) research. We have noted that friction is proportional to the normal force, but not to the area in contact, a somewhat counterintuitive notion. When two rough surfaces are in contact, the actual contact area is a tiny fraction of the total area since only high spots touch. When a greater normal force is exerted, the actual contact area increases, and it is found that the friction is proportional to this area.

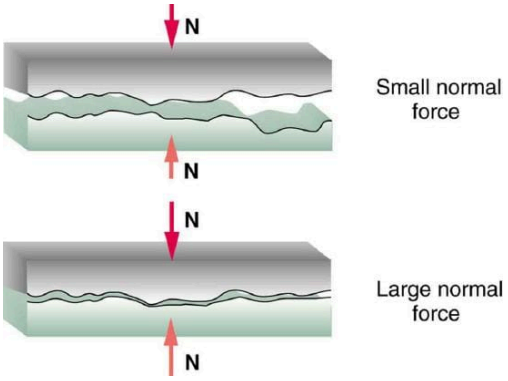


Figure 1.2.8.3: Two rough surfaces in contact have a much smaller area of actual contact than their total area. When there is a greater normal force as a result of a greater applied force, the area of actual contact increases as does friction.

But the atomic-scale view promises to explain far more than the simpler features of friction. The mechanism for how heat is generated is now being determined. In other words, why do surfaces get warmer when rubbed? Essentially, atoms are linked with one another to form lattices. When surfaces rub, the surface atoms adhere and cause atomic lattices to vibrate—essentially creating sound waves that penetrate the material. The sound waves diminish with distance and their energy is converted into heat. Chemical reactions that are related to frictional wear can also occur between atoms and molecules on the surfaces. Figure 1.2.8.4 shows how the tip of a probe drawn across another material is deformed by atomic-scale friction. The force needed to drag the tip can be measured and is found to be related to shear stress, which will be discussed later in this chapter. The variation in shear stress is remarkable (more than a factor of 10^{12}) and difficult to predict theoretically, but shear stress is yielding a fundamental understanding of a large-scale phenomenon known since ancient times—friction.

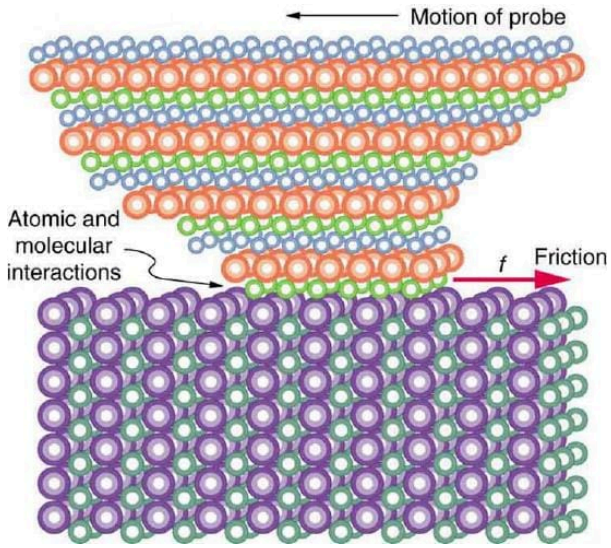


Figure 1.2.8.4: The tip of a probe is deformed sideways by frictional force as the probe is dragged across a surface. Measurements of how the force varies for different materials are yielding fundamental insights into the atomic nature of friction.

Section Summary

- Friction is a contact force between systems that opposes the motion or attempted motion between them. Simple friction is proportional to the normal force ***N*** pushing the systems together. (A normal force is always perpendicular to the contact surface between systems.) Friction depends on both of the materials involved. The magnitude of static friction f_{s} between systems stationary relative to one another is given by

$$f_{\mathrm{s}} \leq \mu_{\mathrm{s}} N,$$

  where μ_{s} is the coefficient of static friction, which depends on both of the materials.
- The kinetic friction force f_{k} between systems moving relative to one another is given by

$$f_{\mathrm{k}} = \mu_{\mathrm{k}} N,$$

where μ_{k} is the coefficient of kinetic friction, which also depends on both materials.

Glossary

friction
: a force that opposes relative motion or attempts at motion between systems in contact

kinetic friction
: a force that opposes the motion of two systems that are in contact and moving relative to one another

static friction
: a force that opposes the motion of two systems that are in contact and are not moving relative to one another

magnitude of static friction
: $f_{\mathrm{s}} \leq \mu_{\mathrm{s}} N$, where μ_{s} is the coefficient of static friction and N is the magnitude of the normal force

magnitude of kinetic friction
: $f_{\mathrm{k}} = \mu_{\mathrm{k}} N$, where μ_{k} is the coefficient of kinetic friction

This page titled 1.2.8: Friction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.8: Friction by Andrew Park is licensed CC BY 4.0.

1.2.9: Newton's Universal Law of Gravitation

Learning Objectives

- Explain Earth's gravitational force.
- Describe the gravitational effect of the Moon on Earth.
- Explain sensation of weightlessness in space.

What do aching feet, a falling apple, and the orbit of the Moon have in common? Each is caused by the gravitational force. Our feet are strained by supporting our weight—the force of Earth's gravity on us. An apple falls from a tree because of the same force acting a few meters above Earth's surface. And the Moon orbits Earth because gravity is able to supply the necessary centripetal force at a distance of hundreds of millions of meters. In fact, the same force causes planets to orbit the Sun, stars to orbit the center of the galaxy, and galaxies to cluster together. Gravity is another example of underlying simplicity in nature. It is the weakest of the four basic forces found in nature, and in some ways the least understood. It is a force that acts at a distance, without physical contact, and is expressed by a formula that is valid everywhere in the universe, for masses and distances that vary from the tiny to the immense.

Sir Isaac Newton was the first scientist to precisely define the gravitational force, and to show that it could explain both falling bodies and astronomical motions. See Figure 1.2.9.1. But Newton was not the first to suspect that the same force caused both our weight and the motion of planets. His forerunner Galileo Galilei had contended that falling bodies and planetary motions had the same cause. Some of Newton's contemporaries, such as Robert Hooke, Christopher Wren, and Edmund Halley, had also made some progress toward understanding gravitation. But Newton was the first to propose an exact mathematical form and to use that form to show that the motion of heavenly bodies should be conic sections—circles, ellipses, parabolas, and hyperbolas. This theoretical prediction was a major triumph—it had been known for some time that moons, planets, and comets follow such paths, but no one had been able to propose a mechanism that caused them to follow these paths and not others.

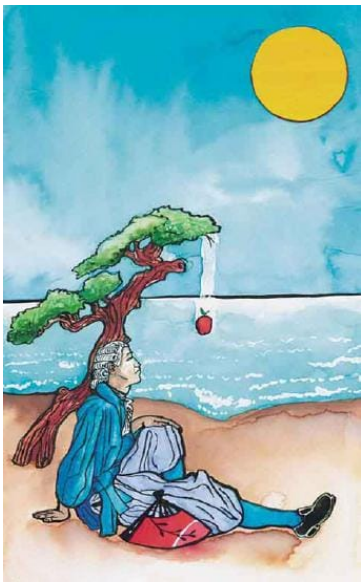


Figure 1.2.9.1: According to early accounts, Newton was inspired to make the connection between falling bodies and astronomical motions when he saw an apple fall from a tree and realized that if the gravitational force could extend above the ground to a tree, it might also reach the Sun. The inspiration of Newton's apple is a part of worldwide folklore and may even be based in fact. Great importance is attached to it because Newton's universal law of gravitation and his laws of motion answered very old questions about nature and gave tremendous support to the notion of underlying simplicity and unity in nature. Scientists still expect underlying simplicity to emerge from their ongoing inquiries into nature.

The gravitational force is relatively simple. It is always attractive, and it depends only on the masses involved and the distance between them. Stated in modern language, **Newton's universal law of gravitation** states that every particle in the universe attracts every other particle with a force along a line joining them. The force is directly proportional to the product of their masses

and inversely proportional to the square of the distance between them.

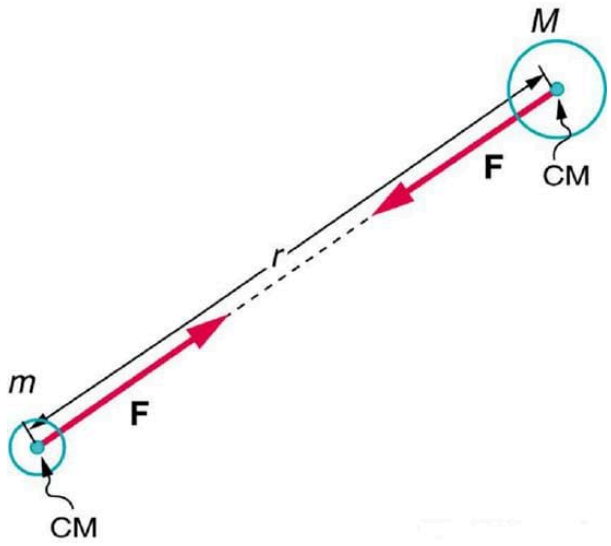


Figure 1.2.9.2: Gravitational attraction is along a line joining the centers of mass of these two bodies. The magnitude of the force is the same on each, consistent with Newton's third law.

MISCONCEPTION ALERT

The magnitude of the force on each object (one has larger mass than the other) is the same, consistent with Newton's third law.

The bodies we are dealing with tend to be large. To simplify the situation we assume that the body acts as if its entire mass is concentrated at one specific point called the **center of mass** (CM). For two bodies having masses m and M with a distance r between their centers of mass, the equation for Newton's universal law of gravitation is

$$F = G\frac{mM}{r^2},$$

where F is the magnitude of the gravitational force and G is a proportionality factor called the **gravitational constant**. G is a universal gravitational constant—that is, it is thought to be the same everywhere in the universe. It has been measured experimentally to be

$$G = 6.673 \times 10^{-11}\frac{\text{N} \cdot \text{m}^2}{\text{kg}^2}$$

in SI units. Note that the units of G are such that a force in newtons is obtained from $F = G\frac{mM}{r^2}$, when considering masses in kilograms and distance in meters. For example, two 1.000 kg masses separated by 1.000 m will experience a gravitational attraction of 6.673×10^{-11} N. This is an extraordinarily small force. The small magnitude of the gravitational force is consistent with everyday experience. We are unaware that even large objects like mountains exert gravitational forces on us. In fact, our body weight is the force of attraction of the *entire Earth* on us with a mass of 5.98×10^{24} kg.

Recall that the acceleration due to gravity g is about $9.80\ \text{m/s}^2$ on Earth. We can now determine why this is so. The weight of an object *mg* is the gravitational force between it and Earth. Substituting *mg* for F in Newton's universal law of gravitation gives

$$mg = G\frac{mM}{r^2},$$

where M is the mass of the object, M is the mass of Earth, and r is the distance to the center of Earth (the distance between the centers of mass of the object and Earth). See Figure 1.2.9.3. The mass M of the object cancels, leaving an equation for g:

$$g = G\frac{M}{r^2}.$$

Substituting known values for Earth's mass and radius (to three significant figures),

$$g = \left(6.67 \times 10^{-11} \frac{\text{N} \cdot \text{m}^2}{\text{kg}^2}\right) \times \frac{5.98 \times 10^{24} \text{ kg}}{\left(6.38 \times 10^{6} \text{ m}\right)^2},$$

and we obtain a value for the acceleration of a falling body:

$$g = 9.80 \text{ m/s}^2.$$

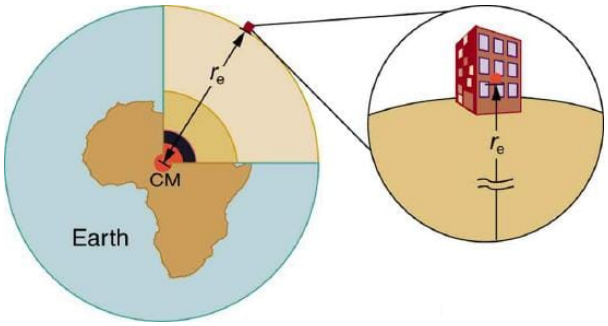


Figure 1.2.9.3: The distance between the centers of mass of Earth and an object on its surface is very nearly the same as the radius of Earth, because Earth is so much larger than the object.

This is the expected value *and is independent of the body's mass*. Newton's law of gravitation takes Galileo's observation that all masses fall with the same acceleration a step further, explaining the observation in terms of a force that causes objects to fall—in fact, in terms of a universally existing force of attraction between masses.

TAKE-HOME EXPERIMENT

Take a marble, a ball, and a spoon and drop them from the same height. Do they hit the floor at the same time? If you drop a piece of paper as well, does it behave like the other objects? Explain your observations.

Example 1.2.9.1: Earth's Gravitational Force on a Mass

(a) Determine the weight of a 5.00 kg rock when on Earth's surface.

(b) Determine the weight of a 5.00 kg rock when 3620 km above the surface of Earth.

Strategy for (a)

Use acceleration due to gravity near Earth's surface and Newton's second law.

Solution for (a)

$$F = mg = 5.00 \text{ kg} \times 9.80 \text{ m/s}^2 = 49.0 \text{ N}$$

Strategy for (b)

Use Newton's universal law of gravitation. Remember that distance is from the center of Earth. In this case the distance is 6380 km + 3620 km = 10,000 km = 1.00×10^7 m.

Solution for (b)

$$F = G\frac{mM}{r^2} = \left(6.673 \times 10^{-11} \frac{\text{N} \cdot \text{m}^2}{\text{kg}^2}\right) \frac{(5.00 \text{ kg})\left(5.98 \times 10^{24} \text{ kg}\right)}{\left(1.00 \times 10^{7} \text{ m}\right)^2} = 20.0 \text{ N}$$

Tides

Ocean tides are one very observable result of the Moon's gravity acting on Earth. Figure 1.2.9.4 is a simplified drawing of the Moon's position relative to the tides. Because water easily flows on Earth's surface, a high tide is created on the side of Earth nearest to the Moon, where the Moon's gravitational pull is strongest. Why is there also a high tide on the opposite side of Earth? The answer is that Earth is pulled toward the Moon more than the water on the far side, because Earth is closer to the Moon. So the water on the side of Earth closest to the Moon is pulled away from Earth, and Earth is pulled away from water on the far side. As Earth rotates, the tidal bulge (an effect of the tidal forces between an orbiting natural satellite and the primary planet that it orbits) keeps its orientation with the Moon. Thus there are two tides per day (the actual tidal period is about 12 hours and 25.2 minutes), because the Moon moves in its orbit each day as well).



Figure 1.2.9.4: The Moon causes ocean tides by attracting the water on the near side more than Earth, and by attracting Earth more than the water on the far side. The distances and sizes are not to scale. For this simplified representation of the Earth-Moon system, there are two high and two low tides per day at any location, because Earth rotates under the tidal bulge.

The Sun also affects tides, although it has about half the effect of the Moon. However, the largest tides, called spring tides, occur when Earth, the Moon, and the Sun are aligned. The smallest tides, called neap tides, occur when the Sun is at a 90° angle to the Earth-Moon alignment.

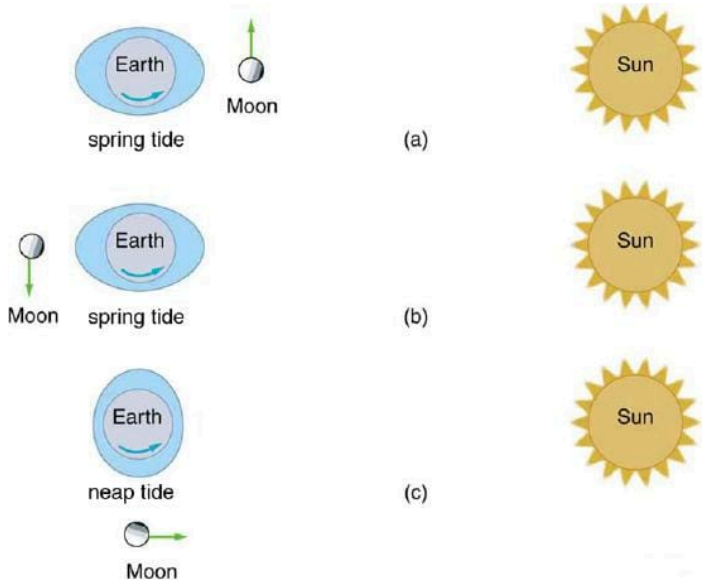


Figure 1.2.9.5: (a, b) Spring tides: The highest tides occur when Earth, the Moon, and the Sun are aligned. (c) Neap tide: The lowest tides occur when the Sun lies at 90° to the Earth-Moon alignment. Note that this figure is not drawn to scale.

Tides are not unique to Earth but occur in many astronomical systems. The most extreme tides occur where the gravitational force is the strongest and varies most rapidly, such as near black holes (see Figure 1.2.9.6). A few likely candidates for black holes have been observed in our galaxy. These have masses greater than the Sun but have diameters only a few kilometers across. The tidal forces near them are so great that they can actually tear matter from a companion star.

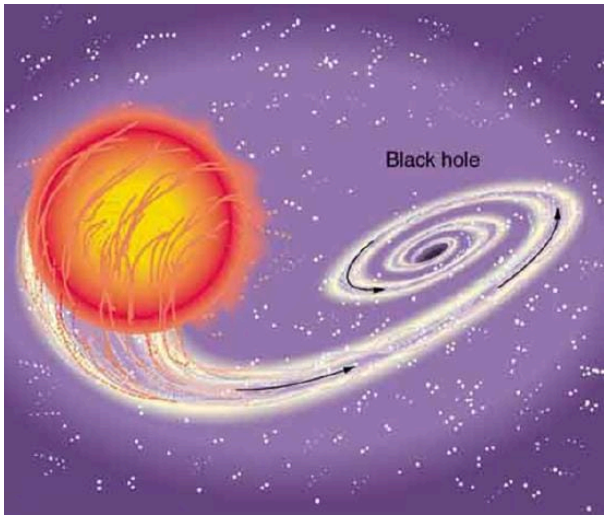


Figure 1.2.9.6: A black hole is an object with such strong gravity that not even light can escape it. This black hole was created by the supernova of one star in a two-star system. The tidal forces created by the black hole are so great that it tears matter from the companion star. This matter is compressed and heated as it is sucked into the black hole, creating light and X-rays observable from Earth.

”Weightlessness” and Microgravity

You may have seen images of astronauts in International Space Station, floating in their environment. What is the effect of “weightlessness” upon an astronaut who is in orbit for months? Or what about the effect of weightlessness upon plant growth? Weightlessness doesn’t mean that an astronaut is not being acted upon by the gravitational force. There is no “zero gravity” in an astronaut’s orbit. The term just means that the astronaut is in free-fall, accelerating with the acceleration due to gravity. If an elevator cable breaks, the passengers inside will be in free fall and will experience weightlessness. You can experience short periods of weightlessness in some rides in amusement parks, and the commercial “zero-G” airplane rides simulate this experience of weightlessness about 20 seconds at a time.



Figure 1.2.9.7: Astronauts experiencing weightlessness on board the International Space Station. (credit: NASA)

Microgravity refers to an environment in which the acceleration of a body due to non-gravitational forces is small compared with that produced by Earth on its surface. Many interesting biology and physics topics have been studied over the past three decades in the presence of microgravity. Of immediate concern is the effect on astronauts of extended times in outer space, such as at the International Space Station. Researchers have observed that muscles will atrophy (waste away) in this environment. There is also a corresponding loss of bone mass. Study continues on cardiovascular adaptation to space flight. On Earth, blood pressure is usually higher in the feet than in the head, because the higher column of blood exerts a downward force on it, due to gravity. When standing, 70% of your blood is below the level of the heart, while in a horizontal position, just the opposite occurs. What difference does the absence of this pressure differential have upon the heart?

Some findings in human physiology in space can be clinically important to the management of diseases back on Earth. On a

somewhat negative note, spaceflight is known to affect the human immune system, possibly making the crew members more vulnerable to infectious diseases. Experiments flown in space also have shown that some bacteria grow faster in microgravity than they do on Earth. However, on a positive note, studies indicate that microbial antibiotic production can increase by a factor of two in space-grown cultures. One hopes to be able to understand these mechanisms so that similar successes can be achieved on the ground. In another area of physics space research, inorganic crystals and protein crystals have been grown in outer space that have much higher quality than any grown on Earth, so crystallography studies on their structure can yield much better results.

Plants have evolved with the stimulus of gravity and with gravity sensors. Roots grow downward and shoots grow upward. Plants might be able to provide a life support system for long duration space missions by regenerating the atmosphere, purifying water, and producing food. Some studies have indicated that plant growth and development are not affected by gravity, but there is still uncertainty about structural changes in plants grown in a microgravity environment.

Section Summary

- Newton's universal law of gravitation: Every particle in the universe attracts every other particle with a force along a line joining them. The force is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. In equation form, this is

$$F = G\frac{mM}{r^2},$$

  where F is the magnitude of the gravitational force. G is the gravitational constant, given by $G = 6.673 \times 10^{-11}\ \mathrm{N \cdot m^2/kg^2}$.

- Newton's law of gravitation applies universally.

Glossary

gravitational constant, *G*
: a proportionality factor used in the equation for Newton's universal law of gravitation; it is a universal constant—that is, it is thought to be the same everywhere in the universe

center of mass
: the point where the entire mass of an object can be thought to be concentrated

microgravity
: an environment in which the acceleration of a body due to non-gravitational forces is small compared with that produced by Earth on its surface

Newton's universal law of gravitation
: every particle in the universe attracts every other particle with a force along a line joining them; the force is directly proportional to the product of their masses and inversely proportional to the square of the distance between them

This page titled 1.2.9: Newton's Universal Law of Gravitation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.9: Newton's Universal Law of Gravitation by Andrew Park is licensed CC BY 4.0.

1.2.10: Centripetal Force

Learning Objectives

- Explain the role of centripetal force in a uniform circular motion.
- Calculate centripetal force and acceleration for simple situations.

Any force or combination of forces can cause a centripetal or radial acceleration. Just a few examples are the tension in the rope on a tether ball, the force of Earth's gravity on the Moon, friction between roller skates and a rink floor, a banked roadway's force on a car, and forces on the tube of a spinning centrifuge.

Any net force causing uniform circular motion is called a **centripetal force**. The direction of a centripetal force is toward the center of curvature, the same as the direction of centripetal acceleration. According to Newton's second law of motion, net force is mass times acceleration: $F_{\text{net}} = ma$. For uniform circular motion, the acceleration is the centripetal acceleration— $a = a_c$. Thus, the magnitude of centripetal force F_c is

$$F_c = ma_c.$$

By using the expressions for centripetal acceleration a_c from $a_c = \frac{v^2}{r}$, we get an expression for the centripetal force $\mathrm{F_c}$ in terms of mass, velocity, and radius of curvature:

$$F_c = m\frac{v^2}{r}.$$

Note that if you solve the first expression for r, you get

$$r = \frac{mv^2}{F_c}.$$

This implies that for a given mass and velocity, a large centripetal force causes a small radius of curvature—that is, a tight curve.

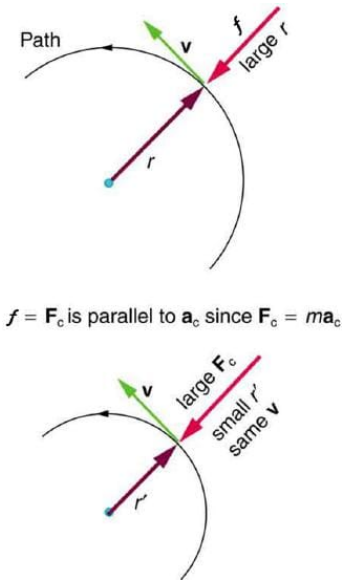


Figure 1.2.10.1: The frictional force supplies the centripetal force and is numerically equal to it. Centripetal force is perpendicular to velocity and causes uniform circular motion. The larger the F_c, the smaller the radius of curvature r and the sharper the curve. The second curve has the same v, but a larger F_c produces a smaller r'.

Example 1.2.10.1: What Coefficient of Friction Do Car Tires Need on a Flat Curve?

(a) Calculate the centripetal force exerted on a 900 kg car that negotiates a 500 m radius curve at 25.0 m/s.

(b) Assuming an unbanked curve, find the minimum static coefficient of friction, between the tires and the road, static friction being the reason that keeps the car from slipping (see Figure 1.2.10.2).

Strategy and Solution for (a)

We know that $F_c = \frac{mv^2}{r}$. Thus,

$$F_c = \frac{mv^2}{r} = \frac{(900\ \text{kg})(25.0\ \text{m/s})^2}{(500\ \text{m})} = 1125\ \text{N}.$$

Strategy for (b)

Figure 1.2.10.2 shows the forces acting on the car on an unbanked (level ground) curve. Friction is to the left, keeping the car from slipping, and because it is the only horizontal force acting on the car, the friction is the centripetal force in this case. We know that the maximum static friction (at which the tires roll but do not slip) is $\mu_s N$, where μ_s is the static coefficient of friction and N is the normal force. The normal force equals the car's weight on level ground, so that $N = mg$. Thus the centripetal force in this situation is

$$F_c = f = \mu_s N = \mu_s mg.$$

Now we have a relationship between centripetal force and the coefficient of friction. Using the expression for $\mathrm{F_c}$

$$F_c = m\frac{v^2}{r},$$
$$m\frac{v^2}{r} = \mu_s mg.$$

We solve this for μ_s, noting that mass cancels, and obtain

$$\mu_s = \frac{v^2}{rg}.$$

Solution for (b)

Substituting the knowns,

$$\mu_s = \frac{(25.0\ \text{m/s})^2}{(500\ \text{m})\,(9.80\ \text{m/s}^2)} = 0.13.$$

(Because coefficients of friction are approximate, the answer is given to only two digits.)

Discussion

The coefficient of friction found in part (b) is much smaller than is typically found between tires and roads. The car will still negotiate the curve if the coefficient is greater than 0.13, because static friction is a responsive force, being able to assume a value less than but no more than $\mu_s N$. A higher coefficient would also allow the car to negotiate the curve at a higher speed, but if the coefficient of friction is less, the safe speed would be less than 25 m/s. Note that mass cancels, implying that in this example, it does not matter how heavily loaded the car is to negotiate the turn. Mass cancels because friction is assumed proportional to the normal force, which in turn is proportional to mass. If the surface of the road were banked, the normal force would be less as will be discussed below.

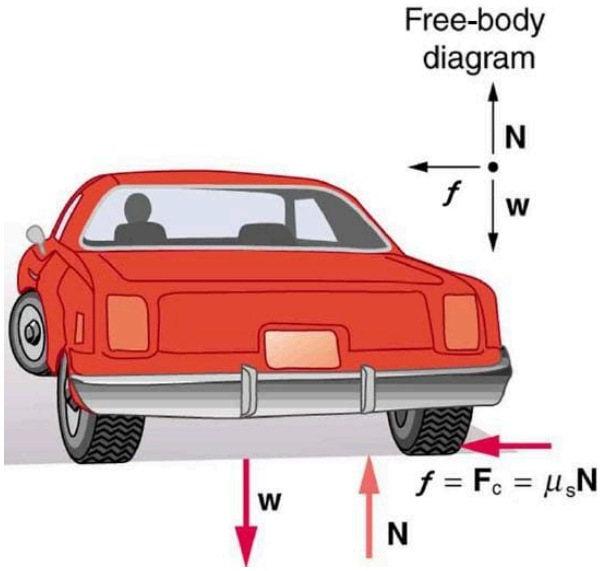


Figure 1.2.10.2: This car on level ground is moving away and turning to the left. The centripetal force causing the car to turn in a circular path is due to friction between the tires and the road. A minimum coefficient of friction is needed, or the car will move in a larger-radius curve and leave the roadway.

In the case of **banked curves**, where the slope of the road helps you negotiate the curve, some or all of the necessary centripetal force is provided by the normal force. See Figure 1.2.10.3. The greater the angle θ, the faster you can take the curve. Race tracks for bikes as well as cars, for example, often have steeply banked curves. In an "ideally banked curve," the angle θ is such that you can negotiate the curve at a certain speed without the aid of friction between the tires and the road. Conceptually, for **ideal banking**, the net external force equals the horizontal centripetal force in the absence of friction. The components of the normal force N in the horizontal and vertical directions must equal the centripetal force and the weight of the car, respectively.

Figure 1.2.10.3 shows a free body diagram for a car on a frictionless banked curve. If the angle θ is ideal for the speed and radius, then the net external force will equal the necessary centripetal force. The only two external forces acting on the car are its weight w and the normal force of the road N. (A frictionless surface can only exert a force perpendicular to the surface—that is, a normal force.) These two forces must add to give a net external force that is horizontal toward the center of curvature and has magnitude mv^2/r. We omit detailed calculations, which require trigonometry.

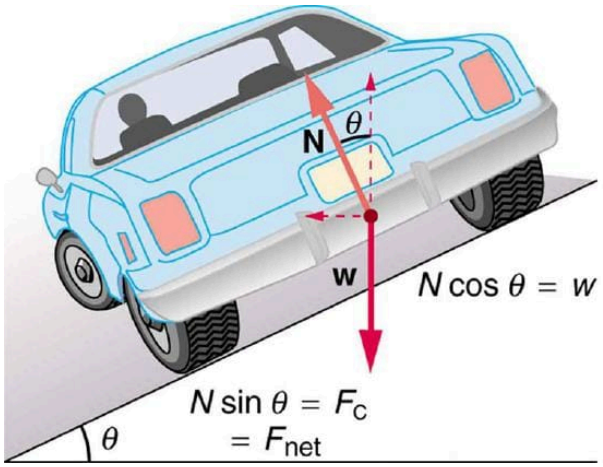


Figure 1.2.10.3: The car on this banked curve is moving away and turning to the left.

Section Summary

- Centripetal force $\mathrm{F_c}$ is any force causing uniform circular motion. It is a "center-seeking" force that always points toward the center of rotation. It is perpendicular to linear velocity v and has magnitude

$$F_c = ma_c,$$

which can also be expressed as

$$F_c = m\frac{v^2}{r}.$$

Glossary

centripetal force
: any net force causing uniform circular motion

ideal banking
: the sloping of a curve in a road, where the angle of the slope allows the vehicle to negotiate the curve at a certain speed without the aid of friction between the tires and the road; the net external force on the vehicle equals the horizontal centripetal force in the absence of friction

banked curve
: the curve in a road that is sloping in a manner that helps a vehicle negotiate the curve

This page titled 1.2.10: Centripetal Force is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.10: Centripetal Force by Andrew Park is licensed CC BY 4.0.

1.2.11: Dynamics (Exercise)

Conceptual Questions

2.1: Development of Force Concept

1. Propose a force standard different from the example of a stretched spring discussed in the text. Your standard must be capable of producing the same force repeatedly.

2. What properties do forces have that allow us to classify them as vectors?

2.2: Newton's First Law of Motion- Inertia

3. How are inertia and mass related?

4. What is the relationship between weight and mass? Which is an intrinsic, unchanging property of a body?

2.3: Newton's Second Law of Motion- Force and Acceleration

5. Which statement is correct? (a) Net force causes motion. (b) Net force causes change in motion. Explain your answer and give an example.

6. Why can we neglect forces such as those holding a body together when we apply Newton's second law of motion?

7. Explain how the choice of the "system of interest" affects which forces must be considered when applying Newton's second law of motion.

8. Describe a situation in which the net external force on a system is not zero, yet its speed remains constant.

9. A system can have a nonzero velocity while the net external force on it *is* zero. Describe such a situation.

10. A rock is thrown straight up. What is the net external force acting on the rock when it is at the top of its trajectory?

11. (a) Give an example of different net external forces acting on the same system to produce different accelerations. (b) Give an example of the same net external force acting on systems of different masses, producing different accelerations. (c) What law accurately describes both effects? State it in words and as an equation.

12. If the acceleration of a system is zero, are no external forces acting on it? What about internal forces? Explain your answers.

13. If a constant, nonzero force is applied to an object, what can you say about the velocity and acceleration of the object?

14. The gravitational force on the basketball in Figure 1.2.11.1 is ignored. When gravity *is* taken into account, what is the direction of the net external force on the basketball—above horizontal, below horizontal, or still horizontal?

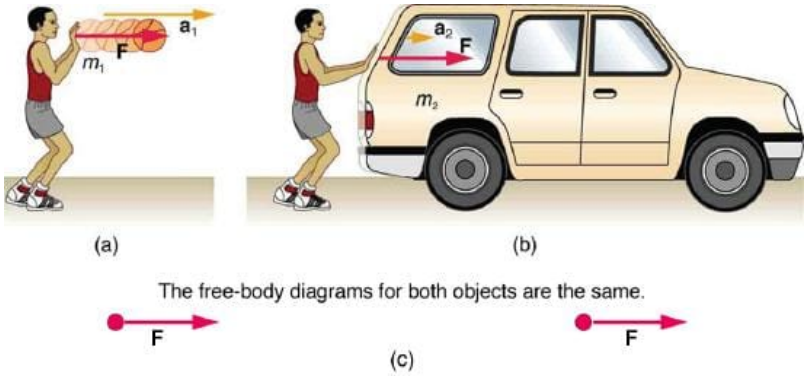


Figure 1.2.11.1: The same force exerted on systems of different masses produces different accelerations. (a) A basketball player pushes on a basketball to make a pass. (The effect of gravity on the ball is ignored.) (b) The same player exerts an identical force on a stalled SUV and produces a far smaller acceleration (even if friction is negligible). (c) The free-body diagrams are identical, permitting direct comparison of the two situations. A series of patterns for the free-body diagram will emerge as you do more problems.

2.4: Newton's Third Law of Motion- Symmetry in Forces

15. When you take off in a jet aircraft, there is a sensation of being pushed back into the seat. Explain why you move backward in the seat—is there really a force backward on you? (The same reasoning explains whiplash injuries, in which the head is apparently thrown backward.)

16. A device used since the 1940s to measure the kick or recoil of the body due to heart beats is the "ballistocardiograph." What physics principle(s) are involved here to measure the force of cardiac contraction? How might we construct such a device?

17. Describe a situation in which one system exerts a force on another and, as a consequence, experiences a force that is equal in magnitude and opposite in direction. Which of Newton's laws of motion apply?

18. Why does an ordinary rifle recoil (kick backward) when fired? The barrel of a recoilless rifle is open at both ends. Describe how Newton's third law applies when one is fired. Can you safely stand close behind one when it is fired?

19. An American football lineman reasons that it is senseless to try to out-push the opposing player, since no matter how hard he pushes he will experience an equal and opposite force from the other player. Use Newton's laws and draw a free-body diagram of an appropriate system to explain how he can still out-push the opposition if he is strong enough.

20. Newton's third law of motion tells us that forces always occur in pairs of equal and opposite magnitude. Explain how the choice of the "system of interest" affects whether one such pair of forces cancels.

2.5: Normal Force and Tension

21. Find at least one example of each case from everyday experiences (find an example that is not in the textbook!): (a) a situation where normal force on a person is equal to the person's weight, (b) a situation where normal force on a person is greater than the person's weight, and (c) a situation where normal force on a person is less than the person's weight. Also, can you find an example (that is not in the textbook!) which covers each of all three cases, depending on changing circumstances?

22. As noted in the text, the normal force and the tension force can only act in one direction: the normal force can only push, and the tension force can only pull. The numerical value of the force in these given directions is determined by what force is necessary to keep a body in the observed state of motion (either not accelerating, in presence of other forces, or accelerating at a particular rate). Under certain conditions, the normal force would go to zero (but not negative, as direction can't reverse), as would the tension force.

23. Explain qualitatively what you would see as: (a) the normal force goes to zero, and (b) the tension force goes to zero.

2.6: Spring Force- Hooke's Law

24. Describe a system which undergoes an oscillation under a Hooke's law force.

25. Imagine an object under a hypothetical force behaving opposite to Hooke's law force. That is, instead of restoring force ($F = -kx$, with the negative sign indicating that the force acts in the opposite direction of displacement), the force reinforces an initial displacement: $F = +kx$. What would happen to this object if it is placed at the equilibrium ($x = 0$) and then very slightly disturbed?

2.7: Friction

26. Define normal force. What is its relationship to friction when friction behaves simply?

27. The glue on a piece of tape can exert forces. Can these forces be a type of simple friction? Explain, considering especially that tape can stick to vertical walls and even to ceilings.

28. When you learn to drive, you discover that you need to let up slightly on the brake pedal as you come to a stop or the car will stop with a jerk. Explain this in terms of the relationship between static and kinetic friction.

29. When you push a piece of chalk across a chalkboard, it sometimes screeches because it rapidly alternates between slipping and sticking to the board. Describe this process in more detail, in particular explaining how it is related to the fact that kinetic friction is less than static friction. (The same slip-grab process occurs when tires screech on pavement.)

2.8: Newton's Universal Law of Gravitation

30. Action at a distance, such as is the case for gravity, was once thought to be illogical and therefore untrue. What is the ultimate determinant of the truth in physics, and why was this action ultimately accepted?

31. Two friends are having a conversation. Anna says a satellite in orbit is in freefall because the satellite keeps falling toward Earth. Tom says a satellite in orbit is not in freefall because the acceleration due to gravity is not $9.80\ \mathrm{m/s^2}$. Who do you agree with and why?

32. Newton's laws of motion and gravity were among the first to convincingly demonstrate the underlying simplicity and unity in nature. Many other examples have since been discovered, and we now expect to find such underlying order in complex situations. Is there proof that such order will always be found in new explorations?

2.9: Centripetal Force

33. If you wish to reduce the stress (which is related to centripetal force) on high-speed tires, would you use large- or small-diameter tires? Explain.

34. Define centripetal force. Can any type of force (for example, tension, gravitational force, friction, and so on) be a centripetal force? Can any combination of forces be a centripetal force?

35. If centripetal force is directed toward the center, why do you feel that you are 'thrown' away from the center as a car goes around a curve? Explain.

36. Race car drivers routinely cut corners as shown in Figure 1.2.11.2. Explain how this allows the curve to be taken at the greatest speed.

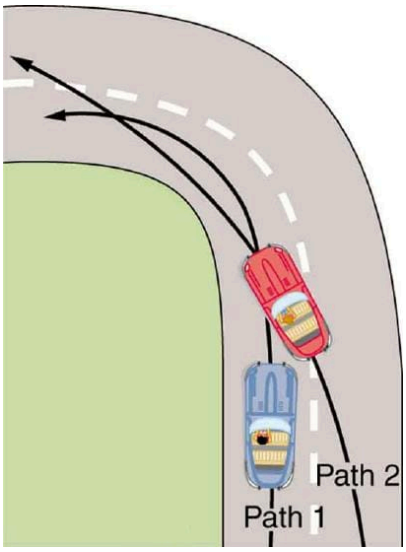


Figure 1.2.11.2: Two paths around a race track curve are shown. Race car drivers will take the inside path (called cutting the corner) whenever possible because it allows them to take the curve at the highest speed.

37. A number of amusement parks have rides that make vertical loops like the one shown in Figure 1.2.11.3. For safety, the cars are attached to the rails in such a way that they cannot fall off. If the car goes over the top at just the right speed, gravity alone will supply the centripetal force. What other force acts and what is its direction if:

(a) The car goes over the top at faster than this speed?

(b)The car goes over the top at slower than this speed?

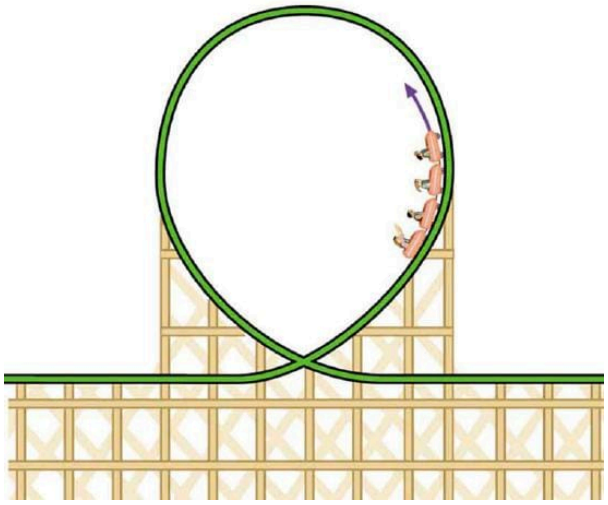


Figure 1.2.11.3: Amusement rides with a vertical loop are an example of a form of curved motion.

38. What is the direction of the force exerted by the car on the passenger as the car goes over the top of the amusement ride pictured in Figure 1.2.11.3 under the following circumstances:

(a) The car goes over the top at such a speed that the gravitational force is the only force acting?

(b) The car goes over the top faster than this speed?

(c) The car goes over the top slower than this speed?

39. Suppose a child is riding on a merry-go-round at a distance about halfway between its center and edge. She has a lunch box resting on wax paper, so that there is very little friction between it and the merry-go-round. Which path shown in Figure 1.2.11.4 will the lunch box take when she lets go? The lunch box leaves a trail in the dust on the merry-go-round. Is that trail straight, curved to the left, or curved to the right? Explain your answer.

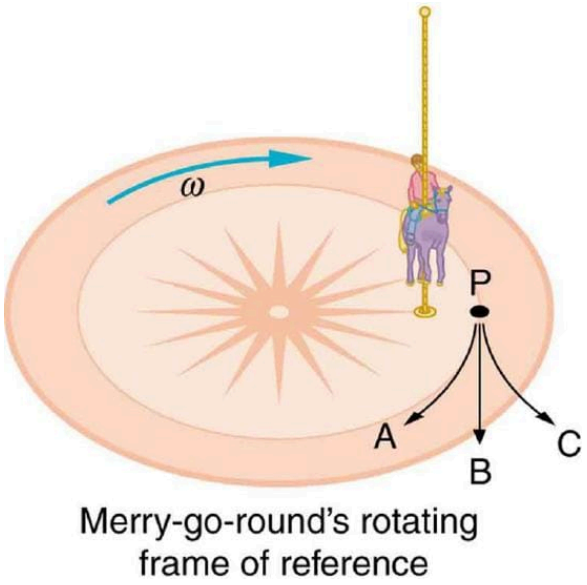


Figure 1.2.11.4: A child riding on a merry-go-round releases her lunch box at point P. This is a view from above the clockwise rotation. Assuming it slides with negligible friction, will it follow path A, B, or C, as viewed from Earth's frame of reference? What will be the shape of the path it leaves in the dust on the merry-go-round?

40. Do you feel yourself thrown to either side when you negotiate a curve that is ideally banked for your car's speed? What is the direction of the force exerted on you by the car seat?

41. Suppose a mass is moving in a circular path on a frictionless table as shown in figure. In the Earth's frame of reference, there is no centrifugal force pulling the mass away from the centre of rotation, yet there is a very real force stretching the string attaching the mass to the nail. Using concepts related to centripetal force and Newton's third law, explain what force stretches the string, identifying its physical origin.

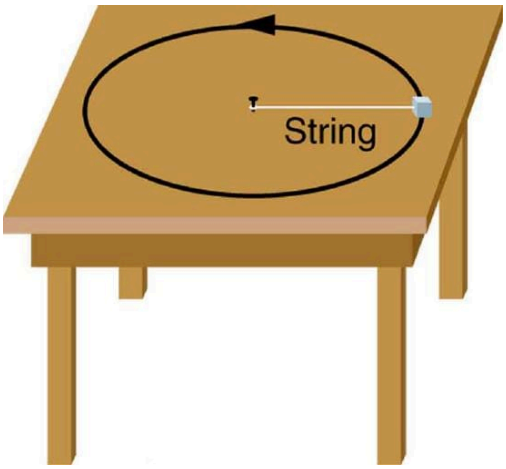


Figure 1.2.11.5: A mass attached to a nail on a frictionless table moves in a circular path. The force stretching the string is real and not fictional. What is the physical origin of the force on the string?

This page titled 1.2.11: Dynamics (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 3.11: Dynamics (Exercise) by Andrew Park is licensed CC BY 4.0.

CHAPTER 2: MECHANICS II - ENERGY AND MOMENTUM, OSCILLATIONS AND WAVES, ROTATION, AND FLUIDS

2.1: Work and Energy

2.1.1: Introduction to Work and Energy

2.1.2: Work- The Scientific Definition

2.1.3: Kinetic Energy and the Work-Energy Theorem

2.1.4: Gravitational Potential Energy

2.1.5: Conservative Forces, Potential Energy, and Conservation of Energy

2.1.6: Spring Potential Energy

2.1.7: Power

2.1.8: Work and Energy (Exercise)

2.2: Impulse and Momentum

2.2.1: Introduction to Linear Momentum and Collisions

2.2.2: Linear Momentum and Force

2.2.3: Impulse

2.2.4: Conservation of Momentum

2.2.5: Elastic Collisions in One Dimension

2.2.6: Inelastic Collisions in One Dimension

2.2.7: Impulse and Momentum (Exercise)

2.3: Oscillations and Waves

2.3.1: Introduction to Oscillatory Motion and Waves

2.3.2: Period and Frequency in Oscillations

2.3.3: Simple Harmonic Motion- A Special Periodic Motion

2.3.4: Forced Oscillations and Resonance

2.3.5: Waves

2.3.6: Wave Interference- Standing Waves and Beats

2.3.7: Sound

2.3.8: Speed of Sound, Frequency, and Wavelength

2.3.9: Doppler Effect and Sonic Booms

2.3.10: Oscillations and Waves (Exercise)

This page titled Chapter 2: Mechanics II - Energy and Momentum, Oscillations and Waves, Rotation, and Fluids is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

- Unit 2: Mechanics II - Energy and Momentum, Oscillations and Waves, Rotation, and Fluids by Andrew Park is licensed CC BY 4.0.

2.1: WORK AND ENERGY

2.1.1: Introduction to Work and Energy
2.1.2: Work- The Scientific Definition
2.1.3: Kinetic Energy and the Work-Energy Theorem
2.1.4: Gravitational Potential Energy
2.1.5: Conservative Forces, Potential Energy, and Conservation of Energy
2.1.6: Spring Potential Energy
2.1.7: Power
2.1.8: Work and Energy (Exercise)

This page titled 2.1: Work and Energy is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

2.1.1: Introduction to Work and Energy



Figure 2.1.1.1: How many forms of energy can you identify in this photograph of a wind farm in Iowa? (credit: Jürgen from Sandesneben, Germany, Wikimedia Commons)

Energy plays an essential role both in everyday events and in scientific phenomena. You can no doubt name many forms of energy, from that provided by our foods, to the energy we use to run our cars, to the sunlight that warms us on the beach. You can also cite examples of what people call energy that may not be scientific, such as someone having an energetic personality. Not only does energy have many interesting forms, it is involved in almost all phenomena, and is one of the most important concepts of physics. What makes it even more important is that the total amount of energy in the universe is constant. Energy can change forms, but it cannot appear from nothing or disappear without a trace. Energy is thus one of a handful of physical quantities that we say is *conserved*.

Conservation of energy (as physicists like to call the principle that energy can neither be created nor destroyed) is based on experiment. Even as scientists discovered new forms of energy, conservation of energy has always been found to apply. Perhaps the most dramatic example of this was supplied by Einstein when he suggested that mass is equivalent to energy (his famous equation $E = mc^2$).

From a societal viewpoint, energy is one of the major building blocks of modern civilization. Energy resources are key limiting factors to economic growth. The world use of energy resources, especially oil, continues to grow, with ominous consequences economically, socially, politically, and environmentally. We will briefly examine the world's energy use patterns at the end of this chapter.

There is no simple, yet accurate, scientific definition for energy. Energy is characterized by its many forms and the fact that it is conserved. We can loosely define **energy** as the ability to do work, admitting that in some circumstances not all energy is available to do work. Because of the association of energy with work, we begin the chapter with a discussion of work. Work is intimately related to energy and how energy moves from one system to another or changes form.

This page titled 2.1.1: Introduction to Work and Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 4.1: Introduction to Work and Energy by Andrew Park is licensed CC BY 4.0.

2.1.2: Work- The Scientific Definition

Learning Objectives

- Explain how an object must be displaced for a force on it to do work.
- Explain how relative directions of force and displacement determine whether the work done is positive, negative, or zero.

What It Means to Do Work

The scientific definition of work differs in some ways from its everyday meaning. Certain things we think of as hard work, such as writing an exam or carrying a heavy load on level ground, are not work as defined by a scientist. The scientific definition of work reveals its relationship to energy—whenever work is done, energy is transferred.

For work, in the scientific sense, to be done, a force must be exerted and there must be motion or displacement in the direction of the force.

Formally, the **work** done on a system by a constant force is defined to be *the product of the component of the force in the direction of motion times the distance through which the force acts*. For one-way motion in one dimension, this is expressed in equation form as

$$W = F_{\parallel} d$$

Where W is work, d is the distance the force acts, and $F_{\parallel}$ is the force parallel to the direction of motion. We can also write this more simply as

$$W = Fd$$

as long as one keeps in mind that the force is in the same direction as the distance.

To find the work done on a system that undergoes motion that is not one-way or that is in two or three dimensions, we would divide the motion into one-way one-dimensional segments and add up the work done over each segment.

WHAT IS WORK?

The work done on a system by a constant force is *the product of the component of the force in the direction of motion times the distance through which the force acts*. For one-way motion in one dimension, this is expressed in equation form as

$$W = Fd$$

Where W is work, d is the distance the force acts, F is the force parallel to the direction of motion.

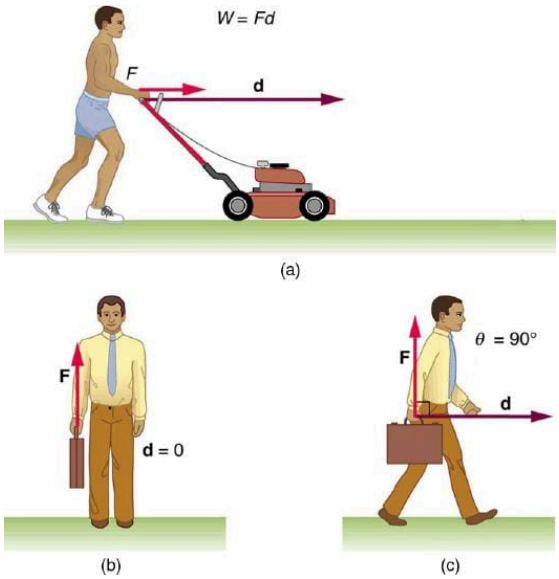


Figure 2.1.2.1: Examples of work. (a) Note that the force is in the direction of motion. (b) A person holding a briefcase does no work on it, because there is no motion. No energy is transferred to or from the briefcase. (c) The person moving the briefcase horizontally at a constant speed does no work on it, and transfers no energy to it. The force acting on the briefcase (against gravity) is perpendicular to the direction of motion.

To examine what the definition of work means, let us consider the other situations shown in Figure 2.1.2.1. The person holding the briefcase in Figure 2.1.2.1(b) does no work, for example. Here $d = 0$, so $W = 0$. Why is it you get tired just holding a load? The answer is that your muscles are doing work against one another, *but they are doing no work on the system of interest* (the "briefcase-Earth system"). There must be motion for work to be done, and there must be a component of the force in the direction of the motion. For example, the person carrying the briefcase on level ground in Figure 2.1.2.1(c) does no work on it, because the force is perpendicular to the motion.

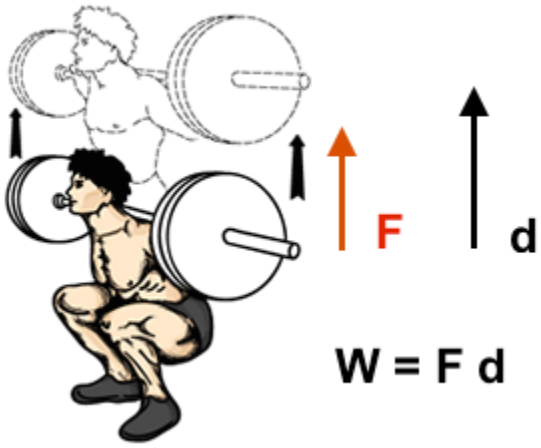


Figure 2.1.2.2: Work done when lifting a mass. Work is positive on the way up because force and displacement point in the same direction. On the way down work is negative due to force and displacement pointing in opposite directions (force up and displacement down).

Work can be positive or negative. In Figure 2.1.2.2 work done lifting the mass is positive because both force and displacement are in the same direction. Likewise, when the mass is lowered the work done is negative because the force and displacement are in *opposite* directions. We will soon see that positive work adds energy to the system and negative work removes energy from a system.

Calculating Work

Work and energy have the same units. From the definition of work, we see that those units are force times distance. Thus, in SI units, work and energy are measured in **newton-meters**. A newton-meter is given the special name **joule** (J), and $1\ \mathrm{J} = 1\ \mathrm{N} \cdot \mathrm{m}$

$= 1\ \text{kg} \cdot \text{m}^2/\text{s}^2$. One joule is not a large amount of energy; it would lift a small 100-gram apple a distance of about 1 meter.

Example 2.1.2.1: Calculating the Work You Do to Push a Lawn Mower Across a Large Lawn

How much work is done on the lawn mower by the person in Figure 2.1.2.1(a) if he exerts a constant force of 75.0 N and pushes the mower 25.0 m on level ground? Compare it with this person's average daily intake of 10,000 kJ (about 2400 kcal) of food energy.

Strategy

We can solve this problem by substituting the given values into the definition of work done on a system, stated in the equation $W = Fd$. The force and displacement are given, so that only the work W is unknown. Note that force and displacement are in the same direction.

Solution

The equation for the work is

$$W = Fd.$$

Substituting the known values gives

$$W = (75.0\ \text{N})(25.0\ \text{m}) = 1875\ \text{J} = 1.88 \times 10^3\ \text{J}.$$

The ratio of the work done to the daily consumption is

$$\frac{W}{1.00 \times 10^4\ \text{kJ}} = \frac{1.88 \times 10^3\ \text{J}}{1.00 \times 10^7\ \text{J}} = 1.88 \times 10^{-4}.$$

Discussion

This ratio is a tiny fraction of what the person consumes, but it is typical. Very little of the energy released in the consumption of food is used to do work. Even when we "work" all day long, less than 10% of our food energy intake is used to do work and more than 90% is converted to thermal energy or stored as chemical energy in fat.

Example 2.1.2.2: Calculating the Work Done Lifting a Mass

The 60.0 kg mass shown in Figure 2.1.2.2 is raised at a constant speed through a vertical distance of 0.800 m. What is the force that must be exerted? How much work is done lifting the mass?

Strategy

Constant speed tells us that the upward force must have the same magnitude as the downward force of gravity $F = mg$ (to have a zero net force). Using that information, along with the given values for mass and displacement, we can determine work using $W = Fd$.

Solution

The force is given by

$$F = mg = (60.0\ \text{kg})\left(9.80\ \text{m/s}^2\right) = 588\ \text{N}.$$

The work done is

$$W = Fd = (588\ \text{N})(0.800\ \text{m}) = 470\ \text{J}.$$

Discussion

Note that the work done lowering the mass back to its starting position would be $-470\ \text{J}$because force and displacement point in opposite directions. The total work done raising and lowering the mass is $470\ \text{J} - 470\ \text{J} = 0\ \text{J}$.

Section Summary

- Work is the transfer of energy by a force acting on an object as it is displaced.
- The work W that a force F does on an object is the product of the magnitude F of the force parallel to the motion, times the magnitude d of the displacement. In symbols,

$$W = Fd.$$

- The SI unit for work and energy is the joule (J), where $1\ \text{J} = 1\ \text{N} \cdot \text{m} = 1\ \text{kg} \cdot \text{m}^2/\text{s}^2$.
- The work done by a force is zero if the displacement is either zero or perpendicular to the force.
- The work done is positive if the force and displacement have the same direction, and negative if they have opposite direction.

Glossary

energy
: the ability to do work

work
: the transfer of energy by a force that causes an object to be displaced; the product of the component of the force in the direction of the displacement and the magnitude of the displacement

joule
: SI unit of work and energy, equal to one newton-meter

This page titled 2.1.2: Work- The Scientific Definition is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 4.2: Work- The Scientific Definition by Andrew Park is licensed CC BY 4.0.

2.1.3: Kinetic Energy and the Work-Energy Theorem

Learning Objectives

- Explain work as a transfer of energy and net work as the work done by the net force.
- Explain and apply the work-energy theorem.

Work Transfers Energy

What happens to the work done on a system? Energy is transferred into the system, but in what form? Does it remain in the system or move on? The answers depend on the situation. For example, if a lawn mower is pushed just hard enough to keep it going at a constant speed, then energy put into the mower by the person is removed continuously by friction, and eventually leaves the system in the form of heat transfer. In contrast, work done on a briefcase by a person carrying it up stairs is stored in the briefcase-Earth system and can be recovered at any time. In fact, the building of the pyramids in ancient Egypt is an example of storing energy in a system by doing work on the system. Some of the energy imparted to the stone blocks in lifting them during construction of the pyramids remains in the stone-Earth system and has the potential to do work.

In this section we begin the study of various types of work and forms of energy. We will find that some types of work leave the energy of a system constant, for example, whereas others change the system in some way, such as making it move. We will also develop definitions of important forms of energy, such as the energy of motion.

Net Work and the Work-Energy Theorem

We know from the study of Newton's laws that net force causes acceleration. We will see in this section that work done by the net force gives a system energy of motion, and in the process we will also find an expression for the energy of motion.

Let us start by considering the total, or net, work done on a system. Net work is defined to be the sum of work done by all external forces—that is, **net work** is the work done by the net external force $\mathbf{F}_{\text{net}}$. In equation form, this is $W_{\text{net}} = F_{\text{net}}\, d$.

Consider a one-dimensional situation where a force is used to accelerate an object in a direction parallel to its initial velocity. Such a situation occurs for the package on the roller belt conveyor system shown in Figure 2.1.3.1.

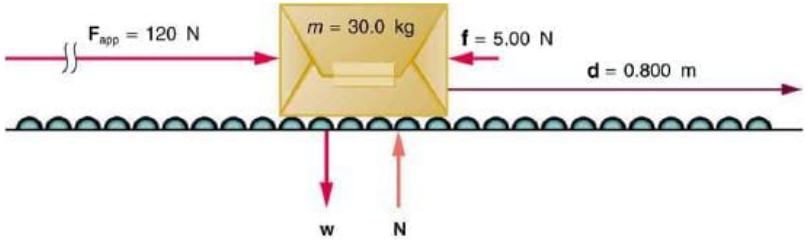


Figure 2.1.3.1: A package on a roller belt is pushed horizontally through a distance **d**.

The force of gravity and the normal force acting on the package are perpendicular to the displacement and do no work. Moreover, they are also equal in magnitude and opposite in direction so they cancel in calculating the net force. The net force arises solely from the horizontal applied force $\mathbf{F}_{\text{app}}$ and the horizontal friction force **f**. Thus, as expected, the net force is parallel to the displacement and the net work is given by

$$W_{\text{net}} = F_{\text{net}}\, d.$$

The effect of the net force $\mathbf{F}_{\text{net}}$ is to accelerate the package from v_0 to v. By using Newton's second law, and doing some algebra, we can reach an interesting conclusion. Substituting $F_{\text{net}} = ma$ from Newton's second law gives

$$W_{\text{net}} = mad.$$

To get a relationship between net work and the speed given to a system by the net force acting on it, we take $d = x - x_0$ and use the equation studied in "Motion Equations for Constant Acceleration in One Dimension" for the change in speed over a distance d if the acceleration has the constant value a; namely, $v^2 = v_0^2 + 2ad$ (note that a appears in the expression for the net work). Solving for acceleration gives $a = \frac{v^2 - v_0^2}{2d}$. When a is substituted into the preceding expression for W_{net} , we obtain

$$W_{\text{net}} = m\left(\frac{v^2 - v_0^2}{2d}\right)d.$$

The d cancels, and we rearrange this to obtain

$$W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2.$$

This expression is called the **work-energy theorem**, and it actually applies *in general* (even for forces that vary in direction and magnitude), although we have derived it for the special case of a constant force parallel to the displacement. The theorem implies that the net work on a system equals the change in the quantity $\frac{1}{2}mv^2$. This quantity is our first example of a form of energy.

THE WORK-ENERGY THEOREM

The net work on a system equals the change in the quantity $\frac{1}{2}mv^2$.

$$W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2$$

The quantity $\frac{1}{2}mv^2$ in the work-energy theorem is defined to be the translational **kinetic energy** (KE) of a mass m moving at a speed v. In equation form, the translational kinetic energy,

$$\text{KE} = \frac{1}{2}mv^2,$$

is the energy associated with translational motion. Kinetic energy is a form of energy associated with the motion of a particle, single body, or system of objects moving together. The work-energy theorem can be compactly written as,

$$W_{\text{net}} = \Delta\text{KE}$$

where $\triangle\text{KE}$ is understood to mean "change in KE."

We are aware that it takes energy to get an object, like a car or the package in Figure 2.1.3.1, up to speed, but it may be a bit surprising that kinetic energy is proportional to speed squared. This proportionality means, for example, that a car traveling at 40 m/s has four times the kinetic energy it has at 20 m/s, helping to explain why high-speed collisions are so devastating. We will now consider a series of examples to illustrate various aspects of work and energy.

Example 2.1.3.1: Calculating the Kinetic Energy of a Package

Suppose a 30.0-kg package on the roller belt conveyor system in Figure 2.1.3.1 is moving at 0.500 m/s. What is its kinetic energy?

Strategy

Because the mass m and speed v are given, the kinetic energy can be calculated from its definition as given in the equation $\text{KE} = \frac{1}{2}mv^2$.

Solution

The kinetic energy is given by

$$\text{KE} = \frac{1}{2}mv^2.$$

Entering known values gives

$$\mathrm{KE} = 0.5(30.0\ \mathrm{kg})(0.500\ \mathrm{m/s})^2,$$

which yields

$$\mathrm{KE} = 3.75\ \mathrm{kg \cdot m^2/s^2} = 3.75\ \mathrm{J}.$$

Discussion

Note that the unit of kinetic energy is the joule, the same as the unit of work, as mentioned when work was first defined. It is also interesting that, although this is a fairly massive package, its kinetic energy is not large at this relatively low speed. This fact is consistent with the observation that people can move packages like this without exhausting themselves.

Example 2.1.3.2: Determining the Work to Accelerate a Package

Suppose that you push on the 30.0-kg package in Figure 2.1.3.1 with a constant force of 120 N through a distance of 0.800 m, and that the opposing friction force averages 5.00 N.

(a) Calculate the net work done on the package. (b) Solve the same problem as in part (a), this time by finding the work done by each force that contributes to the net force.

Strategy and Concept for (a)

This is a motion in one dimension problem, because the downward force (from the weight of the package) and the normal force have equal magnitude and opposite direction, so that they cancel in calculating the net force, while the applied force, friction, and the displacement are all horizontal. (See Figure 2.1.3.1.) As expected, the net work is the net force times distance.

Solution for (a)

The net force is the push force minus friction, or $F_{\mathrm{net}} = 120\ \mathrm{N} - 5.00\ \mathrm{N} = 115\ \mathrm{N}$. Thus the net work is

$$\begin{aligned} W_{\mathrm{net}} &= F_{\mathrm{net}}\, d = (115\ \mathrm{N})(0.800\ \mathrm{m}) \\ &= 92.0\ \mathrm{N \cdot m} = 92.0\ \mathrm{J}. \end{aligned}$$

Discussion for (a)

This value is the net work done on the package. The person actually does more work than this, because friction opposes the motion. Friction does negative work and removes some of the energy the person expends and converts it to thermal energy. The net work equals the sum of the work done by each individual force.

Strategy and Concept for (b)

The forces acting on the package are gravity, the normal force, the force of friction, and the applied force. The normal force and force of gravity are each perpendicular to the displacement, and therefore do no work.

Solution for (b)

The applied force does work.

$$W_{\mathrm{app}} = F_{\mathrm{app}}\, d = (120\ \mathrm{N})(0.800\ \mathrm{m}) = 96.0\ \mathrm{J}$$

The friction force and displacement are in opposite directions. The work done by friction is therefore negative.

$$W_{\mathrm{fr}} = -F_{\mathrm{fr}} d = -(5.00\ \mathrm{N})(0.800\ \mathrm{m}) = -4.00\ \mathrm{J}$$

So the amounts of work done by gravity, by the normal force, by the applied force, and by friction are, respectively,

$$W_{\text{gr}} = 0,$$
$$W_{\text{N}} = 0,$$
$$W_{\text{app}} = 96.0 \text{ J},$$
$$W_{\text{fr}} = -4.00 \text{ J}.$$

The total work done as the sum of the work done by each force is then seen to be

$$W_{\text{total}} = W_{\text{gr}} + W_{\text{N}} + W_{\text{app}} + W_{\text{fr}} = 92.0 \text{ J}.$$

Discussion for (b)

The calculated total work W_{total} as the sum of the work by each force agrees, as expected, with the work W_{net} done by the net force. The work done by a collection of forces acting on an object can be calculated by either approach.

Example 2.1.3.3: Determining Speed from Work and Energy

Find the speed of the package in Figure 2.1.3.1 at the end of the push, using work and energy concepts.

Strategy

Here the work-energy theorem can be used, because we have just calculated the net work, W_{net}, and the initial kinetic energy, $\frac{1}{2}mv_0^2$. These calculations allow us to find the final kinetic energy, $\frac{1}{2}mv^2$, and thus the final speed v.

Solution

The work-energy theorem in equation form is

$$W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2.$$

Solving for $\frac{1}{2}mv^2$ gives

$$\frac{1}{2}mv^2 = W_{\text{net}} + \frac{1}{2}mv_0^2.$$

Thus,

$$\frac{1}{2}mv^2 = 92.0 \text{ J} + 3.75 \text{ J} = 95.75 \text{ J}.$$

Solving for the final speed as requested and entering known values gives

$$v = \sqrt{\frac{2(95.75, J)}{m}} = \sqrt{\frac{191.5 \text{ kg} \cdot \text{m}^2/\text{s}^2}{30.0 \text{ kg}}}$$
$$= 2.53 \text{ m/s}.$$

Discussion

Using work and energy, we not only arrive at an answer, we see that the final kinetic energy is the sum of the initial kinetic energy and the net work done on the package. This means that the work indeed adds to the energy of the package.

Example 2.1.3.4: Work and Energy Can Reveal Distance, Too

How far does the package in Figure 2.1.3.1 coast after the push, assuming friction remains constant? Use work and energy considerations.

Strategy

We know that once the person stops pushing, friction will bring the package to rest. In terms of energy, friction does negative work until it has removed all of the package's kinetic energy. The work done by friction can be expressed as the force of friction times the distance traveled *and* as the change in kinetic energy. Equating both expressions for work gives us a way of finding the distance traveled.

Solution

The normal force and force of gravity cancel in calculating the net force. The horizontal friction force is then the net force, and it acts opposite to the displacement. To reduce the kinetic energy of the package to zero, the work W_{fr} by friction must be equal to the change in kinetic energy.

$$W_{\text{fr}} = -F_{\text{fr}}d = \Delta\text{KE}$$
$$-F_{\text{fr}}d = \Delta\text{KE}$$

and so

$$-F_{\text{fr}}d = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2$$

Initial speed is the speed at the instant the push stops (the result determined in the previous example). The final speed is zero for this case and leads to the expression below.

$$F_{\text{fr}}d = \frac{1}{2}mv_0^2$$

Solving for distance, we obtain the result shown below.

$$d = \frac{95.75\ \text{J}}{5.00\ \text{N}} = 19.2\ \text{m}$$

Discussion

This is a reasonable distance for a package to coast on a relatively friction-free conveyor system. Note that the work done by friction is negative (the force is in the opposite direction of motion), so it removes the kinetic energy.

Some of the examples in this section can be solved without considering energy, but at the expense of missing out on gaining insights about what work and energy are doing in this situation. On the whole, solutions involving energy are generally shorter and easier than those using kinematics and dynamics alone.

Section Summary

- The net work W_{net} is the work done by the net force acting on an object.
- Work done on an object transfers energy to the object.
- The translational kinetic energy of an object of mass m moving at speed v is $\text{KE} = \frac{1}{2}mv^2$.
- The work-energy theorem states that the net work W_{net} on a system changes its kinetic energy, $W_{\text{net}} = \Delta\text{KE} = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2$.

Glossary

net work
: work done by the net force, or vector sum of all the forces, acting on an object

work-energy theorem
: the result, based on Newton's laws, that the net work done on an object is equal to its change in kinetic energy

kinetic energy
: the energy an object has by reason of its motion, equal to $\frac{1}{2}mv^2$ for the translational (i.e., non-rotational) motion of an object of mass m moving at speed v

This page titled 2.1.3: Kinetic Energy and the Work-Energy Theorem is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 4.3: Kinetic Energy and the Work-Energy Theorem by Andrew Park is licensed CC BY 4.0.

2.1.4: Gravitational Potential Energy

Learning Objectives

- Explain gravitational potential energy in terms of work done against gravity.
- Show that the gravitational potential energy of an object near surface of Earth is given by product of its mass, height, and gravitational acceleration.
- Show how knowledge of the potential energy as a function of position can be used to simplify calculations and explain physical phenomena.

Work Done Against Gravity

Consider lifting a weight (of, say, 5 kg, which weighs about 50 N) from the floor to a table surface at a height of 1 meter from the floor. In lifting this weight, you have done work, both in the usual sense (if you do this lifting of weight over and over, you will eventually get tired) and in the scientific sense we defined in the earlier section ($W = Fd$, and you applied an upward force as you displaced the weight upward by 1 meter). One thing is different in this scenario, however. Earlier, in describing work-kinetic energy theorem, we said that net work done on an object changes the kinetic energy of the object ($\Delta \mathrm{KE} = W_{\mathrm{net}}$). But in this case, if you imagine lifting the weight very slowly up to the height of the table, you will find that the kinetic energy of the weight does not change. It starts at zero and remains at zero throughout this very slow motion. What is going on here?

The key difference is that, for this weight, there are now two forces acting on it as it is being lifted. The gravitational force pulls downward, while the force you apply is directed upward. If you are lifting the weight very slowly, then its acceleration is very close to zero, so the net force on it is zero ($F_{\mathrm{net}} = 0$), which means the *net work* done on the weight is also zero ($W_{\mathrm{net}} = F_{\mathrm{net}}\, d$; the applied upward force does positive work, while the downward gravitational force does negative work).

While this is a correct description (the negative work done by the gravity cancels out the positive work done by the applied force), this description misses something. Specifically, you can see that this weight that has been lifted by 1 meter *does* have more energy than a weight that was resting on the floor. This is how you can see it: if you let go of the weight (after removing the table), as the weight originally at 1-meter height reaches the floor, it gains a substantial kinetic energy as it continues to be pulled by gravity. Rather than saying that this kinetic energy simply appeared as the weight fell through the 1-meter height, we opt for the description that this weight already had this energy, when it was at rest at the 1-meter height, just in a form that was not kinetic energy. We call this energy associated with the height of objects the **gravitational potential energy**, PE_g. And in this example, the energy originally came from the work done by the applied force—rather than considering positive work done by the applied force to be cancelled out by the negative work done by the gravity, we say that this positive work done is *stored* as positive PE_g.

This phenomenon of potential energy turns out to be quite general and goes beyond the gravitational force. There are other forces in nature where negative work done by these forces turns into positive potential energy. We call these forces **conservative forces**, which will be the subject of the next section. For the purpose of this section, following is the quick derivation of formula for PE_g (everything in terms of symbols now). In lifting the weight of mass m through a height h slowly, you have to apply a force equal to its weight, mg. So the work done is $W = Fd = mgh$, and since it's this work done that is stored as gravitational potential energy, the gravitational potential energy must be:

$$\mathrm{PE}_g = mgh.$$

Converting Between Potential Energy and Kinetic Energy

Gravitational potential energy may be converted to other forms of energy, such as kinetic energy. If we release the mass, gravitational force will do an amount of work equal to mgh on it, thereby increasing its kinetic energy by that same amount (by the work-energy theorem). We will find it more useful to consider just the conversion of PE_{g} to **KE** without explicitly considering the intermediate step of work. (See Example 2.1.4.2.) This shortcut makes it is easier to solve problems using energy (if possible) rather than explicitly using forces.

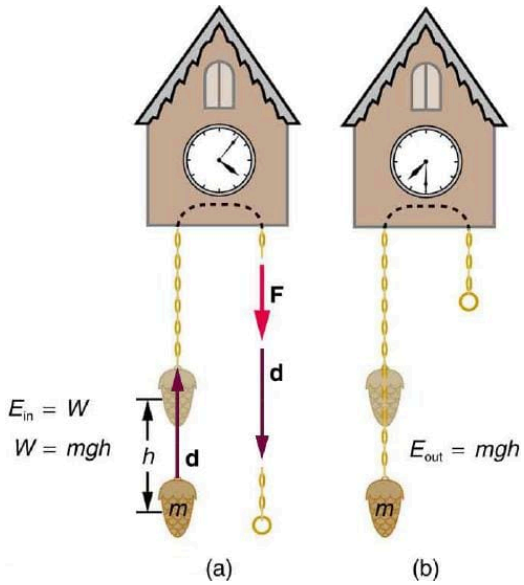


Figure 2.1.4.1: (a) The work done to lift the weight is stored in the mass-Earth system as gravitational potential energy. (b) As the weight moves downward, this gravitational potential energy is transferred to the cuckoo clock.

More precisely, we define the *change* in gravitational potential energy $\Delta \mathrm{PE_g}$ to be

$$\Delta \mathrm{PE_g} = \mathrm{PE_{gf}} - \mathrm{PE_{gi}} = mgh_\mathrm{f} - mgh_\mathrm{i} = mg\left(h_\mathrm{f} - h_\mathrm{i}\right) = mg\Delta h$$

Note that Δh is positive when the final height is greater than the initial height, and vice versa. For example, if a 0.500-kg mass hung from a cuckoo clock is raised 1.00 m, then its change in gravitational potential energy is

$$\begin{aligned} mg\Delta h &= (0.500\ \mathrm{kg})\left(9.80\ \mathrm{m/s^2}\right)(1.00\ \mathrm{m}) \\ &= 4.90\ \mathrm{kg \cdot m^2/s^2} = 4.90\ \mathrm{J}. \end{aligned}$$

Note that the units of gravitational potential energy turn out to be joules, the same as for work and other forms of energy. As the clock runs, the mass is lowered. We can think of the mass as gradually giving up its 4.90 J of gravitational potential energy, *without directly considering the force of gravity that does the work*.

Using Potential Energy to Simplify Calculations

The equation $\Delta \mathrm{PE_g} = mg\Delta h$ applies for any path that has a change in height of $\Delta \mathrm{h}$, not just when the mass is lifted straight up. (See Figure 2.1.4.2.) It is much easier to calculate $mg\Delta h$ (a simple multiplication) than it is to calculate the work done along a complicated path. The idea of gravitational potential energy has the double advantage that it is very broadly applicable and it makes calculations easier. From now on, we will consider that any change in vertical position $\Delta \mathrm{h}$ of a mass m is accompanied by a change in gravitational potential energy $mg\Delta h$, and we will avoid the equivalent but more difficult task of calculating work done by or against the gravitational force.

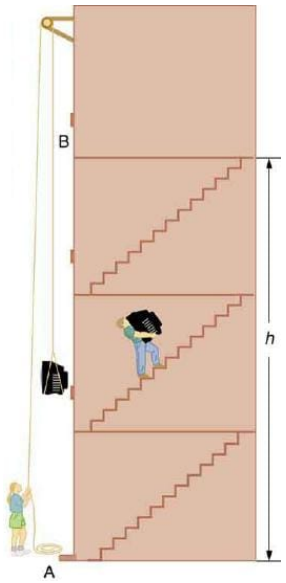


Figure 2.1.4.2: The change in gravitational potential energy ($\Delta \mathrm{PE_g}$) between points A and B is independent of the path. $\Delta \mathrm{PE_g} = mg\Delta h$ for any path between the two points. Gravity is one of a small class of forces where the work done by or against the force depends only on the starting and ending points, not on the path between them.

Example 2.1.4.1: The Force to Stop Falling

A 60.0-kg person jumps onto the floor from a height of 3.00 m. If he lands stiffly (with his knee joints compressing by 0.500 cm), calculate the force on the knee joints.

Strategy

This person's energy is brought to zero in this situation by the work done on him by the floor as he stops. The initial $\mathrm{PE_g}$ is transformed into **KE** as he falls. The work done by the floor reduces this kinetic energy to zero.

Solution

The work done on the person by the floor as he stops is given by

$$W = -Fd,$$

with a minus sign because the displacement while stopping and the force from floor are in opposite directions. The floor removes energy from the system, so it does negative work.

The kinetic energy the person has upon reaching the floor is the amount of potential energy lost by falling through height $\Delta \mathrm{h}$:

$$\mathrm{KE} = -\Delta \mathrm{PE_g} = -mg\Delta h,$$

The distance d that the person's knees bend is much smaller than the height $\Delta \mathrm{h}$ of the fall, so the additional change in gravitational potential energy during the knee bend is ignored.

The work W done by the floor on the person stops the person and brings the person's kinetic energy to zero:

$$W = -\mathrm{KE} = mg\Delta h.$$

Combining this equation with the expression for W gives

$$-Fd = mg\Delta h.$$

Recalling that $\Delta \mathrm{h}$ is negative because the person fell *down*, the force on the knee joints is given by

$$F = -\frac{mg\Delta h}{d} = -\frac{(60.0\ \text{kg})\,(9.80\ \text{m/s}^2)\,(-3.00\ \text{m})}{5.00 \times 10^{-3}\ \text{m}} = 3.53 \times 10^5\ \text{N}.$$

Discussion

Such a large force (500 times more than the person's weight) over the short impact time is enough to break bones. A much better way to cushion the shock is by bending the legs or rolling on the ground, increasing the time over which the force acts. A bending motion of 0.5 m this way yields a force 100 times smaller than in the example. A kangaroo's hopping shows this method in action. The kangaroo is the only large animal to use hopping for locomotion, but the shock in hopping is cushioned by the bending of its hind legs in each jump.(See Figure 2.1.4.3.)



Figure 2.1.4.3: The work done by the ground upon the kangaroo reduces its kinetic energy to zero as it lands. However, by applying the f the ground on the hind legs over a longer distance, the impact on the bones is reduced. (credit: Chris Samuel, Flickr)

Example 2.1.4.2: Finding the Speed of a Roller Coaster from its Height

(a) What is the final speed of the roller coaster shown in Figure 2.1.4.4 if it starts from rest at the top of the 20.0 m hill and work done by frictional forces is negligible? (b) What is its final speed (again assuming negligible friction) if its initial speed is 5.00 m/s?

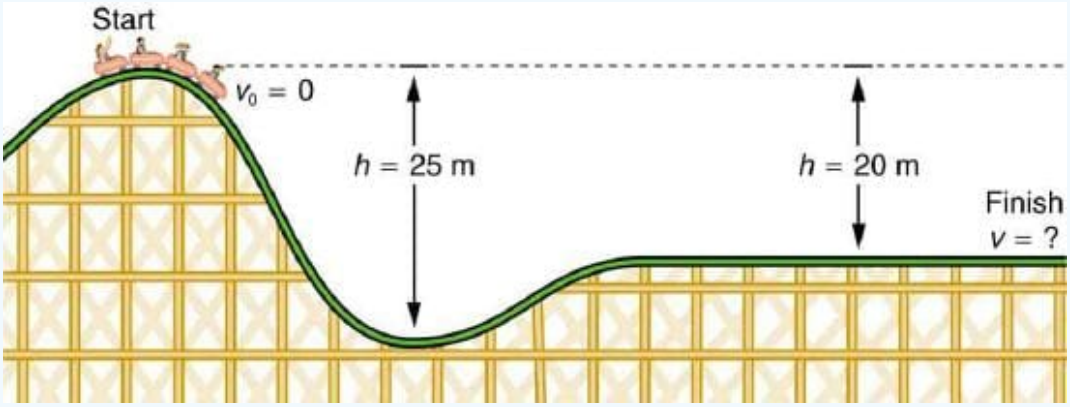


Figure 2.1.4.4: The speed of a roller coaster increases as gravity pulls it downhill and is greatest at its lowest point. Viewed in terms of energy, the roller-coaster-Earth system's gravitational potential energy is converted to kinetic energy. If work done by friction is negligible, all ΔPE_g is converted to **KE**.

Strategy

The roller coaster loses potential energy as it goes downhill. We neglect friction, so that the remaining force exerted by the track is the normal force, which is perpendicular to the direction of motion and does no work. The net work on the roller coaster is then done by gravity alone. The *loss* of gravitational potential energy from moving *downward* through a distance h equals the *gain* in kinetic energy. This can be written in equation form as $-\Delta \mathrm{PE_g} = \Delta \mathrm{KE}$. Using the equations for $\mathrm{PE_g}$ and **KE**, we can solve for the final speed v, which is the desired quantity.

Solution for (a)

$$-\Delta \mathrm{PE_g} = \Delta \mathrm{KE}$$

becomes

$$-\left(\mathrm{PE_{gf}} - \mathrm{PE_{gi}}\right) = \frac{1}{2}mv_\mathrm{f}^2 - \frac{1}{2}mv_\mathrm{i}^2$$

Noting that the initial speed is zero, and using the formula for gravitational potential energy, we have

$$-\left(mgh_\mathrm{f} - mgh_\mathrm{i}\right) = \frac{1}{2}mv_\mathrm{f}^2$$

Taking the initial height as zero gives

$$-mgh_\mathrm{f} = \frac{1}{2}mv_\mathrm{f}^2$$

Solving for the final speed, we find that mass cancels and that

$$v_\mathrm{f} = \sqrt{-2gh_\mathrm{f}}$$

Substituting known values,

$$v_\mathrm{f} = \sqrt{-2gh_\mathrm{f}} = \sqrt{(-2)\left(9.80\ \mathrm{m/s^2}\right)(-20.0\ \mathrm{m})} = 19.8\ \mathrm{m/s}$$

Please note that the height used above was negative. This is because we chose the initial height to be our reference level for gravitational potential energy. Our final height is 20 meters *below* that reference level.

Solution for (b)

Again

$$-\Delta \mathrm{PE_g} = \Delta \mathrm{KE}.$$

Thus,

$$-\left(\mathrm{PE_{gf}} - \mathrm{PE_{gi}}\right) = \frac{1}{2}mv_\mathrm{f}^2 - \frac{1}{2}mv_\mathrm{i}^2$$

Taking the initial height as zero leads to

$$-mgh_\mathrm{f} = \frac{1}{2}mv_\mathrm{f}^2 - \frac{1}{2}mv_\mathrm{i}^2$$

Rearranging gives

$$\frac{1}{2}mv_\mathrm{f}^2 = -mgh_\mathrm{f} + \frac{1}{2}mv_\mathrm{i}^2$$

This means that the final kinetic energy is the sum of the initial kinetic energy and the gravitational potential energy. Mass again cancels, and

$$v_\mathrm{f} = \sqrt{-2gh_\mathrm{f} + v_\mathrm{i}^2}$$

Now, substituting known values gives

$$v_{\mathrm{f}} = \sqrt{-2gh_{\mathrm{f}} + v_{\mathrm{i}}^2} = \sqrt{(-2)\,(9.80\ \mathrm{m/s^2})\,(-20.0\ \mathrm{m}) + (5.00\ \mathrm{m/s})^2} = 20.4\ \mathrm{m/s}.$$

Discussion and Implications

First, note that mass cancels. This is quite consistent with observations that all objects fall at the same rate if friction is negligible. Second, only the speed of the roller coaster is considered; there is no information about its direction at any point. This reveals another general truth. When friction is negligible, the speed of a falling body depends only on its initial speed and height, and not on its mass or the path taken. For example, the roller coaster will have the same final speed whether it falls 20.0 m straight down or takes a more complicated path like the one in the figure. Third, and perhaps unexpectedly, the final speed in part (b) is greater than in part (a), but by far less than 5.00 m/s. Finally, note that speed can be found at *any* height along the way by simply using the appropriate value of h at the point of interest.

We have seen that work done by or against the gravitational force depends only on the starting and ending points, and not on the path between, allowing us to define the simplifying concept of gravitational potential energy. We can do the same thing for a few other forces, and we will see that this leads to a formal definition of the law of conservation of energy.

Section Summary

- Work done against gravity in lifting an object becomes potential energy of the object-Earth system.
- The change in gravitational potential energy, $\Delta\mathrm{PE_g}$, is $\Delta\mathrm{PE_g} = mg\Delta h$, with $\Delta\mathrm{h}$ being the change in height and g the acceleration due to gravity.
- The gravitational potential energy of an object near Earth's surface is due to its position in the mass-Earth system. Only differences in gravitational potential energy, $\Delta\mathrm{PE_g}$, have physical significance.
- As an object descends without friction, its gravitational potential energy changes into kinetic energy corresponding to increasing speed, so that $\Delta\mathrm{KE} = -\Delta\mathrm{PE_g}$.

Glossary

gravitational potential energy
: energy associated with height of objects on the Earth

This page titled 2.1.4: Gravitational Potential Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 4.4: Gravitational Potential Energy by Andrew Park is licensed CC BY 4.0.

2.1.5: Conservative Forces, Potential Energy, and Conservation of Energy

Learning Objectives

- Define conservative force, potential energy, and mechanical energy.
- Apply conservation of mechanical energy to simple physical situations.
- Explain the law of the conservation of energy.
- Describe some of the many forms of energy.

Potential Energy and Conservative Forces

Work is done by a force, and some forces, such as weight, have special characteristics. A **conservative force** is one, like the gravitational force, for which work done by or against it depends only on the starting and ending points of a motion and not on the path taken. This happens if the force is a function of position alone and not time, velocity, or other parameters. We can define a **potential energy** (**PE**) for any conservative force, just as we did for the gravitational force. For example, when you wind up a toy, an egg timer, or an old-fashioned watch, you do work against its spring and store energy in it. (We treat these springs as ideal, in that we assume there is no friction and no production of thermal energy.) This stored energy is recoverable as work, and it is useful to think of it as potential energy contained in the spring. Indeed, the reason that the spring has this characteristic is that its force is *conservative*. That is, a conservative force results in stored or potential energy. Gravitational potential energy is one example, as is the energy stored in a spring. We will also see how conservative forces are related to the conservation of energy.

POTENTIAL ENERGY AND CONSERVATIVE FORCES

Potential energy is the energy a system has due to position, shape, or configuration. It is stored energy that is completely recoverable.

A conservative force is a force that is a function of position alone, with the result that the work done by the force depends only on the starting and ending points of a motion and not on the particular path taken.

Change in potential energy comes from the work done against a conservative force. Or flipping this around, potential energy increases as a conservative force does a negative work. That is,

$$\Delta \mathrm{PE} = -W_c,$$

where W_c is work done by a conservative force.

Conservation of Mechanical Energy

Consider an object in a system. If it gains any kinetic energy, this is a result of a net work done on the object, according to the work-kinetic energy theorem. If only conservative forces, such as the gravitational force or a spring force, do work in this system, then as kinetic energy increases with the net work done by the conservative forces, the system loses potential energy. That is, $\Delta \mathrm{KE} = W_c = -\Delta \mathrm{PE}$. In other words,

$$\Delta \mathrm{KE} + \Delta \mathrm{PE} = 0.$$

This equation means that the total kinetic and potential energy is constant for any process involving only conservative forces. That is,

$$\left.\begin{array}{l} \mathrm{KE} + \mathrm{PE} = \text{ constant} \\ \text{or} \\ \mathrm{KE_i} + \mathrm{PE_i} = \mathrm{KE_f} + \mathrm{PE_f} \end{array}\right\} \text{(conservative forces only)},$$

where i and f denote initial and final values. This equation is a form of the work-energy theorem for conservative forces; it is known as the **conservation of mechanical energy** principle. Remember that this applies to the extent that all the forces are conservative, so that friction is negligible. The total kinetic plus potential energy of a system is defined to be its **mechanical energy**, (**KE+PE**). In a system that experiences only conservative forces, there is a potential energy associated with each force, and the energy only changes form between **KE** and the various types of **PE**, with the total energy remaining constant.

Example 2.1.5.1: Using Conservation of Mechanical Energy to Calculate the Speed of a Toy Car

A 0.100-kg toy car is propelled by a compressed spring, as shown in Figure 2.1.5.1. The car follows a track that rises 0.180 m above the starting point. The compressed spring has a potential energy of 0.200 J. Assuming work done by friction to be negligible, find (a) how fast the car is going before it starts up the slope and (b) how fast it is going at the top of the slope.

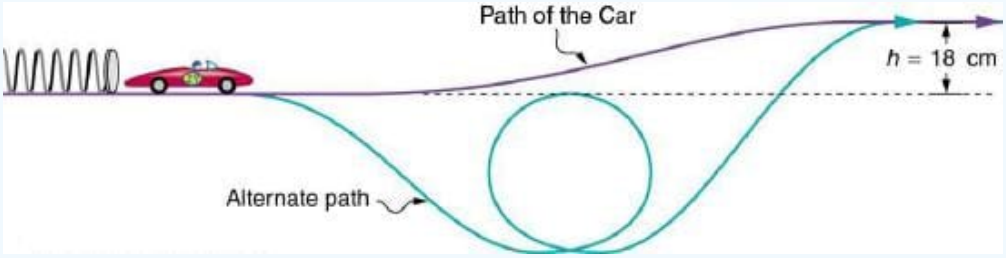


Figure 2.1.5.1: A toy car is pushed by a compressed spring and coasts up a slope. Assuming negligible friction, the potential energy in the spring is first completely converted to kinetic energy, and then to a combination of kinetic and gravitational potential energy as the car rises. The details of the path are unimportant because all forces are conservative—the car would have the same final speed if it took the alternate path shown.

Strategy

The spring force and the gravitational force are conservative forces, so conservation of mechanical energy can be used. Thus,

$$\mathrm{KE_i} + \mathrm{PE_i} = \mathrm{KE_f} + \mathrm{PE_f}$$

or

$$\frac{1}{2}mv_\mathrm{i}^2 + mgh_\mathrm{i} + \mathrm{PE_{si}} = \frac{1}{2}mv_\mathrm{f}^2 + mgh_\mathrm{f} + \mathrm{PE_{sf}}$$

where h is the height (vertical position) and $\mathrm{PE_s}$ is the potential energy of the spring. This general statement looks complex but becomes much simpler when we start considering specific situations. First, we must identify the initial and final conditions in a problem; then, we enter them into the last equation to solve for an unknown.

Solution for (a)

This part of the problem is limited to conditions just before the car is released and just after it leaves the spring. Take the initial height to be zero, so that both h_i and h_f are zero. Furthermore, the initial speed v_i is zero and the final compression of the spring is zero, and so several terms in the conservation of mechanical energy equation are zero and it simplifies to

$$\mathrm{PE_{si}} = \frac{1}{2}mv_\mathrm{f}^2.$$

In other words, the initial potential energy in the spring is converted completely to kinetic energy in the absence of friction. Solving for the final speed and entering known values yields

$$v_\mathrm{f} = \sqrt{\frac{2\mathrm{PE_{si}}}{m}} = \sqrt{\frac{(2)(0.200\ \mathrm{J})}{0.100\ \mathrm{kg}}} = 2.00\ \mathrm{m/s}.$$

Solution for (b)

One method of finding the speed at the top of the slope is to consider conditions just before the car is released and just after it reaches the top of the slope, completely ignoring everything in between. Doing the same type of analysis to find which terms are zero, the conservation of mechanical energy becomes

$$\mathrm{PE}_{\mathrm{si}} = \frac{1}{2}mv_{\mathrm{f}}^2 + mgh_{\mathrm{f}}.$$

This form of the equation means that the spring's initial potential energy is converted partly to gravitational potential energy and partly to kinetic energy. The final speed at the top of the slope will be less than at the bottom. Solving for vfsize 12{v rSub { size 8{f} } } {} and substituting known values gives

$$v_{\mathrm{f}} = \sqrt{\frac{(2)\,(\mathrm{PE}_{\mathrm{si}} - mgh_{\mathrm{f}})}{m}} = \sqrt{\frac{0.0472\ \mathrm{J}}{0.100\ \mathrm{kg}}} = 0.687\ \mathrm{m/s}$$

Discussion

Another way to solve this problem is to realize that the car's kinetic energy before it goes up the slope is converted partly to potential energy—that is, to take the final conditions in part (a) to be the initial conditions in part (b).

Note that, for conservative forces, we do not directly calculate the work they do; rather, we consider their effects through their corresponding potential energies, just as we did in Example 2.1.5.1. Note also that we do not consider details of the path taken—only the starting and ending points are important (as long as the path is not impossible). This assumption is usually a tremendous simplification, because the path may be complicated and forces may vary along the way.

Conservation of Total Energy

Energy, as we have noted, is conserved, making it one of the most important physical quantities in nature. The **law of conservation of energy** can be stated as follows:

Total energy is constant in any process. It may change in form or be transferred from one system to another, but the total remains the same.

We have explored some forms of energy and some ways it can be transferred from one system to another, in the case of work done by conservative forces transforming between potential energy and kinetic energy. Together, these make up the mechanical energy (**KE+PE**), and the mechanical energy is not always conserved, because it can be transformed into other forms of energy through work done by non-conservative forces. Here, we list some of the many forms energy can take.

Some of the Many Forms of Energy

Here are some of the many forms of energy. You probably have heard of some of these before; many of these will be covered in later chapters, but let us detail a few here. **Electrical energy** is a common form that is converted to many other forms and does work in a wide range of practical situations. Fuels, such as gasoline and food, carry **chemical energy** that can be transferred to a system through oxidation. Chemical fuel can also produce electrical energy, such as in batteries. Batteries can in turn produce light, which is a very pure form of energy. Most energy sources on Earth are in fact stored energy from the energy we receive from the Sun. We sometimes refer to this as **radiant energy**, or electromagnetic radiation, which includes visible light, infrared, and ultraviolet radiation. **Nuclear energy** comes from processes that convert measurable amounts of mass into energy. Nuclear energy is transformed into the energy of sunlight, into electrical energy in power plants, and into the energy of the heat transfer and blast in weapons. Atoms and molecules inside all objects are in random motion. This internal mechanical energy from the random motions is called **thermal energy**, because it is related to the temperature of the object. These and all other forms of energy can be converted into one another and can do work.

PROBLEM-SOLVING STRATEGIES FOR ENERGY PROBLEMS

You will find the following problem-solving strategies useful whenever you deal with energy. The strategies help in organizing and reinforcing energy concepts. In fact, they are used in the examples presented in this chapter. The familiar general problem-solving strategies presented earlier—involving identifying physical principles, knowns, and unknowns, checking units, and so on—continue to be relevant here.

Step 1. Determine the system of interest and identify what information is given and what quantity is to be calculated. A sketch will help.

Step 2. Examine all the forces involved and determine whether you know or are given the potential energy from the work done by the forces. Then use step 3 or step 4.

Step 3. If you know the potential energies for the forces that enter into the problem, then forces are all conservative, and you can apply conservation of mechanical energy simply in terms of potential and kinetic energy. Use $mathrmKE_{\mathrm{i}} + \mathrm{PE}_{\mathrm{i}} = \mathrm{KE}_{\mathrm{f}} + \mathrm{PE}_{\mathrm{f}}.$ or $\Delta\mathrm{KE} + \Delta\mathrm{PE} = 0$ as starting point and use formulas for kinetic energy and potential energies.

Step 4. If you know that the total mechanical energy changes (usually there will be some hint of work done by non-conservative forces, energy input from other forms of energy, or even an explicit statement of how much work is done on the system, changing its total mechanical energy), then you can use the information given in the problem to say that, the change of total mechanical energy is due to these non-conservative forces: $\Delta\mathrm{KE} + \Delta\mathrm{PE} = W_{\mathrm{nc}}$. Make sure W_{nc} does not include work done by conservative forces, since those are already accounted for in the potential energy.

Step 5. You have already identified the types of work and energy involved (in step 2). Before solving for the unknown, *eliminate terms wherever possible* to simplify the algebra. For example, choose $h = 0$ at either the initial or final point, so that PE_{g} is zero there. Then solve for the unknown in the customary manner.

Step 6. *Check the answer to see if it is reasonable*. Once you have solved a problem, reexamine the forms of work and energy to see if you have set up the conservation of energy equation correctly. For example, work done against friction should be negative, potential energy at the bottom of a hill should be less than that at the top, and so on. Also check to see that the numerical value obtained is reasonable. For example, the final speed of a skateboarder who coasts down a 3-m-high ramp could reasonably be 20 km/h, but *not* 80 km/h.

Transformation of Energy

The transformation of energy from one form into others is happening all the time. The chemical energy in food is converted into thermal energy through metabolism; light energy is converted into chemical energy through photosynthesis. In a larger example, the chemical energy contained in coal is converted into thermal energy as it burns to turn water into steam in a boiler. This thermal energy in the steam in turn is converted to mechanical energy as it spins a turbine, which is connected to a generator to produce electrical energy. (In all of these examples, not all of the initial energy is converted into the forms mentioned. This important point is discussed later in this section.)

Another example of energy conversion occurs in a solar cell. Sunlight impinging on a solar cell (see Figure 2.1.5.2) produces electricity, which in turn can be used to run an electric motor. Energy is converted from the primary source of solar energy into electrical energy and then into mechanical energy.



Figure 2.1.5.2: Solar energy is converted into electrical energy by solar cells, which is used to run a motor in this solar-power aircraft. (credit: NASA)

Section Summary

- A conservative force is one for which work depends only on the starting and ending points of a motion, not on the path taken.
- We can define potential energy (**PE**) for any conservative force, just as we defined PE_g for the gravitational force.
- Mechanical energy is defined to be **KE+PE** for a conservative force.
- When only conservative forces act on and within a system, the total mechanical energy is conserved, $\Delta\mathrm{KE} + \Delta\mathrm{PE} = 0$.
- The law of conservation of energy states that the total energy, including the mechanical energy and other forms of energy, is constant in any process. Energy may change in form or be transferred from one system to another, but the total remains the same.
- Commonly encountered forms of energy include electric energy, chemical energy, radiant energy, nuclear energy, and thermal energy.

Glossary

conservative force
: a force that is a function of position alone, with the result that the work done by the force depends only on the starting and ending points of a motion and not on the particular path taken

potential energy
: energy due to position, shape, or configuration

conservation of mechanical energy
: the rule that the sum of the kinetic energies and potential energies remains constant if only conservative forces act on and within a system

mechanical energy
: the sum of kinetic energy and potential energy

law of conservation of energy
: the general law that total energy is constant in any process; energy may change in form or be transferred from one system to another, but the total remains the same

electrical energy
: the energy carried by a flow of charge

chemical energy
: the energy in a substance stored in the bonds between atoms and molecules that can be released in a chemical reaction

radiant energy
: the energy carried by electromagnetic waves

nuclear energy
: energy released by changes within atomic nuclei, such as the fusion of two light nuclei or the fission of a heavy nucleus

thermal energy
: the energy within an object due to the random motion of its atoms and molecules that accounts for the object's temperature

This page titled 2.1.5: Conservative Forces, Potential Energy, and Conservation of Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 4.5: Conservative Forces, Potential Energy, and Conservation of Energy by Andrew Park is licensed CC BY 4.0.

2.1.6: Spring Potential Energy

Learning Objectives

- Explain the work done in deforming a spring.
- Describe the potential energy stored in a deformed spring.

Hooke's Law, $F = -kx$, describes force exerted by a spring being deformed. Here, F is the restoring force, x is the displacement from equilibrium or **deformation**, and k is a constant related to the difficulty in deforming the system. The minus sign indicates the restoring force is in the direction opposite to the displacement.

In order to produce a deformation, work must be done. That is, a force must be exerted through a distance, whether you pluck a guitar string or compress a car spring. If the only result is deformation, and no work goes into thermal, sound, or kinetic energy, then all the work is initially stored in the deformed object as some form of potential energy. The potential energy stored in a spring is $\mathrm{PE_{el}} = \frac{1}{2}kx^2$. Here, we generalize the idea to elastic potential energy for a deformation of any system that can be described by Hooke's law. Hence,

$$\mathrm{PE_{el}} = \frac{1}{2}kx^2,$$

where $\mathrm{PE_{el}}$ is the **elastic potential energy** stored in any deformed system that obeys Hooke's law and has a displacement x from equilibrium and a force constant k.

It is possible to find the work done in deforming a system in order to find the energy stored. This work is performed by an applied force F_{app}. The applied force is exactly opposite to the restoring force (action-reaction), and so $F_{\mathrm{app}} = kx$. Figure 2.1.6.1 shows a graph of the applied force versus deformation x for a system that can be described by Hooke's law. Work done on the system is force multiplied by distance, which equals the area under the curve or $(1/2)kx^2$ (Method A in the figure). Another way to determine the work is to note that the force increases linearly from 0 to kx, so that the average force is $(1/2)kx$, the distance moved is x, and thus $W = F_{\mathrm{app}}d = [(1/2)kx](x) = (1/2)kx^2$ (Method B in the figure).

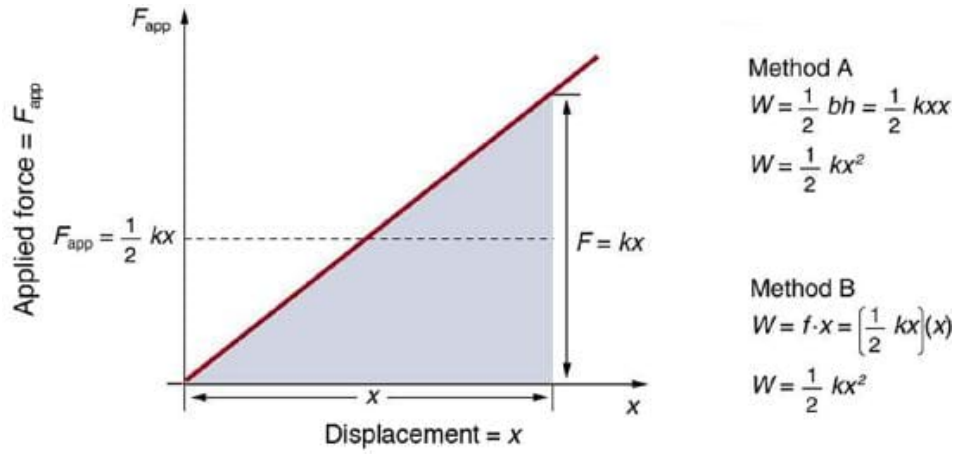


Figure 2.1.6.1: A graph of applied force versus distance for the deformation of a system that can be described by Hooke's law is displayed. The work done on the system equals the area under the graph or the area of the triangle, which is half its base multiplied by its height, or $W = (1/2)kx^2$.

Example 2.1.6.1: Calculating Stored Energy: A Tranquilizer Gun Spring

We can use a toy gun's spring mechanism to ask and answer two simple questions: (a) How much energy is stored in the spring of a tranquilizer gun that has a force constant of 50.0 N/m and is compressed 0.150 m? (b) If you neglect friction and the mass of the spring, at what speed will a 2.00-g projectile be ejected from the gun?

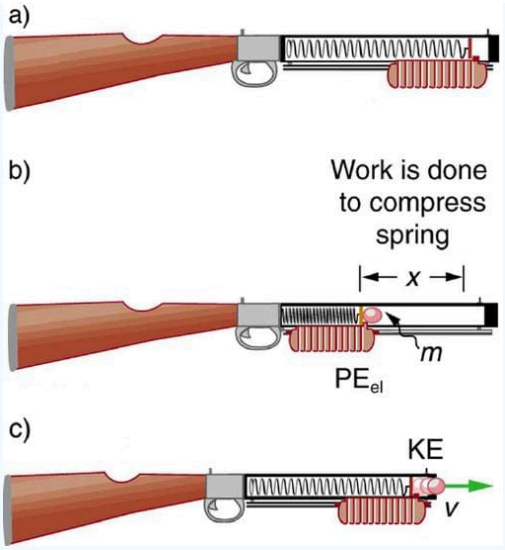


Figure 2.1.6.2: (a) In this image of the gun, the spring is uncompressed before being cocked. (b) The spring has been compressed a distance x, and the projectile is in place. (c) When released, the spring converts elastic potential energy $\mathrm{PE_{el}}$ into kinetic energy.

Strategy for a

(a): The energy stored in the spring can be found directly from elastic potential energy equation, because k and x are given.

Solution for a

Entering the given values for k and x yields

$$\begin{aligned}\mathrm{PE_{el}} &= \frac{1}{2}kx^2 = \frac{1}{2}(50.0\ \mathrm{N/m})(0.150\ \mathrm{m})^2 = 0.563\ \mathrm{N\cdot m} \\ &= 0.563\ \mathrm{J}\end{aligned}$$

Strategy for b

Because there is no friction, the potential energy is converted entirely into kinetic energy. The expression for kinetic energy can be solved for the projectile's speed.

Solution for b

1. Identify known quantities:

$$\mathrm{KE_f} = \mathrm{PE_{el}} \text{ or } 1/2mv^2 = (1/2)kx^2 = \mathrm{PE_{el}} = 0.563\ \mathrm{J}$$

2. Solve for v:

$$v = \left[\frac{2\mathrm{PE_{el}}}{m}\right]^{1/2} = \left[\frac{2(0.563\ \mathrm{J})}{0.002\ \mathrm{kg}}\right]^{1/2} = 23.7(\ \mathrm{J/kg})^{1/2}$$

3. Convert units: 23.7m/s23.7 m/s

Discussion

(a) and (b): This projectile speed is impressive for a tranquilizer gun (more than 80 km/h). The numbers in this problem seem reasonable. The force needed to compress the spring is small enough for an adult to manage, and the energy imparted to the dart is small enough to limit the damage it might do. Yet, the speed of the dart is great enough for it to travel an acceptable distance.

Exercise 2.1.6.1

Envision holding the end of a ruler with one hand and deforming it with the other. When you let go, you can see the oscillations of the ruler. In what way could you modify this simple experiment to increase the rigidity of the system?

Answer available at the end of this section.

Exercise 2.1.6.2

If you apply a deforming force on an object and let it come to equilibrium, what happened to the work you did on the system?

Answer available at the end of this section.

Section Summary

- Hooke's law describes force exerted by a spring being deformed,

$$F = -kx,$$

  where F is the restoring force, x is the displacement from equilibrium or deformation, and k is the force constant of the system.

- Elastic potential energy $\mathrm{PE}_{\mathrm{el}}$ stored in the deformation of a system that can be described by Hooke's law is given by

$$\mathrm{PE}_{\mathrm{el}} = (1/2)kx^2.$$

Glossary

deformation
: displacement from equilibrium

elastic potential energy
: potential energy stored as a result of deformation of an elastic object, such as the stretching of a spring

Answers

Exercise 2.1.6.1

You could hold the ruler at its midpoint so that the part of the ruler that oscillates is half as long as in the original experiment.

Exercise 2.1.6.2

It was stored in the object as potential energy.

This page titled 2.1.6: Spring Potential Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 4.6: Spring Potential Energy by Andrew Park is licensed CC BY 4.0.

2.1.7: Power

Learning Objectives

- Understand the relationship among power, energy, and work done.
- Examine power consumption and calculations of the cost of energy consumed.

What is Power?

Power—the word conjures up many images: a professional football player muscling aside his opponent, a dragster roaring away from the starting line, a volcano blowing its lava into the atmosphere, or a rocket blasting off, as in Figure 2.1.7.1.



Figure 2.1.7.1: This powerful rocket on the Space Shuttle *Endeavor* did work and consumed energy at a very high rate. (credit: NASA)

These images of power have in common the rapid performance of work, consistent with the scientific definition of **power** (**P**) as the rate at which work is done.

Definition: POWER

Power is the rate at which work is done.

$$P = \frac{W}{t}$$

The SI unit for power is the **watt** (**W**), where 1 watt equals 1 joule/second (1 W=1 J/s).

Because work is energy transfer, power is also the rate at which energy is expended. A 60-W light bulb, for example, expends 60 J of energy per second. Great power means a large amount of work or energy developed in a short time. For example, when a powerful car accelerates rapidly, it does a large amount of work and consumes a large amount of fuel in a short time.

Calculating Power from Energy

Example 2.1.7.1: Calculating the Power to Climb Stairs

What is the power output for a 60.0-kg woman who runs up a 3.00 m high flight of stairs in 3.50 s, starting from rest but having a final speed of 2.00 m/s? (See Figure 2.1.7.2.)

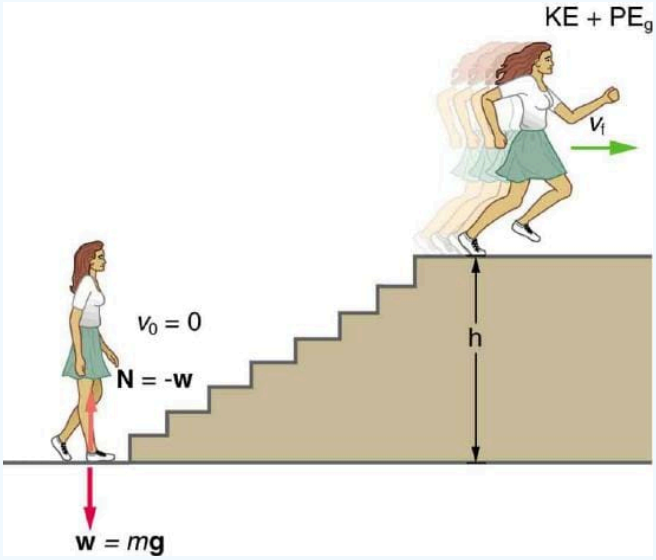


Figure 2.1.7.2: When this woman runs upstairs starting from rest, she converts the chemical energy originally from food into kinetic energy and gravitational potential energy. Her power output depends on how fast she does this.

Strategy and Concept

The work going into mechanical energy is **W= KE + PE**. At the bottom of the stairs, we take both **KE** and $\mathrm{PE_g}$ as initially zero; thus, $W = \mathrm{KE_f} + \mathrm{PE_g} = \frac{1}{2}mv_f^2 + mgh$, where h is the vertical height of the stairs. Because all terms are given, we can calculate **W** and then divide it by time to get power.

Solution

Substituting the expression for **W** into the definition of power given in the previous equation, $P = W/t$ yields

$$P = \frac{W}{t} = \frac{\frac{1}{2}mv_f^2 + mgh}{t}.$$

Entering known values yields

$$\begin{aligned} P &= \frac{0.5(60.0\ \mathrm{kg})(2.00\ \mathrm{m/s})^2 + (60.0\ \mathrm{kg})\,(9.80\ \mathrm{m/s^2})\,(3.00\ \mathrm{m})}{3.50\ \mathrm{s}} \\ &= \frac{120\ \mathrm{J} + 1764\ \mathrm{J}}{3.50\ \mathrm{s}} \\ &= 538\ \mathrm{W}. \end{aligned}$$

Discussion

The woman does 1764 J of work to move up the stairs compared with only 120 J to increase her kinetic energy; thus, most of her power output is required for climbing rather than accelerating.

It is impressive that this woman's useful power output is slightly less than 1 **horsepower** (1 hp=746 W)! People can generate more than a horsepower with their leg muscles for short periods of time by rapidly converting available blood sugar and oxygen into work output. (A horse can put out 1 hp for hours on end.) Once oxygen is depleted, power output decreases and the person begins to breathe rapidly to obtain oxygen to metabolize more food—this is known as the *aerobic* stage of exercise. If the woman

climbed the stairs slowly, then her power output would be much less, although the amount of work done would be the same.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—MEASURE YOUR POWER RATING

Determine your own power rating by measuring the time it takes you to climb a flight of stairs. We will ignore the gain in kinetic energy, as the above example showed that it was a small portion of the energy gain. Don't expect that your output will be more than about 0.5 hp.

Examples of Power

Examples of power are limited only by the imagination, because there are as many types as there are forms of work and energy. (See Table 2.1.7.1 for some examples.) Sunlight reaching Earth's surface carries a maximum power of about 1.3 kilowatts per square meter ($\mathrm{kW/m^2}$). A tiny fraction of this is retained by Earth over the long term. Our consumption rate of fossil fuels is far greater than the rate at which they are stored, so it is inevitable that they will be depleted. Power implies that energy is transferred, perhaps changing form. It is never possible to change one form completely into another without losing some of it as thermal energy. For example, a 60-W incandescent bulb converts only 5 W of electrical power to light, with 55 W dissipating into thermal energy. Furthermore, the typical electric power plant converts only 35 to 40% of its fuel into electricity. The remainder becomes a huge amount of thermal energy that must be dispersed as heat transfer, as rapidly as it is created. A coal-fired power plant may produce 1000 megawatts; 1 megawatt (MW) is 10^6 W of electric power. But the power plant consumes chemical energy at a rate of about 2500 MW, creating heat transfer to the surroundings at a rate of 1500 MW. (See Figure 2.1.7.3.)



Figure 2.1.7.3: Tremendous amounts of electric power are generated by coal-fired power plants such as this one in China, but an even larger amount of power goes into heat transfer to the surroundings. The large cooling towers here are needed to transfer heat as rapidly as it is produced. The transfer of heat is not unique to coal plants but is an unavoidable consequence of generating electric power from any fuel—nuclear, coal, oil, natural gas, or the like. (credit: Kleinolive, Wikimedia Commons)

Table 2.1.7.1*: Power Output or Consumption*

Object or Phenomenon	Power in Watts
Supernova (at peak)	5×10^{37}
Milky Way galaxy	10^{37}
Crab Nebula pulsar	10^{28}

Object or Phenomenon	Power in Watts
The Sun	4×10^{26}
Volcanic eruption (maximum)	4×10^{15}
Lightning bolt	2×10^{12}
Nuclear power plant (total electric and heat transfer)	3×10^{9}
Aircraft carrier (total useful and heat transfer)	10^{8}
Dragster (total useful and heat transfer)	2×10^{6}
Car (total useful and heat transfer)	8×10^{4}
Football player (total useful and heat transfer)	5×10^{3}
Clothes dryer	4×10^{3}
Person at rest (all heat transfer)	100
Typical incandescent light bulb (total useful and heat transfer)	60
Heart, person at rest (total useful and heat transfer)	8
Electric clock	3
Pocket calculator	10^{-3}

Power and Energy Consumption

We usually have to pay for the energy we use. It is interesting and easy to estimate the cost of energy for an electrical appliance if its power consumption rate and time used are known. The higher the power consumption rate and the longer the appliance is used, the greater the cost of that appliance. The power consumption rate is $P = W/t = E/t$, where ***E*** is the energy supplied by the

electricity company. So the energy consumed over a time t is

$$E = Pt.$$

Electricity bills state the energy used in units of **kilowatt-hours** ($\mathrm{kW \cdot h}$), which is the product of power in kilowatts and time in hours. This unit is convenient because electrical power consumption at the kilowatt level for hours at a time is typical.

Example 2.1.7.2: Calculating Energy Costs

What is the cost of running a 0.200-kW computer 6.00 h per day for 30.0 d if the cost of electricity is \$0.120 per $\mathrm{kW \cdot h}$?

Strategy

Cost is based on energy consumed; thus, we must find ***E*** from $E = Pt$ and then calculate the cost. Because electrical energy is expressed in $\mathrm{kW \cdot h}$, at the start of a problem such as this it is convenient to convert the units into kW and hours.

Solution

The energy consumed in $\mathrm{kW \cdot h}$ is

$$\begin{aligned} E = Pt &= (0.200\ \mathrm{kW})(6.00\ \mathrm{h/d})(30.0\ \mathrm{d}) \\ &= 36.0\ \mathrm{kW \cdot h}, \end{aligned}$$

and the cost is simply given by

$$\mathrm{cost} = (36.0\ \mathrm{kW \cdot h})(\$0.120\ \mathrm{per\ kW \cdot h}) = \$4.32\ \mathrm{per\ month}.$$

Discussion

The cost of using the computer in this example is neither exorbitant nor negligible. It is clear that the cost is a combination of power and time. When both are high, such as for an air conditioner in the summer, the cost is high.

The motivation to save energy has become more compelling with its ever-increasing price. Armed with the knowledge that energy consumed is the product of power and time, you can estimate costs for yourself and make the necessary value judgments about where to save energy. Either power or time must be reduced. It is most cost-effective to limit the use of high-power devices that normally operate for long periods of time, such as water heaters and air conditioners. This would not include relatively high power devices like toasters, because they are on only a few minutes per day. It would also not include electric clocks, in spite of their 24-hour-per-day usage, because they are very low power devices. It is sometimes possible to use devices that have greater efficiencies—that is, devices that consume less power to accomplish the same task. One example is the compact fluorescent light bulb, which produces over four times more light per watt of power consumed than its incandescent cousin.

Modern civilization depends on energy, but current levels of energy consumption and production are not sustainable. The likelihood of a link between global warming and fossil fuel use (with its concomitant production of carbon dioxide), has made reduction in energy use as well as a shift to non-fossil fuels of the utmost importance. Even though energy in an isolated system is a conserved quantity, the final result of most energy transformations is waste heat transfer to the environment, which is no longer useful for doing work. As we will discuss in more detail later, the potential for energy to produce useful work has been “degraded” in the energy transformation.

Section Summary

- Power is the rate at which work is done, or in equation form, for the average power ***P*** for work ***W*** done over a time t, $P = W/t$.
- The SI unit for power is the watt (W), where $1\ \mathrm{W} = 1\ \mathrm{J/s}$.
- The power of many devices such as electric motors is also often expressed in horsepower (hp), where $1\mathrm{hp} = 746\ \mathrm{W}$.

Glossary

power
: the rate at which work is done

watt
: (W) SI unit of power, with $1\ \mathrm{W} = 1\ \mathrm{J/s}$

horsepower
: an older non-SI unit of power, with $1\ \mathrm{hp} = 746\ \mathrm{W}$

kilowatt-hour
: $(\mathrm{kW \cdot h})$unit used primarily for electrical energy provided by electric utility companies

This page titled 2.1.7: Power is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 4.7: Power by Andrew Park is licensed CC BY 4.0.

2.1.8: Work and Energy (Exercise)

Conceptual Questions

3.1: Work- The Scientific Definition

1. Give an example of something we think of as work in everyday circumstances that is not work in the scientific sense. Is energy transferred or changed in form in your example? If so, explain how this is accomplished without doing work.

2. Give an example of a situation in which there is a force and a displacement, but the force does no work. Explain why it does no work.

3. Describe a situation in which a force is exerted for a long time but does no work. Explain.

3.2: Kinetic Energy and the Work-Energy Theorem

4. The person in Figure 2.1.8.1 does work on the lawn mower. Under what conditions would the mower gain energy? Under what conditions would it lose energy?

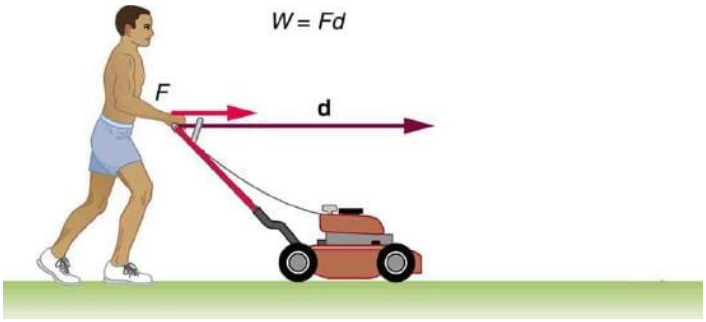


Figure 2.1.8.1

5. Work done on a system puts energy into it. Work done by a system removes energy from it. Give an example for each statement.

6. When solving for speed in Example 4.2.3, we kept only the positive root. Why?

3.3: Gravitational Potential Energy

7. In Example 4.3.2, we calculated the final speed of a roller coaster that descended 20 m in height and had an initial speed of 5 m/s downhill. Suppose the roller coaster had had an initial speed of 5 m/s *uphill* instead, and it coasted uphill, stopped, and then rolled back down to a final point 20 m below the start. We would find in that case that it had the same final speed. Explain in terms of conservation of energy.

8. Does the work you do on a book when you lift it onto a shelf depend on the path taken? On the time taken? On the height of the shelf? On the mass of the book?

3.4: Conservative Forces, Potential Energy, and Conservation of Energy

9. What is a conservative force?

10. The force exerted by a diving board is conservative, provided the internal friction is negligible. Assuming friction is negligible, describe changes in the potential energy of a diving board as a swimmer dives from it, starting just before the swimmer steps on the board until just after his feet leave it.

11. Define mechanical energy. What is the relationship of mechanical energy to nonconservative forces? What happens to mechanical energy if only conservative forces act?

12. What is the relationship of potential energy to conservative force?

13. Consider the following scenario. A car for which friction is *not* negligible accelerates from rest down a hill, running out of gasoline after a short distance. The driver lets the car coast farther down the hill, then up and over a small crest. He then coasts down that hill into a gas station, where he brakes to a stop and fills the tank with gasoline. Identify the forms of energy the car has, and how they are changed and transferred in this series of events. (See Figure 2.1.8.2.)

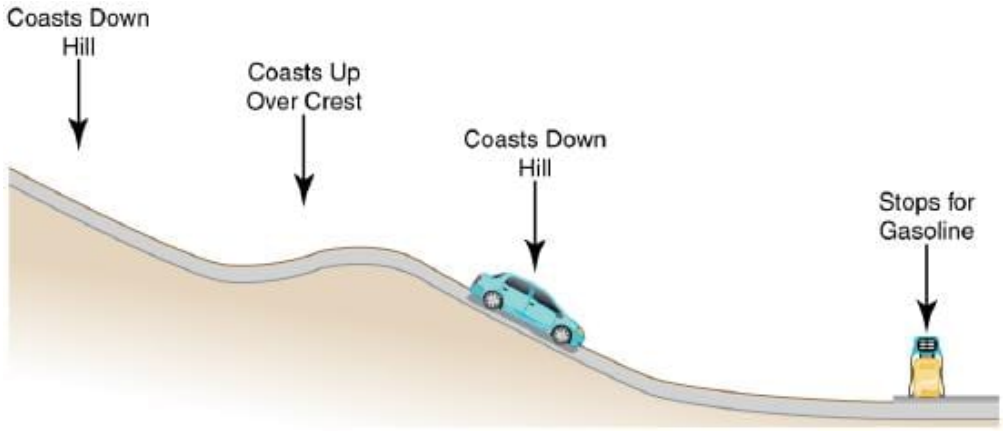


Figure 2.1.8.2: A car experiencing non-negligible friction coasts down a hill, over a small crest, then downhill again, and comes to a stop at a gas station.

14. Describe the energy transfers and transformations for a javelin, starting from the point at which an athlete picks up the javelin and ending when the javelin is stuck into the ground after being thrown

15. List four different forms or types of energy. Give one example of a conversion from each of these forms to another form.

16. List the energy conversions that occur when riding a bicycle.

3.5: Spring Potential Energy

17. Describe a system in which elastic potential energy is stored.

3.6: Power

18. Most electrical appliances are rated in watts. Does this rating depend on how long the appliance is on? (When off, it is a zero-watt device.) Explain in terms of the definition of power.

19. Explain, in terms of the definition of power, why energy consumption is sometimes listed in kilowatt-hours rather than joules. What is the relationship between these two energy units?

20. A spark of static electricity, such as that you might receive from a doorknob on a cold dry day, may carry a few hundred watts of power. Explain why you are not injured by such a spark.

This page titled 2.1.8: Work and Energy (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 4.8: Work and Energy (Exercise) by Andrew Park is licensed CC BY 4.0.

2.2: IMPULSE AND MOMENTUM

2.2.1: Introduction to Linear Momentum and Collisions
2.2.2: Linear Momentum and Force
2.2.3: Impulse
2.2.4: Conservation of Momentum
2.2.5: Elastic Collisions in One Dimension
2.2.6: Inelastic Collisions in One Dimension
2.2.7: Impulse and Momentum (Exercise)

This page titled 2.2: Impulse and Momentum is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

2.2.1: Introduction to Linear Momentum and Collisions



Figure 2.2.1.1: Each rugby player has great momentum, which will affect the outcome of their collisions with each other and the ground. (credit: ozzzie, Flickr)

We use the term momentum in various ways in everyday language, and most of these ways are consistent with its precise scientific definition. We speak of sports teams or politicians gaining and maintaining the momentum to win. We also recognize that momentum has something to do with collisions. For example, looking at the rugby players in the photograph colliding and falling to the ground, we expect their momenta to have great effects in the resulting collisions. Generally, momentum implies a tendency to continue on course—to move in the same direction—and is associated with great mass and speed.

Momentum, like energy, is important because it is conserved. Only a few physical quantities are conserved in nature, and studying them yields fundamental insight into how nature works, as we shall see in our study of momentum.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

This page titled 2.2.1: Introduction to Linear Momentum and Collisions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 5.1: Introduction to Linear Momentum and Collisions by Andrew Park is licensed CC BY 4.0.

2.2.2: Linear Momentum and Force

Learning Objectives

- Define linear momentum.
- Explain the relationship between momentum and force.
- State Newton's second law of motion in terms of momentum.
- Calculate momentum given mass and velocity.

Linear Momentum

The scientific definition of linear momentum is consistent with most people's intuitive understanding of momentum: a large, fast-moving object has greater momentum than a smaller, slower object. **Linear momentum** is defined as the product of a system's mass multiplied by its velocity. In symbols, linear momentum is expressed as

$$\boldsymbol{p} = m\boldsymbol{v}$$

Momentum is directly proportional to the object's mass and also its velocity. Thus the greater an object's mass or the greater its velocity, the greater its momentum. Momentum $\boldsymbol{p}$ is a vector having the same direction as the velocity $\boldsymbol{v}$. The SI unit for momentum is $\mathrm{kg \cdot m/s}$.

Definition: LINEAR MOMENTUM

Linear momentum is defined as the product of a system's mass multiplied by its velocity:

$$\boldsymbol{p} = m\boldsymbol{v}.$$

Example 2.2.2.1: Calculating Momentum: A Football Player and a Football

(a) Calculate the momentum of a 110-kg football player running at 8.00 m/s. (b) Compare the player's momentum with the momentum of a hard-thrown 0.410-kg football that has a speed of 25.0 m/s.

Strategy

No information is given regarding direction, and so we can calculate only the magnitude of the momentum, p. (As usual, a symbol that is in italics is a magnitude, whereas one that is italicized and is boldfaced or has an arrow is a vector.) In both parts of this example, the magnitude of momentum can be calculated directly from the definition of momentum given in the equation, which becomes

$$p = mv$$

when only magnitudes are considered.

Solution for (a)

To determine the momentum of the player, substitute the known values for the player's mass and speed into the equation.

$$p_{\text{player}} = (110\ \mathrm{kg})(8.00\ \mathrm{m/s}) = 880\ \mathrm{kg \cdot m/s}$$

Solution for (b)

To determine the momentum of the ball, substitute the known values for the ball's mass and speed into the equation.

$$p_{\text{ball}} = (0.410\ \text{kg})(25.0\ \text{m/s}) = 10.3\ \text{kg} \cdot \text{m/s}$$

The ratio of the player's momentum to that of the ball is

$$\frac{p_{\text{player}}}{p_{\text{ball}}} = \frac{880}{10.3} = 85.9.$$

Discussion

Although the ball has greater velocity, the player has a much greater mass. Thus the momentum of the player is much greater than the momentum of the football, as you might guess. As a result, the player's motion is only slightly affected if he catches the ball. We shall quantify what happens in such collisions in terms of momentum in later sections.

Momentum and Newton's Second Law

The importance of momentum, unlike the importance of energy, was recognized early in the development of classical physics. Momentum was deemed so important that it was called the "quantity of motion." Newton actually stated his **second law of motion** in terms of momentum: The net external force equals the change in momentum of a system divided by the time over which it changes. Using symbols, this law is

$$\boldsymbol{F}_{\text{net}} = \frac{\Delta \boldsymbol{p}}{\Delta t},$$

where $\boldsymbol{F}_{\text{net}}$ is the net external force, $\Delta \boldsymbol{p}$ is the change in momentum, and Δt is the change in time.

NEWTON'S SECOND LAW OF MOTION IN TERMS OF MOMENTUM

The net external force equals the change in momentum of a system divided by the time over which it changes: $\boldsymbol{F}_{\text{net}} = \Delta \boldsymbol{p}/\Delta t$.

MAKING CONNECTIONS: FORCE AND MOMENTUM

Force and momentum are intimately related. Force acting over time can change momentum, and Newton's second law of motion, can be stated in its most broadly applicable form in terms of momentum. Momentum continues to be a key concept in the study of atomic and subatomic particles in quantum mechanics.

This statement of Newton's second law of motion includes the more familiar $\boldsymbol{F}_{\text{net}} = m\boldsymbol{a}$ as a special case. We can derive this form as follows. First, note that the change in momentum $\Delta \boldsymbol{p}$ is given by

$$\Delta \boldsymbol{p} = \Delta(m\boldsymbol{v}).$$

If the mass of the system is constant, then

$$\Delta(m\boldsymbol{v}) = m\Delta \boldsymbol{v}.$$

So that for constant mass, Newton's second law of motion becomes

$$\boldsymbol{F}_{\text{net}} = \frac{\Delta \boldsymbol{p}}{\Delta t} = m\frac{\Delta \boldsymbol{v}}{\Delta t}$$

Because $\frac{\Delta v}{\Delta t} = \boldsymbol{a}$, we get the familiar equation

$$\boldsymbol{F}_{\text{net}} = m\boldsymbol{a}$$

when the mass of the system is constant.

Newton's second law of motion stated in terms of momentum is more generally applicable because it can be applied to systems

where the mass is changing, such as rockets, as well as to systems of constant mass. We will consider systems with varying mass in some detail; however, the relationship between momentum and force remains useful when mass is constant, such as in the following example.

Example 2.2.2.2: Calculating Force: Venus Williams' Racquet

During the 2007 French Open, Venus Williams hit the fastest recorded serve in a premier women's match, reaching a speed of 58 m/s (209 km/h). What is the average force exerted on the 0.057-kg tennis ball by Venus Williams' racquet, assuming that the ball's speed just after impact is 58 m/s, that the initial horizontal component of the velocity before impact is negligible, and that the ball remained in contact with the racquet for 5.0 ms (milliseconds)?

Strategy

This problem involves only one dimension because the ball starts from having no horizontal velocity component before impact. Newton's second law stated in terms of momentum is then written as $\boldsymbol{F}_{\text{net}} = \Delta\boldsymbol{p}/\Delta t$.

As noted above, when mass is constant, the change in momentum is given by

$$\Delta p = m\Delta v = m\left(v_{\text{f}} - v_{\text{i}}\right).$$

In this example, the velocity just after impact and the change in time are given; thus, once Δp is calculated, $F_{\text{net}} = \frac{\Delta p}{\Delta t}$ can be used to find the force.

Solution

To determine the change in momentum, substitute the values for the initial and final velocities into the equation above.

$$\begin{aligned}\Delta p &= m\left(v_{\text{f}} - v_{\text{i}}\right)\\ &= (0.057\ \text{kg})(58\ \text{m/s} - 0\ \text{m/s})\\ &= 3.306\ \text{kg}\cdot\text{m/s} \approx 3.3\ \text{kg}\cdot\text{m/s}\end{aligned}$$

Now the magnitude of the net external force can determined by using $F_{\text{net}} = \frac{\Delta p}{\Delta t}$:

$$\begin{aligned}F_{\text{net}} &= \frac{\Delta p}{\Delta t} = \frac{3.306\ \text{kg}\cdot\text{m/s}}{5.0\times10^{-3}\ \text{s}}\\ &= 661\ \text{N} \approx 660\ \text{N},\end{aligned}$$

where we have retained only two significant figures in the final step.

Discussion

This quantity was the average force exerted by Venus Williams' racquet on the tennis ball during its brief impact (note that the ball also experienced the 0.56-N force of gravity, but that force was not due to the racquet). This problem could also be solved by first finding the acceleration and then using $F_{\text{net}} = ma$, but one additional step would be required compared with the strategy used in this example.

Section Summary

- Linear momentum (*momentum* for brevity) is defined as the product of a system's mass multiplied by its velocity.
- In symbols, linear momentum $\boldsymbol{p}$ is defined to be

$$\boldsymbol{p} = m\boldsymbol{v},$$

  where mm is the mass of the system and $\boldsymbol{v}$ is its velocity.
- The SI unit for momentum is $\text{kg}\cdot\text{m/s}$.
- Newton's second law of motion in terms of momentum states that the net external force equals the change in momentum of a system divided by the time over which it changes.
- In symbols, Newton's second law of motion is defined to be

$$F_{\text{net}} = \frac{\Delta p}{\Delta t},$$

where F_{net} is the net external force, Δp is the change in momentum, and Δt is the change time.

Glossary

linear momentum
: the product of mass and velocity

second law of motion
: physical law that states that the net external force equals the change in momentum of a system divided by the time over which it changes

This page titled 2.2.2: Linear Momentum and Force is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 5.2: Linear Momentum and Force by Andrew Park is licensed CC BY 4.0.

2.2.3: Impulse

Learning Objectives

- Define impulse.
- Describe effects of impulses in everyday life.
- Calculate average force and impulse given mass, velocity, and time.

The effect of a force on an object depends on how long it acts, as well as how great the force is. In previous examples, a very large force acting for a short time had a great effect on the momentum of the tennis ball. A small force could cause the same **change in momentum**, but it would have to act for a much longer time. For example, if the ball were thrown upward, the gravitational force (which is much smaller than the tennis racquet's force) would eventually reverse the momentum of the ball. Quantitatively, the effect we are talking about is the change in momentum $\Delta \boldsymbol{p}$.

By rearranging the equation $\boldsymbol{F}_{\text{net}} = \frac{\Delta \boldsymbol{p}}{\Delta t}$ to be

$$\Delta \boldsymbol{p} = \boldsymbol{F}_{\text{net}} \Delta t,$$

we can see how the change in momentum equals the average net external force multiplied by the time this force acts. The quantity $\boldsymbol{F}_{\text{net}}\ \Delta t$, force times duration, is given the name **impulse**. Net impulse on an object gives its change in momentum.

IMPULSE: CHANGE IN MOMENTUM

Change in momentum equals the average net external force multiplied by the time this force acts: $\Delta \boldsymbol{p} = \boldsymbol{F}_{\text{net}} \Delta t$.

The quantity $\boldsymbol{F}_{\text{net}}\ \Delta t$ is given the name impulse.

There are many ways in which an understanding of impulse can save lives, or at least limbs. The dashboard padding in a car, and certainly the airbags, allow the net force on the occupants in the car to act over a much longer time when there is a sudden stop. The momentum change is the same for an occupant, whether an air bag is deployed or not, but the force (to bring the occupant to a stop) will be much less if it acts over a larger time. Cars today have many plastic components. One advantage of plastics is their lighter weight, which results in better gas mileage. Another advantage is that a car will crumple in a collision, especially in the event of a head-on collision. A longer collision time means the force on the car will be less. Deaths during car races decreased dramatically when the rigid frames of racing cars were replaced with parts that could crumple or collapse in the event of an accident.

Bones in a body will fracture if the force on them is too large. If you jump onto the floor from a table, the force on your legs can be immense if you land stiff-legged on a hard surface. Rolling on the ground after jumping from the table, or landing with a parachute, extends the time over which the force (on you from the ground) acts.

Our definition of impulse includes an assumption that the force is constant over the time interval Δt. *Forces are usually not constant*. Forces vary considerably even during the brief time intervals considered. It is, however, possible to find an average effective force F_{eff} that produces the same result as the corresponding time-varying force. Figure $2.2.3.1$ shows a graph of what an actual force looks like as a function of time for a ball bouncing off the floor. The area under the curve has units of momentum and is equal to the impulse or change in momentum between times t_1 and t_2. That area is equal to the area inside the rectangle bounded by F_{eff}, t_1, and t_2. Thus the impulses and their effects are the same for both the actual and effective forces.

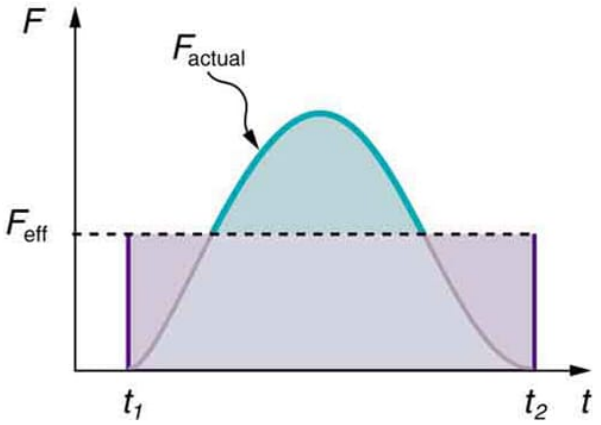


Figure 2.2.3.1: A graph of force versus time with time along the x-axis and force along the y-axis for an actual force and an equivalent effective force. The areas under the two curves are equal.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—HAND MOVEMENT AND IMPULSE

Try catching a ball while "giving" with the ball, pulling your hands toward your body. Then, try catching a ball while keeping your hands still. Hit water in a tub with your full palm. After the water has settled, hit the water again by diving your hand with your fingers first into the water. (Your full palm represents a swimmer doing a belly flop and your diving hand represents a swimmer doing a dive.) Explain what happens in each case and why. Which orientations would you advise people to avoid and why?

MAKING CONNECTIONS: CONSTANT FORCE AND CONSTANT ACCELERATION

The assumption of a constant force in the definition of impulse is analogous to the assumption of a constant acceleration in kinematics. In both cases, nature is adequately described without the use of calculus.

Section Summary

- Impulse is force times duration, and the net impulse, equal to the net external force multiplied by duration, gives the change in momentum:

$$\Delta \boldsymbol{p} = \boldsymbol{F}_{\text{net}}\, \Delta t.$$

- For non-constant net external force, you need to use time-averaged net external force to calculate the impulse over the duration.

Glossary

change in momentum
: the difference between the final and initial momentum; the mass times the change in velocity

impulse
: the average net external force times the time it acts; equal to the change in momentum

This page titled 2.2.3: Impulse is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 5.3: Impulse by Andrew Park is licensed CC BY 4.0.

2.2.4: Conservation of Momentum

Learning Objectives

- Describe the principle of conservation of momentum.
- Explain conservation of momentum with examples.

In the previous sections in this chapter, changes in momentum for an object were considered in terms of impulse and force. But for certain conditions, the total momentum is conserved (stays constant). What are these conditions?

The answer to this question entails considering a sufficiently large system. It is always possible to find a larger system in which total momentum is constant, even if momentum changes for components of the system. If a football player runs into the goalpost in the end zone, there will be a force on him that causes him to bounce backward. However, the Earth also recoils —conserving momentum—because of the force applied to it through the goalpost. Because Earth is many orders of magnitude more massive than the player, its recoil is immeasurably small and can be neglected in any practical sense, but it is real nevertheless.

Consider what happens if the masses of two colliding objects are more similar (i.e, both objects are about the same mass) than the masses of a football player and Earth—for example, one car bumping into another, as shown in Figure 2.2.4.1. Both cars are coasting in the same direction when the lead car (labeled m_2) is bumped by the trailing car (labeled m_1). The only unbalanced force on each car is the force of the collision. (Assume that the effects due to friction are negligible.) Car 1 slows down as a result of the collision, losing some momentum, while car 2 speeds up and gains some momentum. We shall now show that the total momentum of the two-car system remains constant.

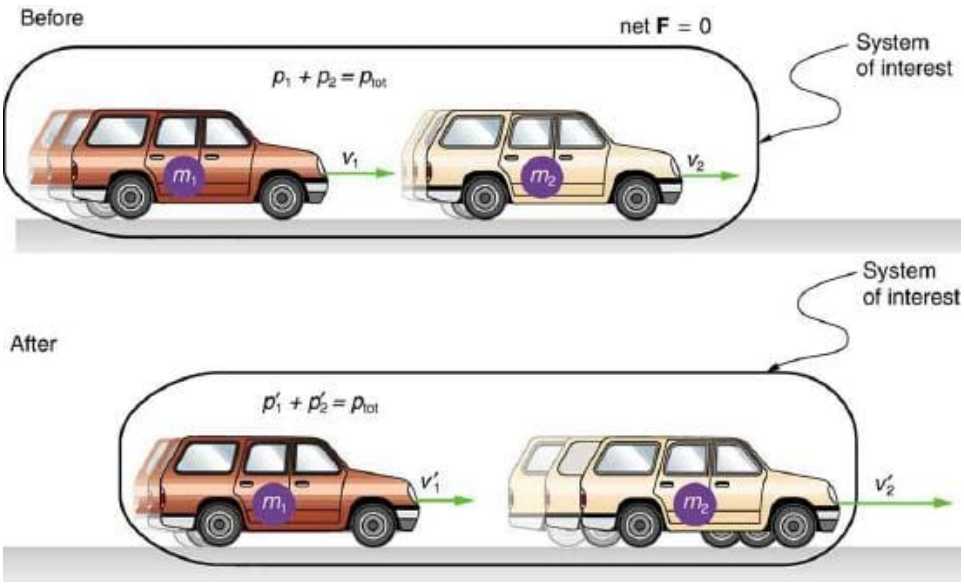


Figure 2.2.4.1: A car of mass m_1 moving with a velocity of v_1 bumps into another car of mass m_2 and velocity v_2 that it is following. As a result, the first car slows down to a velocity of v_1' and the second speeds up to a velocity of v_2'. The momentum of each car is changed, but the total momentum p_{tot} of the two cars is the same before and after the collision (if you assume friction is negligible).

Using the definition of impulse, the change in momentum of car 1 is given by

$$\Delta p_1 = F_1 \Delta t,$$

where F_1 is the force on car 1 due to car 2, and Δt is the time the force acts (the duration of the collision). Similarly, the change in momentum of car 2 is

$$\Delta p_2 = F_2 \Delta t,$$

where F_2 is the force on car 2 due to car 1, and, intuitively, the duration of the collision Δt is the same for both cars. We know from Newton's third law that $F_2 = -F_1$, and so

$$\Delta p_2 = -F_1 \Delta t = -\Delta p_1.$$

Thus, the changes in momentum cancel out, and

$$\Delta p_1 + \Delta p_2 = 0.$$

Because the changes in momentum add to zero, the total momentum of the two-car system is constant. That is,

$$p_1 + p_2 = \text{ constant } .$$

Because of Newton's third law, when two objects in a system interact with each other, it does not change the total momentum of the system. In the absence of an external force (that is, force due to an object outside of the system, so that we can ignore the impulse due to the reaction force), the total momentum of the system is constant, or conserved. We call this the **conservation of momentum principle**.

We have noted that the motion in perpendicular directions—x, y, and —are independent. Momentum conservation follows this paradigm, and components of total momentum along each direction are conserved separately. For example, during projectile motion and where air resistance is negligible, momentum is conserved in the horizontal direction because horizontal forces are zero and momentum is unchanged. But along the vertical direction, the net vertical force is not zero and the momentum of the projectile is not conserved. (See Figure 2.2.4.2.) However, if the momentum of the projectile-Earth system is considered in the vertical direction, we find that the total momentum is conserved.

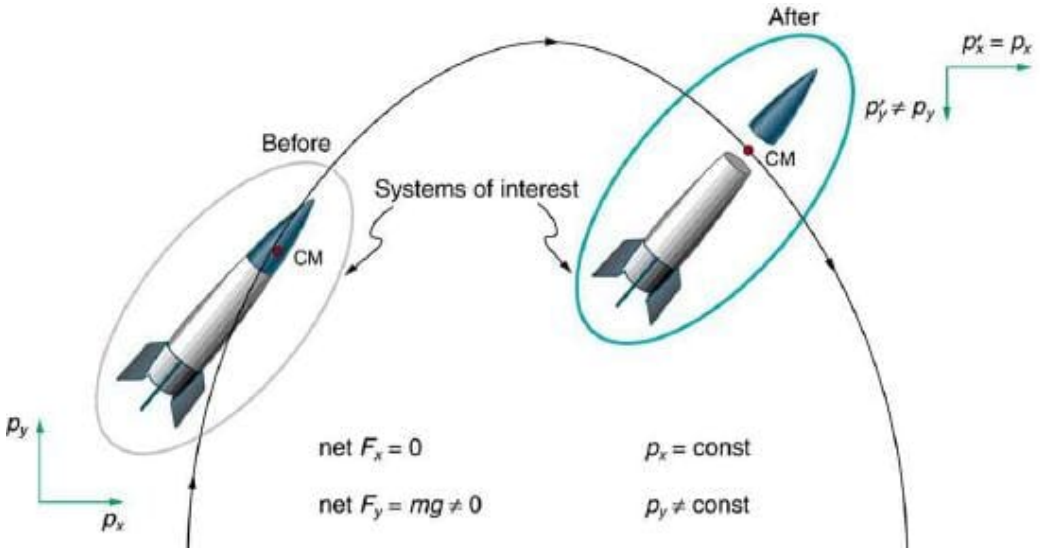


Figure 2.2.4.2: The horizontal component of a projectile's momentum is conserved if air resistance is negligible, even in this case where a space probe separates. The forces causing the separation are internal to the system, so that the net external horizontal force Fx–net is still zero. The vertical component of the momentum is not conserved, because the net vertical force Fy–net is not zero. In the vertical direction, the space probe-Earth system needs to be considered and we find that the total momentum is conserved. The center of mass of the space probe takes the same path it would if the separation did not occur.

The conservation of momentum principle can be applied to systems as different as a comet striking Earth and a gas containing huge numbers of atoms and molecules. Conservation of momentum is violated only when the net external force is not zero. But another larger system can always be considered in which momentum is conserved by simply including the source of the external force. For example, in the collision of two cars considered above, the two-car system conserves momentum while each one-car system does not.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—DROP OF TENNIS BALL AND A BASKETBALL

Hold a tennis ball side by side and in contact with a basketball. Drop the balls together. (Be careful!) What happens? Explain your observations. Now hold the tennis ball above and in contact with the basketball. What happened? Explain your observations. What do you think will happen if the basketball ball is held above and in contact with the tennis ball?

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—TWO TENNIS BALLS IN A BALLISTIC TRAJECTORY

Tie two tennis balls together with a string about a foot long. Hold one ball and let the other hang down and throw it in a ballistic trajectory. Explain your observations. Now mark the center of the string with bright ink or attach a brightly colored sticker to it and throw again. What happened? Explain your observations.

Some aquatic animals such as jellyfish move around based on the principles of conservation of momentum. A jellyfish fills its umbrella section with water and then pushes the water out resulting in motion in the opposite direction to that of the jet of water. Squids propel themselves in a similar manner but, in contrast with jellyfish, are able to control the direction in which they move by aiming their nozzle forward or backward. Typical squids can move at speeds of 8 to 12 km/h.

The ballistocardiograph (BCG) was a diagnostic tool used in the second half of the 20th century to study the strength of the heart. About once a second, your heart beats, forcing blood into the aorta. A force in the opposite direction is exerted on the rest of your body (recall Newton's third law). A ballistocardiograph is a device that can measure this reaction force. This measurement is done by using a sensor (resting on the person) or by using a moving table suspended from the ceiling. This technique can gather information on the strength of the heart beat and the volume of blood passing from the heart. However, the electrocardiogram (ECG or EKG) and the echocardiogram (cardiac ECHO or ECHO; a technique that uses ultrasound to see an image of the heart) are more widely used in the practice of cardiology.

MAKING CONNECTIONS: CONSERVATION OF MOMENTUM AND COLLISION

Conservation of momentum is quite useful in describing collisions. Momentum is crucial to our understanding of atomic and subatomic particles because much of what we know about these particles comes from collision experiments.

Section Summary

- The conservation of momentum principle says that the total momentum is conserved, or constant, in the absence of a net external force, as a direct consequence of Newton's third law.
- During projectile motion and where air resistance is negligible, momentum is conserved in the horizontal direction because horizontal forces are zero.

Glossary

conservation of momentum principle
: when the net external force is zero, the total momentum of the system is conserved or constant

This page titled 2.2.4: Conservation of Momentum is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 5.4: Conservation of Momentum by Andrew Park is licensed CC BY 4.0.

2.2.5: Elastic Collisions in One Dimension

Learning Objectives

- Describe an elastic collision of two objects in one dimension.
- Determine the final velocities in an elastic collision given masses and initial velocities.

Let us consider various types of two-object collisions. These collisions are the easiest to analyze, and they illustrate many of the physical principles involved in collisions. The conservation of momentum principle is very useful here, and it can be used whenever the net external force on a system is zero.

We start with the elastic collision of two objects moving along the same line—a one-dimensional problem. An **elastic collision** is one that also conserves total kinetic energy, in addition to the total momentum. Figure 2.2.5.1 illustrates an elastic collision in which kinetic energy and momentum are conserved.

Truly elastic collisions can only be achieved with subatomic particles, such as electrons striking nuclei. Macroscopic collisions can be very nearly, but not quite, elastic—some kinetic energy is always converted into other forms of energy such as heat transfer due to friction and sound. One macroscopic collision that is nearly elastic is that of two steel blocks on ice. Another nearly elastic collision is that between two carts with spring bumpers on an air track. Icy surfaces and air tracks are nearly frictionless, more readily allowing nearly elastic collisions on them.

Definition: ELASTIC COLLISION

An elastic collision is one that conserves total kinetic energy.

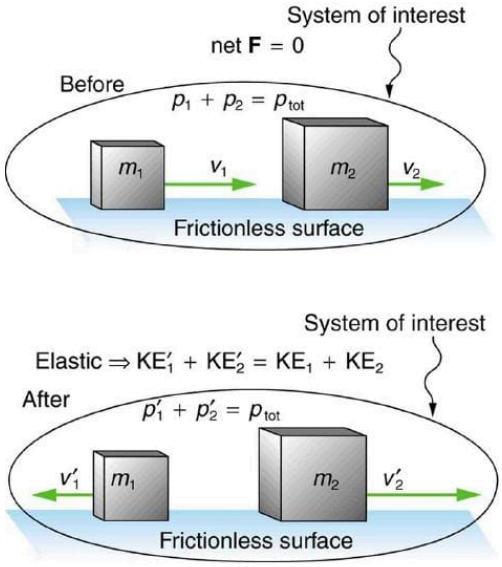


Figure 2.2.5.1: An elastic one-dimensional two-object collision. Momentum and internal kinetic energy are conserved.

Now, to solve problems involving one-dimensional elastic collisions between two objects we can use the equations for conservation of momentum and conservation of internal kinetic energy. First, the equation for conservation of momentum for two objects in a one-dimensional collision is

$$p_1 + p_2 = p'_1 + p'_2 \quad (F_{\text{net}} = 0)$$

or

$$m_1 v_1 + m_2 v_2 = m_1 v_1' + m_2 v_2' \quad (F_{\text{net}} = 0)\,,$$

where the primes (') indicate values after the collision. By definition, an elastic collision conserves total kinetic energy, and so the sum of kinetic energies before the collision equals the sum after the collision. Thus,

$$\frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 = \frac{1}{2} m_1 v_1'^2 + \frac{1}{2} m_2 v_2'^2 \text{ (two-object elastic collision)}$$

expresses the equation for conservation of total kinetic energy in a one-dimensional collision.

Example 2.2.5.1: Calculating Velocities Following an Elastic Collision

Calculate the velocities of two objects following an elastic collision, given that

$$m_1 = 0.500 \text{ kg}, m_2 = 3.50 \text{ kg}, v_1 = 4.00 \text{ m/s}, \text{ and } v_2 = 0.$$

Strategy and Concept

First, visualize what the initial conditions mean—a small object strikes a larger object that is initially at rest. This situation is slightly simpler than the situation shown in Figure 2.2.5.1 where both objects are initially moving. We are asked to find two unknowns (the final velocities v_1' and v_2'). To find two unknowns, we must use two independent equations. Because this collision is elastic, we can use the above two equations. Both can be simplified by the fact that object 2 is initially at rest, and thus $v_2 = 0$. Once we simplify these equations, we combine them algebraically to solve for the unknowns.

Solution

For this problem, note that $v_2 = 0$ and use conservation of momentum. Thus,

$$p_1 = p_1' + p_2'$$

or

$$m_1 v_1 = m_1 v_1' + m_2 v_2'.$$

Using conservation of internal kinetic energy and that $v_2 = 0$,

$$\frac{1}{2} m_1 v_1^2 = \frac{1}{2} m_1 v_1'^2 + \frac{1}{2} m_2 v_2'^2.$$

Solving the first equation (momentum equation) for v_2', we obtain

$$v_2' = \frac{m_1}{m_2} (v_1 - v_1')\,.$$

Substituting this expression into the second equation (internal kinetic energy equation) eliminates the variable v_2', leaving only v_1' as an unknown (the algebra is left as an exercise for the reader). There are two solutions to any quadratic equation; in this example, they are

$$v_1' = 4.00 \text{ m/s}$$

and

$$v_1' = -3.00 \text{ m/s}.$$

As noted when quadratic equations were encountered in earlier chapters, both solutions may or may not be meaningful. In this case, the first solution is the same as the initial condition. The first solution thus represents the situation before the collision and is discarded. The second solution ($v_1' = -3.00$ m/s) is negative, meaning that the first object bounces backward. When this negative value of v_1' is used to find the velocity of the second object after the collision, we get

$$v_2' = \frac{m_1}{m_2}(v_1 - v_1') = \frac{0.500\ \text{kg}}{3.50\ \text{kg}}[4.00 - (-3.00)]\text{m/s}$$

or

$$v_2' = 1.00\ \text{m/s}.$$

Discussion

The result of this example is intuitively reasonable. A small object strikes a larger one at rest and bounces backward. The larger one is knocked forward, but with a low speed. (This is like a compact car bouncing backward off a full-size SUV that is initially at rest.) As a check, try calculating the internal kinetic energy before and after the collision. You will see that the internal kinetic energy is unchanged at 4.00 J. Also check the total momentum before and after the collision; you will find it, too, is unchanged.

The equations for conservation of momentum and internal kinetic energy as written above can be used to describe any one-dimensional elastic collision of two objects. These equations can be extended to more objects if needed.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—ICE CUBES AND ELASTIC COLLISION

Find a few ice cubes which are about the same size and a smooth kitchen tabletop or a table with a glass top. Place the ice cubes on the surface several centimeters away from each other. Flick one ice cube toward a stationary ice cube and observe the path and velocities of the ice cubes after the collision. Try to avoid edge-on collisions and collisions with rotating ice cubes. Have you created approximately elastic collisions? Explain the speeds and directions of the ice cubes using momentum.

Section Summary

- An elastic collision is one that conserves total kinetic energy.
- Conservation of kinetic energy and momentum together allow the final velocities to be calculated in terms of initial velocities and masses in one dimensional two-body collisions.

Glossary

elastic collision
: a collision that also conserves total kinetic energy, in addition to the total momentum

This page titled 2.2.5: Elastic Collisions in One Dimension is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 5.5: Elastic Collisions in One Dimension by Andrew Park is licensed CC BY 4.0.

2.2.6: Inelastic Collisions in One Dimension

Learning Objectives

- Define inelastic collision.
- Explain perfectly inelastic collision.
- Apply an understanding of collisions to everyday situations.
- Determine recoil velocity and loss in kinetic energy given mass and initial velocity.

We have seen that in an elastic collision, total kinetic energy is conserved. An **inelastic collision** is one in which the total kinetic energy changes (it is not conserved). This lack of conservation means that the forces between colliding objects may remove or add total kinetic energy. Work done by internal forces may change the forms of energy within a system. For inelastic collisions, such as when colliding objects stick together, this internal work may transform some kinetic energy into thermal energy. Or it may convert stored energy into total kinetic energy, such as when exploding bolts separate a satellite from its launch vehicle.

Definition: INELASTIC COLLISION

An inelastic collision is one in which the total kinetic energy changes (it is not conserved).

Figure 2.2.6.1 shows an example of an inelastic collision. Two objects that have equal masses head toward one another at equal speeds and then stick together. Their total kinetic energy is initially $\frac{1}{2}mv^2 + \frac{1}{2}mv^2 = mv^2$. The two objects come to rest after sticking together, conserving total momentum. And the total kinetic energy is zero after this collision. A collision in which the objects stick together is called a **perfectly inelastic collision** because it reduces the total kinetic energy as much as possible, while conserving total momentum.

Definition: PERFECTLY INELASTIC COLLISION

A collision in which the objects stick together is sometimes called “perfectly inelastic.”

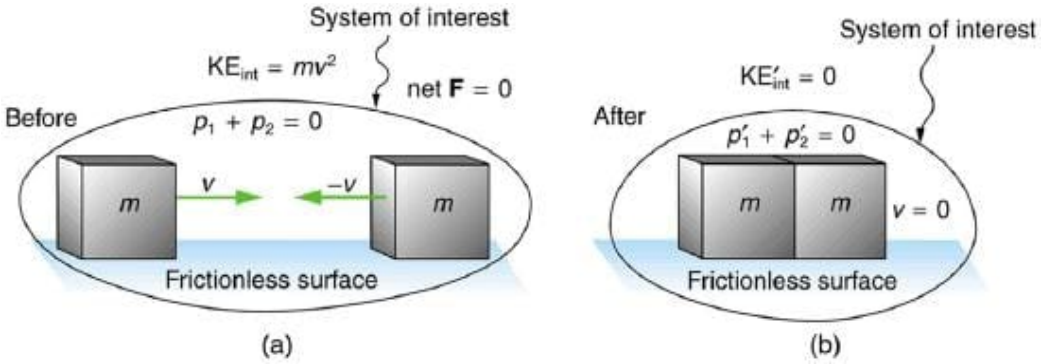


Figure 2.2.6.1: An inelastic one-dimensional two-object collision. Momentum is conserved, but total kinetic energy is not conserved. (a) Two objects of equal mass initially head directly toward one another at the same speed. (b) The objects stick together (a perfectly inelastic collision), and so their final velocity is zero.

Example 2.2.6.1: Calculating Velocity and Change in Kinetic Energy: Inelastic Collision of a Puck and a Goalie

(a) Find the recoil velocity of a 70.0-kg ice hockey goalie, originally at rest, who catches a 0.150-kg hockey puck slapped at him at a velocity of 35.0 m/s. (b) How much kinetic energy is lost during the collision? Assume friction between the ice and the puck-goalie system is negligible. (See Figure 2.2.6.2)

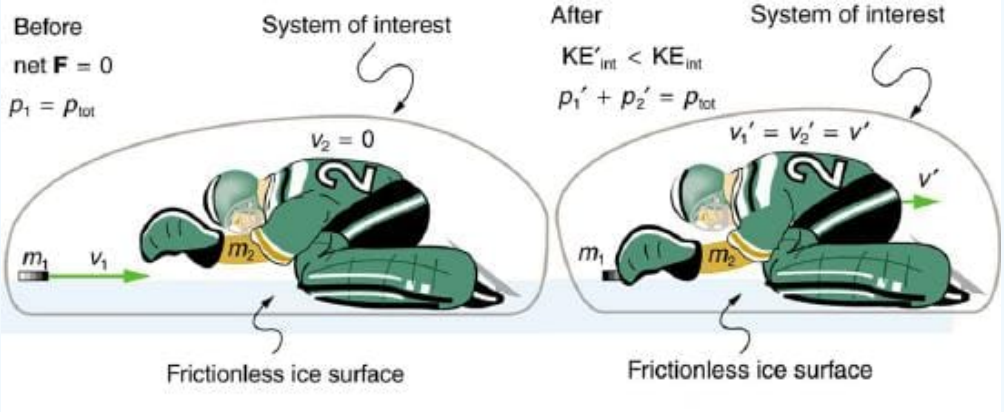


Figure 2.2.6.2: An ice hockey goalie catches a hockey puck and recoils backward. The initial kinetic energy of the puck is almost entirely converted to thermal energy and sound in this inelastic collision.

Strategy

Momentum is conserved because the net external force on the puck-goalie system is zero. We can thus use conservation of momentum to find the final velocity of the puck and goalie system. Note that the initial velocity of the goalie is zero and that the final velocity of the puck and goalie are the same. Once the final velocity is found, the kinetic energies can be calculated before and after the collision and compared as requested.

Solution for (a)

Momentum is conserved because the net external force on the puck-goalie system is zero.

Conservation of momentum is

$$p_1 + p_2 = p_1' + p_2'$$

or

$$m_1 v_1 + m_2 v_2 = m_1 v_1' + m_2 v_2'.$$

Because the goalie is initially at rest, we know $v_2 = 0$. Because the goalie catches the puck, the final velocities are equal, or $v_1' = v_2' = v'$. Thus, the conservation of momentum equation simplifies to

$$m_1 v_1 = (m_1 + m_2)\, v'.$$

Solving for v' yields

$$v' = \frac{m_1}{m_1 + m_2} v_1.$$

Entering known values in this equation, we get

$$v' = \left(\frac{0.150 \text{ kg}}{0.150 \text{ kg} + 70.0 \text{ kg}}\right)(35.0 \text{ m/s}) = 7.48 \times 10^{-2} \text{ m/s}.$$

Discussion for (a)

This recoil velocity is small and in the same direction as the puck's original velocity, as we might expect.

Solution for (b)

Before the collision, the total kinetic energy $\mathrm{KE}_{\mathrm{int}}$ of the system is that of the hockey puck, because the goalie is initially at rest. Therefore, $\mathrm{KE}_{\mathrm{int}}$ is initially

$$\begin{aligned}\mathrm{KE}_{\mathrm{int}} &= \frac{1}{2}mv^2 = \frac{1}{2}(0.150\ \mathrm{kg})(35.0\ \mathrm{m/s})^2 \\ &= 91.9\ \mathrm{J}\end{aligned}.$$

After the collision, the total kinetic energy is

$$\begin{aligned}\mathrm{KE}'_{\mathrm{int}} &= \frac{1}{2}(m+M)v^2 = \frac{1}{2}(70.15\ \mathrm{kg})\left(7.48\times10^{-2}\ \mathrm{m/s}\right)^2 \\ &= 0.196\ \mathrm{J}\end{aligned}.$$

The change in total kinetic energy is thus

$$\begin{aligned}\mathrm{KE}'_{\mathrm{int}} - \mathrm{KE}_{\mathrm{int}} &= 0.196\ \mathrm{J} - 91.9\ \mathrm{J} \\ &= -91.7\ \mathrm{J}\end{aligned}$$

where the minus sign indicates that the energy was lost.

Discussion for (b)

Nearly all of the initial total kinetic energy is lost in this perfectly inelastic collision. $\mathrm{KE}_{\mathrm{int}}$ is mostly converted to thermal energy and sound. Note that in this case, the total kinetic energy does not decrease all the way down to zero, because doing so would violate conservation of momentum, because initial total momentum is not zero. This is the most amount of energy that can be converted from the initial total kinetic energy without violating conservation of momentum.

Example 2.2.6.2: Calculating Final Velocity and Energy Release: Two Carts Collide

During some collisions stored energy may be converted into kinetic energy during a collision. Figure 2.2.6.3 shows a one-dimensional example in which two carts on an air track collide, releasing potential energy from a compressed spring.

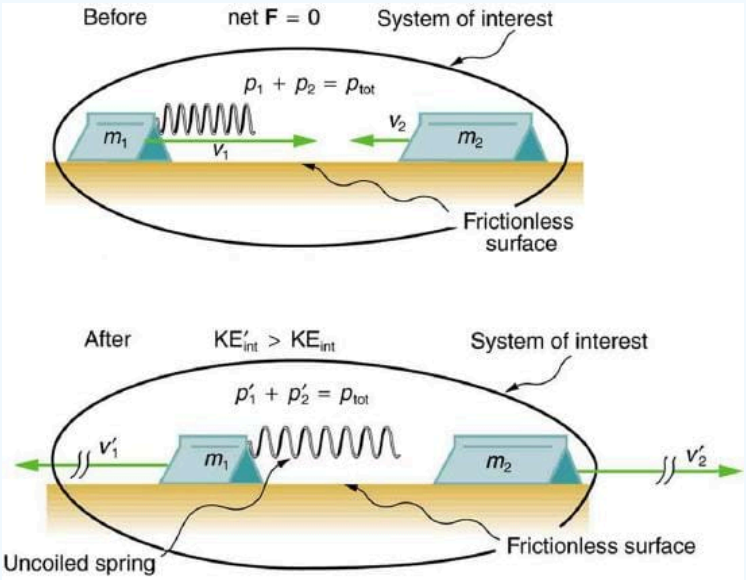


Figure 2.2.6.3: An air track is nearly frictionless, so that momentum is conserved. Motion is one-dimensional. In this collision, examined in Example 2.2.6.2, the potential energy of a compressed spring is released during the collision and is converted to internal kinetic energy.

In the collision pictured in Figure 2.2.6.3, two carts collide inelastically. Cart 1 (denoted m_1 carries a spring which is initially compressed. During the collision, the spring releases its potential energy and converts it to total kinetic energy. The mass of cart 1 and the spring is 0.350 kg, and the cart and the spring together have an initial velocity of $2.00\ \mathrm{m/s}$. Cart 2 (denoted m_2 in Figure 2.2.6.3) has a mass of 0.500 kg and an initial velocity of $-0.500\ \mathrm{m/s}$. After the collision, cart 1 is observed to recoil with a velocity of $-4.00\ \mathrm{m/s}$. (a) What is the final velocity of cart 2? (b) How much energy was released by the spring (assuming all of it was converted into internal kinetic energy)?

Strategy

We can use conservation of momentum to find the final velocity of cart 2, because $F_{\mathrm{net}} = 0$ (the track is frictionless and the force of the spring is internal). Once this velocity is determined, we can compare the total kinetic energy before and after the collision to see how much energy was released by the spring.

Solution for (a)

As before, the equation for conservation of momentum in a two-object system is

$$m_1 v_1 + m_2 v_2 = m_1 v_1' + m_2 v_2'.$$

The only unknown in this equation is v_2'. Solving for v_2' and substituting known values into the previous equation yields

$$\begin{aligned} v_2' &= \frac{m_1 v_1 + m_2 v_2 - m_1 v_1'}{m_2} \\ &= \frac{(0.350\ \mathrm{kg})(2.00\ \mathrm{m/s}) + (0.500\ \mathrm{kg})(-0.500\ \mathrm{m/s})}{0.500\ \mathrm{kg}} - \frac{(0.350\ \mathrm{kg})(-4.00\ \mathrm{m/s})}{0.500\ \mathrm{kg}} \\ &= 3.70\ \mathrm{m/s}. \end{aligned}$$

Solution for (b)

The total kinetic energy before the collision is

$$\begin{aligned} \mathrm{KE}_{\mathrm{int}} &= \frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 \\ &= \frac{1}{2}(0.350\ \mathrm{kg})(2.00\ \mathrm{m/s})^2 + \frac{1}{2}(0.500\ \mathrm{kg})(-0.500\ \mathrm{m/s})^2. \\ &= 0.763\ \mathrm{J} \end{aligned}$$

After the collision, the total kinetic energy is

$$\begin{aligned} \mathrm{KE}_{\mathrm{int}}' &= \frac{1}{2} m_1 v_1'^2 + \frac{1}{2} m_2 v_2'^2 \\ &= \frac{1}{2}(0.350\ \mathrm{kg})(-4.00\ \mathrm{m/s})^2 + \frac{1}{2}(0.500\ \mathrm{kg})(3.70\ \mathrm{m/s})^2 \\ &= 6.22\ \mathrm{J}. \end{aligned}$$

The change in total kinetic energy is thus

$$\begin{aligned} \mathrm{KE}_{\mathrm{int}}' - \mathrm{KE}_{\mathrm{int}} &= 6.22\ \mathrm{J} - 0.763\ \mathrm{J} \\ &= 5.46\ \mathrm{J}. \end{aligned}$$

Discussion

The final velocity of cart 2 is large and positive, meaning that it is moving to the right after the collision. The total kinetic energy in this collision increases by 5.46 J. That energy was released by the spring.

Section Summary

- An inelastic collision is one in which the total kinetic energy changes (it is not conserved).
- A collision in which the objects stick together is sometimes called perfectly inelastic because it reduces total kinetic energy as much as possible while conserving total momentum.

Glossary

inelastic collision
: a collision in which total kinetic energy is not conserved

perfectly inelastic collision
: a collision in which the colliding objects stick together

This page titled 2.2.6: Inelastic Collisions in One Dimension is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 5.6: Inelastic Collisions in One Dimension by Andrew Park is licensed CC BY 4.0.

2.2.7: Impulse and Momentum (Exercise)

Conceptual Questions

4.1: Linear Momentum and Force

1. An object that has a small mass and an object that has a large mass have the same momentum. Which object has the largest kinetic energy?

2. An object that has a small mass and an object that has a large mass have the same kinetic energy. Which mass has the largest momentum?

Professional Application

3. Football coaches advise players to block, hit, and tackle with their feet on the ground rather than by leaping through the air. Using the concepts of momentum, work, and energy, explain how a football player can be more effective with his feet on the ground.

4. How can a small force impart the same momentum to an object as a large force?

4.2: Impulse

Professional Application

5. Explain in terms of impulse how padding reduces forces in a collision. State this in terms of a real example, such as the advantages of a carpeted vs. tile floor for a day care center.

6. While jumping on a trampoline, sometimes you land on your back and other times on your feet. In which case can you reach a greater height and why?

Professional Application

7. Tennis racquets have “sweet spots.” If the ball hits a sweet spot then the player's arm is not jarred as much as it would be otherwise. Explain why this is the case.

4.3: Conservation of Momentum

Professional Application

8. If you dive into water, you reach greater depths than if you do a belly flop. Explain this difference in depth using the concept of conservation of energy. Explain this difference in depth using what you have learned in this chapter.

9. Under what circumstances is momentum conserved?

10. Can momentum be conserved for a system if there are external forces acting on the system? If so, under what conditions? If not, why not?

11. Momentum for a system can be conserved in one direction while not being conserved in another. What is the angle between the directions? Give an example.

Professional Application

12. Explain in terms of momentum and Newton’s laws how a car’s air resistance is due in part to the fact that it pushes air in its direction of motion.

13. Can objects in a system have momentum while the momentum of the system is zero? Explain your answer.

14. Must the total energy of a system be conserved whenever its momentum is conserved? Explain why or why not.

4.4: Elastic Collisions in One Dimension

15. What physical quantities are conserved in an elastic collision?

16. Give examples of elastic (or nearly elastic) collisions from everyday experiences.

4.5: Inelastic Collisions in One Dimension

17. What is an inelastic collision? What is a perfectly inelastic collision?

18. Mixed-pair ice skaters performing in a show are standing motionless at arms length just before starting a routine. They reach out, clasp hands, and pull themselves together by only using their arms. Assuming there is no friction between the blades of their skates and the ice, what is their velocity after their bodies meet?

19. A small pickup truck that has a camper shell slowly coasts toward a red light with negligible friction. Two dogs in the back of the truck are moving and making various inelastic collisions with each other and the walls. What is the effect of the dogs on the motion of the center of mass of the system (truck plus entire load)? What is their effect on the motion of the truck?

This page titled 2.2.7: Impulse and Momentum (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 5.7: Impulse and Momentum (Exercise) by Andrew Park is licensed CC BY 4.0.

2.3: OSCILLATIONS AND WAVES

2.3.1: Introduction to Oscillatory Motion and Waves

2.3.2: Period and Frequency in Oscillations

2.3.3: Simple Harmonic Motion- A Special Periodic Motion

2.3.4: Forced Oscillations and Resonance

2.3.5: Waves

2.3.6: Wave Interference- Standing Waves and Beats

2.3.7: Sound

2.3.8: Speed of Sound, Frequency, and Wavelength

2.3.9: Doppler Effect and Sonic Booms

2.3.10: Oscillations and Waves (Exercise)

This page titled 2.3: Oscillations and Waves is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

2.3.1: Introduction to Oscillatory Motion and Waves



Figure 2.3.1.1: There are at least four types of waves in this picture—only the water waves are evident. There are also sound waves, light waves, and waves on the guitar strings. (credit: John Norton)

What do an ocean buoy, a child in a swing, the cone inside a speaker, a guitar, atoms in a crystal, the motion of chest cavities, and the beating of hearts all have in common? They all **oscillate**—-that is, they move back and forth between two points. Many systems oscillate, and they have certain characteristics in common. All oscillations involve force and energy. You push a child in a swing to get the motion started. The energy of atoms vibrating in a crystal can be increased with heat. You put energy into a guitar string when you pluck it.

Some oscillations create **waves**. A guitar creates sound waves. You can make water waves in a swimming pool by slapping the water with your hand. You can no doubt think of other types of waves. Some, such as water waves, are visible. Some, such as sound waves, are not. But *every wave is a disturbance that moves from its source and carries energy*. Other examples of waves include earthquakes and visible light. Even subatomic particles, such as electrons, can behave like waves.

By studying oscillatory motion and waves, we shall find that a small number of underlying principles describe all of them and that wave phenomena are more common than you have ever imagined.

Glossary

oscillate
: moving back and forth regularly between two points

wave
: a disturbance that moves from its source and carries energy

This page titled 2.3.1: Introduction to Oscillatory Motion and Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.1: Introduction to Oscillatory Motion and Waves by Andrew Park is licensed CC BY 4.0.

2.3.2: Period and Frequency in Oscillations

Learning Objectives

- Understand the relationship between the frequency and the period of oscillations.
- Determine the frequency of oscillations.



Figure 2.3.2.1: The strings on this guitar vibrate at regular time intervals. (credit: JAR)

When you pluck a guitar string, the resulting sound has a steady tone and lasts a long time. Each successive vibration of the string takes the same time as the previous one. We define **periodic motion** to be a motion that repeats itself at regular time intervals, such as exhibited by the guitar string or by an object on a spring moving up and down. The time to complete one oscillation remains constant and is called the **period** T. Its units are usually seconds, but may be any convenient unit of time. The word period refers to the time for some event whether repetitive or not; but we shall be primarily interested in periodic motion, which is by definition repetitive. A concept closely related to period is the frequency of an event. For example, if you get a paycheck twice a month, the frequency of payment is two per month and the period between checks is half a month. **Frequency** f is defined to be the number of events per unit time. For periodic motion, frequency is the number of oscillations per unit time. The relationship between frequency and period is

$$f = \frac{1}{T}.$$

The SI unit for frequency is the *cycle per second*, which is defined to be a *hertz* (Hz):

$$1 \text{ Hz} = 1\frac{\text{cycle}}{\text{sec}} \text{ or } 1 \text{ Hz} = \frac{1}{\text{s}}$$

A cycle is one complete oscillation. Note that a vibration can be a single or multiple event, whereas oscillations are usually repetitive for a significant number of cycles.

Example 2.3.2.1: Determine the Frequency of Two Oscillations: Medical Ultrasound and the Period of Middle C

We can use the formulas presented in this module to determine both the frequency based on known oscillations and the oscillation based on a known frequency. Let's try one example of each. (a) A medical imaging device produces ultrasound by oscillating with a period of 0.400 µs. What is the frequency of this oscillation? (b) The frequency of middle C on a typical musical instrument is 264 Hz. What is the time for one complete oscillation?

Strategy

Both questions (a) and (b) can be answered using the relationship between period and frequency. In question (a), the period T is given and we are asked to find frequency f. In question (b), the frequency f is given and we are asked to find the period T.

Solution a

1. Substitute $0.400\ \mu\text{s}$ for T in $f = \frac{1}{T}$:

$$f = \frac{1}{T} = \frac{1}{0.400 \times 10^{-6}\ \text{s}}.$$

Solve to find

$$f = 2.50 \times 10^{6}\ \text{Hz}.$$

Discussion a

The frequency of sound found in (a) is much higher than the highest frequency that humans can hear and, therefore, is called ultrasound. Appropriate oscillations at this frequency generate ultrasound used for noninvasive medical diagnoses, such as observations of a fetus in the womb.

Solution b

1. Identify the known values:

The time for one complete oscillation is the period T:

$$f = \frac{1}{T}.$$

2. Solve for T:

$$T = \frac{1}{f}.$$

3. Substitute the given value for the frequency into the resulting expression:

$$T = \frac{1}{f} = \frac{1}{264\ \text{Hz}} = \frac{1}{264\text{cycles/s}} = 3.79 \times 10^{-3}\ \text{s} = 3.79\ \text{ms}.$$

Discussion

The period found in (b) is the time per cycle, but this value is often quoted as simply the time in convenient units (ms or milliseconds in this case).

Exercise 2.3.2.1

Identify an event in your life (such as receiving a paycheck) that occurs regularly. Identify both the period and frequency of this event.

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- Periodic motion is a repetitious oscillation.
- The time for one oscillation is the period T.

- The number of oscillations per unit time is the frequency f.
- These quantities are related by

$$f = \frac{1}{T}.$$

Glossary

period
: time it takes to complete one oscillation

periodic motion
: motion that repeats itself at regular time intervals

frequency
: number of events per unit of time

Answers

Exercise 2.3.2.1

Exercise 2.3.2.1

I visit my parents for dinner every other Sunday. The frequency of my visits is 26 per calendar year. The period is two weeks.

This page titled 2.3.2: Period and Frequency in Oscillations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.2: Period and Frequency in Oscillations by Andrew Park is licensed CC BY 4.0.

2.3.3: Simple Harmonic Motion- A Special Periodic Motion

Learning Objectives

- Describe a simple harmonic oscillator.
- Explain the link between simple harmonic motion and waves.

The oscillations of a system in which the net force can be described by Hooke's law are of special importance, because they are very common. They are also the simplest oscillatory systems. **Simple Harmonic Motion** (SHM) is the name given to oscillatory motion for a system where the net force can be described by Hooke's law, and such a system is called a **simple harmonic oscillator**. If the net force can be described by Hooke's law and there is no *damping* (by friction or other non-conservative forces), then a simple harmonic oscillator will oscillate with equal displacement on either side of the equilibrium position, as shown for an object on a spring in Figure 2.3.3.1. The maximum displacement from equilibrium is called the **amplitude** X. The units for amplitude and displacement are the same, but depend on the type of oscillation. For the object on the spring, the units of amplitude and displacement are meters; whereas for sound oscillations, they have units of pressure (and other types of oscillations have yet other units). Because amplitude is the maximum displacement, it is related to the energy in the oscillation.

TAKE-HOME EXPERIMENT: SHM AND THE MARBLE

Find a bowl or basin that is shaped like a hemisphere on the inside. Place a marble inside the bowl and tilt the bowl periodically so the marble rolls from the bottom of the bowl to equally high points on the sides of the bowl. Get a feel for the force required to maintain this periodic motion. What is the restoring force and what role does the force you apply play in the simple harmonic motion (SHM) of the marble?

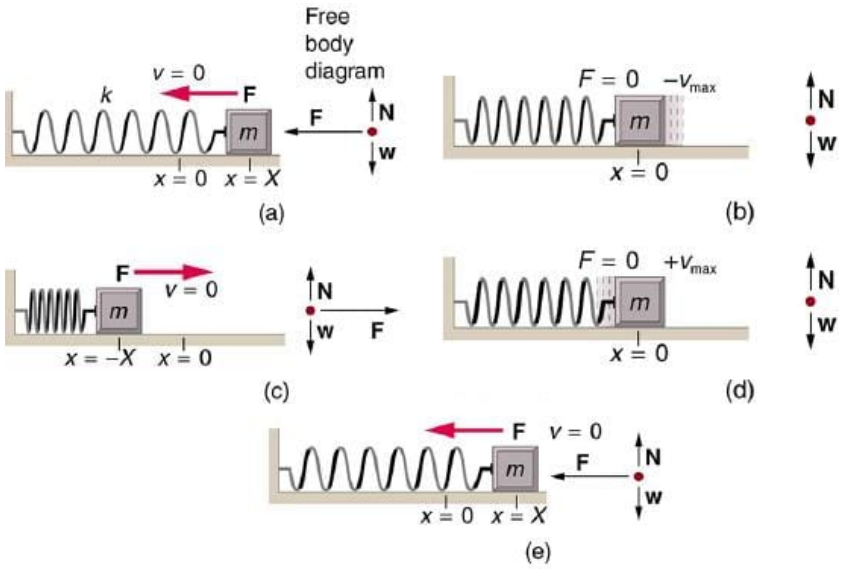


Figure 2.3.3.1: An object attached to a spring sliding on a frictionless surface is an uncomplicated simple harmonic oscillator. When displaced from equilibrium, the object performs simple harmonic motion that has an amplitude X and a period T. The object's maximum speed occurs as it passes through equilibrium. The stiffer the spring is, the smaller the period T. The greater the mass of the object is, the greater the period T.

What is so significant about simple harmonic motion? One special thing is that the period T and frequency f of a simple harmonic oscillator are independent of amplitude. The string of a guitar, for example, will oscillate with the same frequency whether plucked gently or hard. Because the period is constant, a simple harmonic oscillator can be used as a clock.

Two important factors do affect the period of a simple harmonic oscillator. The period is related to how stiff the system is. A very stiff object has a large spring constant k, which causes the system to have a smaller period. For example, you can adjust a

diving board's stiffness—the stiffer it is, the faster it vibrates, and the shorter its period. Period also depends on the mass of the oscillating system. The more massive the system is, the longer the period. For example, a heavy person on a diving board bounces up and down more slowly than a light one.

In fact, the mass m and the spring constant k are the *only* factors that affect the period and frequency of simple harmonic motion of a mass on a spring.

PERIOD OF SIMPLE HARMONIC OSCILLATOR

The *period of a simple harmonic oscillator* is given by

$$T = 2\pi\sqrt{\frac{m}{k}}$$

and, because $f = 1/T$, the *frequency of a simple harmonic oscillator* is

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}.$$

Note that neither T nor f has any dependence on amplitude.

TAKE-HOME EXPERIMENT: MASS AND RULER OSCILLATIONS

Find two identical wooden or plastic rulers. Tape one end of each ruler firmly to the edge of a table so that the length of each ruler that protrudes from the table is the same. On the free end of one ruler tape a heavy object such as a few large coins. Pluck the ends of the rulers at the same time and observe which one undergoes more cycles in a time period, and measure the period of oscillation of each of the rulers.

Example 2.3.3.1: Calculate the Frequency and Period of Oscillations: Bad Shock Absorbers in a Car

If the shock absorbers in a car go bad, then the car will oscillate at the least provocation, such as when going over bumps in the road and after stopping (See Figure 2.3.3.2). Calculate the frequency and period of these oscillations for such a car if the car's mass (including its load) is 900 kg and the spring constant (k) of the suspension system is $6.53 \times 10^4\ \mathrm{N/m}$.

Strategy

The frequency of the car's oscillations will be that of a simple harmonic oscillator as given in the equation $f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$. The mass and the spring constant are both given.

Solution

1. Enter the known values of *k* and *m*:

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = \frac{1}{2\pi}\sqrt{\frac{6.53 \times 10^4\ \mathrm{N/m}}{900\ \mathrm{kg}}}.$$

2. Calculate the frequency:

$$\frac{1}{2\pi}\sqrt{72.6/\mathrm{s}^{-2}} = 1.3656/\mathrm{s}^{-1} \approx 1.36/\mathrm{s}^{-1} = 1.36\ \mathrm{Hz}.$$

3. You could use $T = 2\pi\sqrt{\frac{m}{k}}$ to calculate the period, but it is simpler to use the relationship $T = 1/f$ and substitute the value just found for f:

$$T = \frac{1}{f} = \frac{1}{1.356\ \mathrm{Hz}} = 0.738\ \mathrm{s}.$$

Discussion

The values of T and f both seem about right for a bouncing car. You can observe these oscillations if you push down hard on the end of a car and let go.

The Link between Simple Harmonic Motion and Waves

If a time-exposure photograph of the bouncing car were taken as it drove by, the headlight would make a wavelike streak, as shown in Figure 2.3.3.2. Similarly, Figure 2.3.3.3 shows an object bouncing on a spring as it leaves a wavelike "trace of its position on a moving strip of paper. Both waves are sine functions. All simple harmonic motion is intimately related to sine and cosine waves.

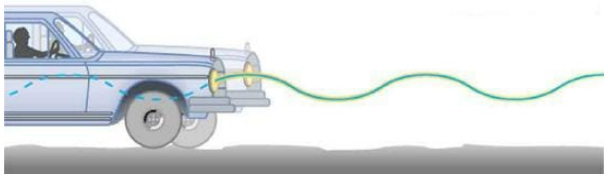


Figure 2.3.3.2: The bouncing car makes a wavelike motion. If the restoring force in the suspension system can be described only by Hooke's law, then the wave is a sine function. (The wave is the trace produced by the headlight as the car moves to the right.)

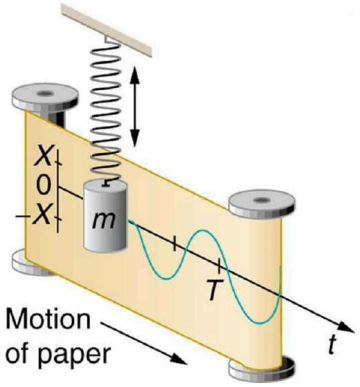


Figure 2.3.3.3: The vertical position of an object bouncing on a spring is recorded on a strip of moving paper, leaving a sine wave.

The displacement $x(t)$, the velocity $v(t)$, and the acceleration $a(t)$ of simple harmonic motion sketches out sinusoidal functions as a function of time. That is, if you sketch out $\sin(t)$ and $\cos(t)$, with t in seconds, for example, what you graph will appear similar to the actual graphs of $x(t)$,$v(t)$, and $a(t)$. You just need to scale them horizontally and vertically, so that they have correct period and amplitude, respectively.

Figure 2.3.3.4 shows the simple harmonic motion of an object on a spring and presents graphs of $x(t)$,$v(t)$, and $a(t)$ versus time.

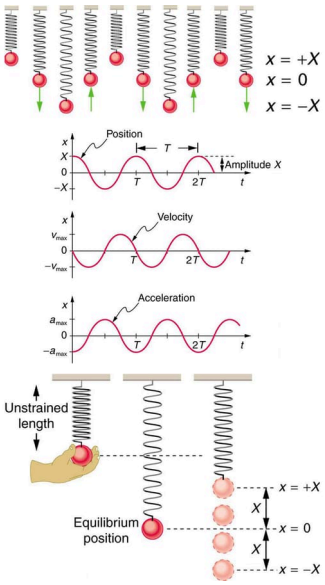


Figure 2.3.3.4: Graphs of $x(t)$,$v(t)$, and $a(t)$ versus t for the motion of an object on a spring. The net force on the object can be described by Hooke's law, and so the object undergoes simple harmonic motion. Note that the initial position has the vertical displacement at its maximum value X; v is initially zero and then negative as the object moves down; and the initial acceleration is negative, back toward the equilibrium position and becomes zero at that point.

The most important point here is that these relationships are valid for all simple harmonic motion. They are very useful in visualizing waves associated with simple harmonic motion, including visualizing how waves add with one another.

Exercise 2.3.3.1

Suppose you pluck a banjo string. You hear a single note that starts out loud and slowly quiets over time. Describe what happens to the sound waves in terms of period, frequency and amplitude as the sound decreases in volume.

Answer available at the end of this section.

Exercise 2.3.3.2

A babysitter is pushing a child on a swing. At the point where the swing reaches x, where would the corresponding point on a wave of this motion be located?

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- Simple harmonic motion is oscillatory motion for a system that can be described only by Hooke's law. Such a system is also called a simple harmonic oscillator.
- Maximum displacement is the amplitude X. The period T and frequency f of a simple harmonic oscillator are given by

  $T = 2\pi\sqrt{\frac{m}{k}}$ and $f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$, where m is the mass of the system.

- Displacement, velocity, and acceleration in simple harmonic motion as a function of time are represented by sinusoidal functions in time, that is functions that are similar to $\sin(t)$ and $\cos(t)$.

Glossary

amplitude
: the maximum displacement from the equilibrium position of an object oscillating around the equilibrium position

simple harmonic motion
: the oscillatory motion in a system where the net force can be described by Hooke's law

simple harmonic oscillator
: a device that implements Hooke's law, such as a mass that is attached to a spring, with the other end of the spring being connected to a rigid support such as a wall

Answers

Exercise 2.3.3.1

Frequency and period remain essentially unchanged. Only amplitude decreases as volume decreases.

Exercise 2.3.3.2

Exercise 2.3.3.2

x is the maximum deformation, which corresponds to the amplitude of the wave. The point on the wave would either be at the very top or the very bottom of the curve.

This page titled 2.3.3: Simple Harmonic Motion- A Special Periodic Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.3: Simple Harmonic Motion- A Special Periodic Motion by Andrew Park is licensed CC BY 4.0.

2.3.4: Forced Oscillations and Resonance

Learning Objectives

- Observe the resonance phenomena in several examples.
- Observe amplitude of a damped harmonic oscillator.
- Understand the origin of damping of resonance.



Figure 2.3.4.1: You can cause the strings in a piano to vibrate simply by producing sound waves from your voice. (credit: Matt Billings, Flickr)

Sit in front of a piano sometime and sing a loud brief note at it with the dampers off its strings. It will sing the same note back at you—the strings, having the same frequencies as your voice, are resonating in response to the forces from the sound waves that you sent to them. Your voice and a piano's strings is a good example of the fact that objects—in this case, piano strings—can be forced to oscillate but oscillate best at their natural frequency. In this section, we shall briefly explore applying a *periodic driving force* acting on a simple harmonic oscillator. The driving force puts energy into the system at a certain frequency, not necessarily the same as the natural frequency of the system. The **natural frequency** is the frequency at which a simple harmonic oscillator oscillates, if there is no damping force, and other than an initial kick to get the motion started, there is no driving force. A **damping force** is a frictional force (friction, air resistance, and other forces that lead to decrease in mechanical energy) which converts the mechanical energy of the oscillatory motion into thermal energy.

Most of us have played with toys involving an object supported on an elastic band, something like the paddle ball suspended from a finger in Figure 2.3.4.2. Imagine the finger in the figure is your finger. At first you hold your finger steady, and the ball bounces up and down with a small amount of damping. If you move your finger up and down slowly, the ball will follow along without bouncing much on its own. As you increase the frequency at which you move your finger up and down, the ball will respond by oscillating with increasing amplitude. When you drive the ball at its natural frequency, the ball's oscillations increase in amplitude with each oscillation for as long as you drive it. The phenomenon of driving a system with a frequency equal to its natural frequency is called **resonance**. A system being driven at its natural frequency is said to **resonate**. As the driving frequency gets progressively higher than the resonant or natural frequency, the amplitude of the oscillations becomes smaller, until the oscillations nearly disappear and your finger simply moves up and down with little effect on the ball.

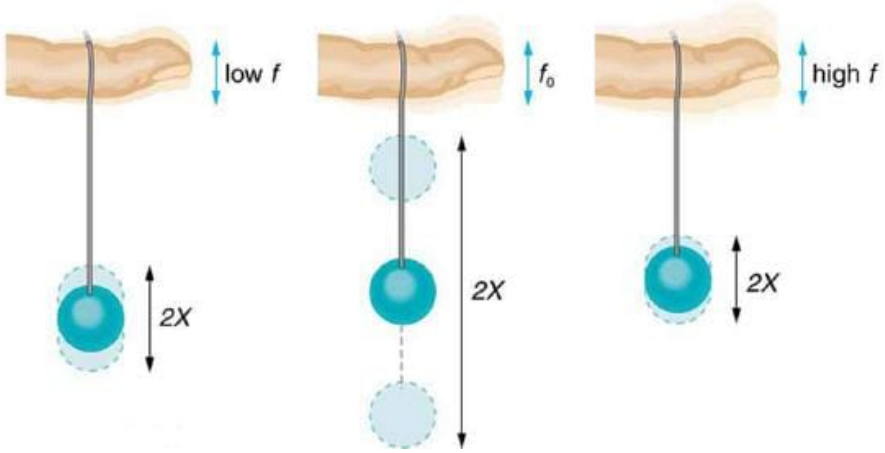


Figure 2.3.4.2: The paddle ball on its rubber band moves in response to the finger supporting it. If the finger moves with the natural frequency f_0 of the ball on the rubber band, then a resonance is achieved, and the amplitude of the ball's oscillations increases dramatically. At higher and lower driving frequencies, energy is transferred to the ball less efficiently, and it responds with lower-amplitude oscillations.

Figure 2.3.4.3 shows a graph of the amplitude of a damped harmonic oscillator as a function of the frequency of the periodic force driving it. There are three curves on the graph, each representing a different amount of damping. All three curves peak at the point where the frequency of the driving force equals the natural frequency of the harmonic oscillator. The highest peak, or greatest response, is for the least amount of damping, because less energy is removed by the damping force.

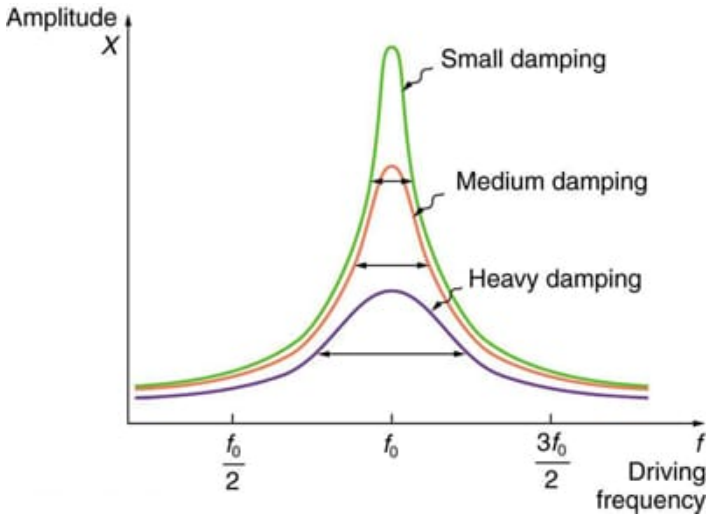


Figure 2.3.4.3: Amplitude of a harmonic oscillator as a function of the frequency of the driving force. The curves represent the same oscillator with the same natural frequency but with different amounts of damping. Resonance occurs when the driving frequency equals the natural frequency, and the greatest response is for the least amount of damping. The narrowest response is also for the least damping.

It is interesting that the widths of the resonance curves shown in Figure 2.3.4.3 depend on damping: the less the damping, the narrower the resonance. The message is that if you want a driven oscillator to resonate at a very specific frequency, you need as little damping as possible. Little damping is the case for piano strings and many other musical instruments. Conversely, if you want small-amplitude oscillations, such as in a car's suspension system, then you want heavy damping. Heavy damping reduces the amplitude, but the tradeoff is that the system responds at more frequencies.

These features of driven harmonic oscillators apply to a huge variety of systems. When you tune a radio, for example, you are adjusting its resonant frequency so that it only oscillates to the desired station's broadcast (driving) frequency. The more selective the radio is in discriminating between stations, the smaller its damping. Magnetic resonance imaging (MRI) is a widely used medical diagnostic tool in which atomic nuclei (mostly hydrogen nuclei) are made to resonate by incoming radio waves (on the order of 100 MHz). A child on a swing is driven by a parent at the swing's natural frequency to achieve maximum amplitude. In all of these cases, the efficiency of energy transfer from the driving force into the oscillator is best at resonance. Speed bumps and gravel roads prove that even a car's suspension system is not immune to resonance. In spite of finely engineered shock absorbers, which ordinarily convert mechanical energy to thermal energy almost as fast as it comes in, speed bumps still cause a

large-amplitude oscillation. On gravel roads that are corrugated, you may have noticed that if you travel at the “wrong” speed, the bumps are very noticeable whereas at other speeds you may hardly feel the bumps at all. Figure 2.3.4.4 shows a photograph of a famous example (the Tacoma Narrows Bridge) of the destructive effects of a driven harmonic oscillation. The Millennium Bridge in London was closed for a short period of time for the same reason while inspections were carried out.

In our bodies, the chest cavity is a clear example of a system at resonance. The diaphragm and chest wall drive the oscillations of the chest cavity which result in the lungs inflating and deflating. The system is critically damped and the muscular diaphragm oscillates at the resonant value for the system, making it highly efficient.



Figure 2.3.4.4: In 1940, the Tacoma Narrows Bridge in Washington state collapsed. Heavy cross winds drove the bridge into oscillations at its resonant frequency. Damping decreased when support cables broke loose and started to slip over the towers, allowing increasingly greater amplitudes until the structure failed (credit: PRI's *Studio 360*, via Flickr)

Exercise 2.3.4.1

A famous magic trick involves a performer singing a note toward a crystal glass until the glass shatters. Explain why the trick works in terms of resonance and natural frequency.

Answer available at the end of this section.

Section Summary

- A system’s natural frequency is the frequency at which the system will oscillate if not affected by driving or damping forces.
- A periodic force driving a harmonic oscillator at its natural frequency produces resonance. The system is said to resonate.
- The less damping a system has, the higher the amplitude of the forced oscillations near resonance. The more damping a system has, the broader response it has to varying driving frequencies.

Glossary

natural frequency
: the frequency at which a simple harmonic oscillator oscillates if it is set in motion without a driving force

damping force
: a frictional force which converts the mechanical energy of the oscillatory motion into thermal energy

resonance
: the phenomenon of driving a system with a frequency equal to the system's natural frequency

resonate
: a system being driven at its natural frequency

Answers

Exercise 2.3.4.1

The performer must be singing a note that corresponds to the natural frequency of the glass. As the sound wave is directed at the glass, the glass responds by resonating at the same frequency as the sound wave. With enough energy introduced into the system, the glass begins to vibrate and eventually shatters.

This page titled 2.3.4: Forced Oscillations and Resonance is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.4: Forced Oscillations and Resonance by Andrew Park is licensed CC BY 4.0.

2.3.5: Waves

Learning Objectives

- State the characteristics of a wave.
- Calculate the velocity of wave propagation, given frequency and wavelength



Figure 2.3.5.1: Waves in the ocean behave similarly to all other types of waves. (credit: Steve Jurveston, Flickr)

What do we mean when we say something is a wave? The most intuitive and easiest wave to imagine is the familiar water wave. More precisely, a **wave** is a disturbance that propagates, or moves from the place it was created. For water waves, the disturbance is in the surface of the water, perhaps created by a rock thrown into a pond or by a swimmer splashing the surface repeatedly. For sound waves, the disturbance is a change in air pressure, perhaps created by the oscillating cone inside a speaker. For earthquakes, there are several types of disturbances, including disturbance of Earth's surface and pressure disturbances under the surface. Even radio waves are most easily understood using an analogy with water waves. Visualizing water waves is useful because there is more to it than just a mental image. Water waves exhibit characteristics common to all waves, such as amplitude, period, frequency and energy. All wave characteristics can be described by a small set of underlying principles.

A wave is a disturbance that propagates, or moves from the place it was created. The simplest waves repeat themselves for several cycles and are associated with simple harmonic motion. Let us start by considering the simplified water wave in Figure 2.3.5.2. The wave is an up and down disturbance of the water surface. It causes a sea gull to move up and down in simple harmonic motion as the wave crests and troughs (peaks and valleys) pass under the bird. The time for one complete up and down motion is the wave's period T. The wave's frequency is $f = 1/T$, as usual. The wave itself moves to the right in the figure. This movement of the wave is actually the disturbance moving to the right, not the water itself (or the bird would move to the right). We define **wave velocity** v_{w} to be the speed at which the disturbance moves. Wave velocity is sometimes also called the *propagation velocity or propagation speed,* because the disturbance propagates from one location to another.

MISCONCEPTION ALERT

Many people think that water waves push water from one direction to another. In fact, the particles of water tend to stay in one location, save for moving up and down due to the energy in the wave. The energy moves forward through the water, but the water stays in one place. If you feel yourself pushed in an ocean, what you feel is the energy of the wave, not a rush of water.

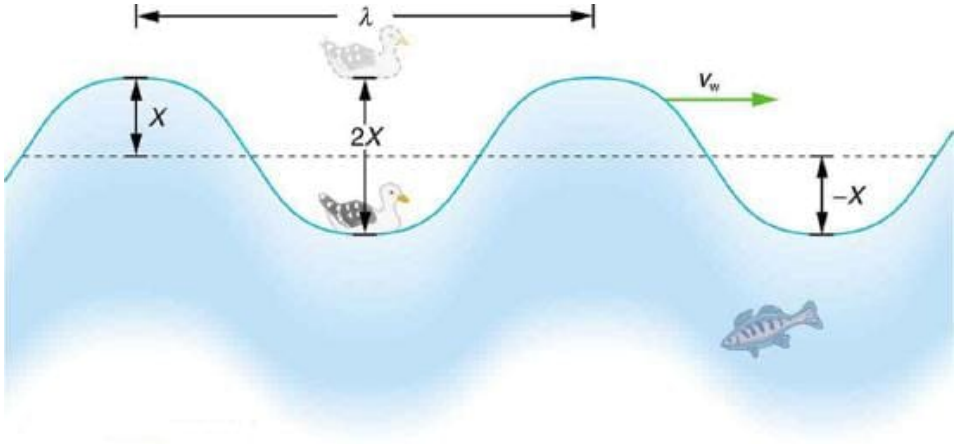


Figure 2.3.5.2: An idealized ocean wave passes under a sea gull that bobs up and down in simple harmonic motion. The wave has a wavelength λ, which is the distance between adjacent identical parts of the wave. The up and down disturbance of the surface propagates parallel to the surface at a speed v_w.

he water wave in the figure also has a length associated with it, called its **wavelength** λ, the distance between adjacent identical parts of a wave. (λ is the distance parallel to the direction of propagation.) The speed of propagation v_w is the distance the wave travels in a given time, which is one wavelength in the time of one period. In equation form, that is

$$v_w = \frac{\lambda}{T}$$

or

$$v_w = f\lambda. \tag{2.3.5.1}$$

This fundamental relationship holds for all types of waves. For water waves, v_w is the speed of a surface wave; for sound, v_w is the speed of sound; and for visible light, v_w is the speed of light, for example.

One note of caution with Equation (2.3.5.1): despite the appearance, wave speed v_w is a *not* a function of frequency f and wavelength λ. When f changes (or λ changes), v_w does not change (you will see this explicitly later with sound and light waves). Wave speed is a *property of the medium* the wave travels in. Unless the medium itself changes, the wave speed remains constant. This is true of sound and light waves, as well as almost all other types of waves you will see in this textbook (water waves, waves on a string, etc.). So, what happens when frequency or wavelength does change, then? The wavelength and frequency change to compensate. For example, if the frequency doubles for a given wave, instead of the wave speed doubling, the wavelength will decrease to half, so that the product $f\lambda$ remains constant.

TAKE-HOME EXPERIMENT: WAVES IN A BOWL

Fill a large bowl or basin with water and wait for the water to settle so there are no ripples. Gently drop a cork into the middle of the bowl. Estimate the wavelength and period of oscillation of the water wave that propagates away from the cork. Remove the cork from the bowl and wait for the water to settle again. Gently drop the cork at a height that is different from the first drop. Does the wavelength depend upon how high above the water the cork is dropped?

Calculate the Velocity of Wave Propagation: Gull in the Ocean

Calculate the wave velocity of the ocean wave in Figure 2.3.5.2 if the distance between wave crests is 10.0 m and the time for a sea gull to bob up and down is 5.00 s.

Strategy

We are asked to find v_w. The given information tells us that $\lambda = 10.0$ m and $T = 5.00$ s. Therefore, we can use $v_w = \frac{\lambda}{T}$ to find the wave velocity.

Solution

1. Enter the known values into $v_{\mathrm{w}} = \frac{\lambda}{T}$:

$$v_{\mathrm{w}} = \frac{10.0\ \mathrm{m}}{5.00\ \mathrm{s}}.$$

2. Solve for v_{w} to find $v_{\mathrm{w}} = 2.00\ \mathrm{m/s}$.

Discussion

This slow speed seems reasonable for an ocean wave. Note that the wave moves to the right in the figure at this speed, not the varying speed at which the sea gull moves up and down.

Transverse and Longitudinal Waves

A simple wave consists of a periodic disturbance that propagates from one place to another. The wave in Figure 2.3.5.3 propagates in the horizontal direction while the surface is disturbed in the vertical direction. Such a wave is called a **transverse wave** or shear wave; in such a wave, the disturbance is perpendicular to the direction of propagation. In contrast, in a **longitudinal wave** or compressional wave, the disturbance is parallel to the direction of propagation. Figure 2.3.5.4 shows an example of a longitudinal wave. The size of the disturbance is its amplitude X and is completely independent of the speed of propagation v_{w}.

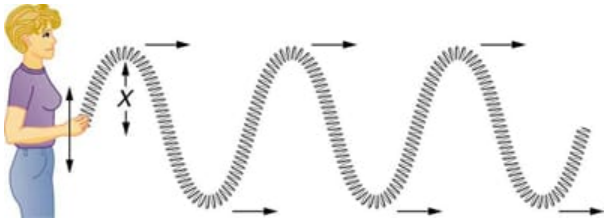


Figure 2.3.5.3: In this example of a transverse wave, the wave propagates horizontally, and the disturbance in the cord is in the vertical direction.

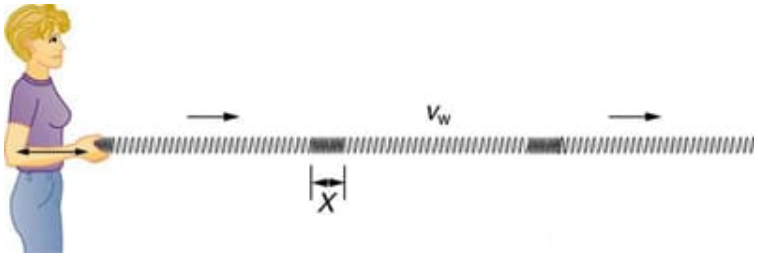


Figure 2.3.5.4: In this example of a longitudinal wave, the wave propagates horizontally, and the disturbance in the cord is also in the horizontal direction.

Waves may be transverse, longitudinal, or *a combination of the two*. (Water waves are actually a combination of transverse and longitudinal. The simplified water wave illustrated in Figure 2.3.5.2 shows no longitudinal motion of the bird.) The waves on the strings of musical instruments are transverse—so are electromagnetic waves, such as visible light.

Sound waves in air and water are longitudinal. Their disturbances are periodic variations in pressure that are transmitted in fluids. Fluids do not have appreciable shear strength, and thus the sound waves in them must be longitudinal or compressional. Sound in solids can be both longitudinal and transverse.

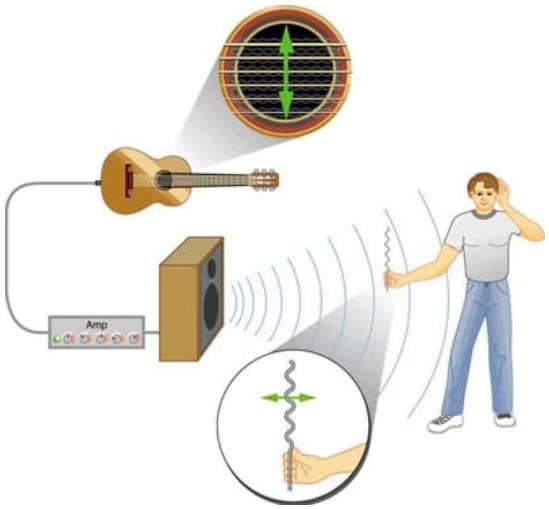


Figure 2.3.5.5: The wave on a guitar string is transverse. The sound wave rattles a sheet of paper in a direction that shows the sound wave is longitudinal.

Earthquake waves under Earth's surface also have both longitudinal and transverse components (called compressional or P-waves and shear or S-waves, respectively). These components have important individual characteristics—they propagate at different speeds, for example. Earthquakes also have surface waves that are similar to surface waves on water.

Exercise 2.3.5.1

Why is it important to differentiate between longitudinal and transverse waves?

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- A wave is a disturbance that moves from the point of creation with a wave velocity v_{w}.
- A wave has a wavelength λ, which is the distance between adjacent identical parts of the wave.
- Wave velocity and wavelength are related to the wave's frequency and period by $v_{\text{w}} = \frac{\lambda}{T}$ or $v_{\text{w}} = f\lambda$. However, wave velocity is a property of the medium and remains constant as wavelength or frequency changes.
- A transverse wave has a disturbance perpendicular to its direction of propagation, whereas a longitudinal wave has a disturbance parallel to its direction of propagation.

Glossary

longitudinal wave
: a wave in which the disturbance is parallel to the direction of propagation

transverse wave
: a wave in which the disturbance is perpendicular to the direction of propagation

wave velocity
: the speed at which the disturbance moves; also called wave speed, propagation velocity, or propagation speed

wavelength
: the distance between adjacent identical parts of a wave

Answers

Exercise 2.3.5.1

Exercise 2.3.5.1

In the different types of waves, energy can propagate in a different direction relative to the motion of the wave. This is important to understand how different types of waves affect the materials around them.

This page titled 2.3.5: Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.5: Waves by Andrew Park is licensed CC BY 4.0.

2.3.6: Wave Interference- Standing Waves and Beats

Learning Objectives

- Distinguish between constructive interference and destructive interference.
- Explain standing waves in terms of interference, nodes, and antinodes.
- Explain the beat phenomenon in terms of interference.



Figure 2.3.6.1: These waves result from the superposition of several waves from different sources, producing a complex pattern. (credit: waterborough, Wikimedia Commons)

Most waves do not look very simple. They look more like the waves in Figure 2.3.6.1 than like the simple water wave considered in "Waves". (Simple waves may be created by a simple harmonic oscillation, and thus have a sinusoidal shape). Complex waves are more interesting, even beautiful, but they look formidable. Most waves appear complex because they result from several simple waves adding together. Luckily, the rules for adding waves are quite simple.

When two or more waves arrive at the same point, they superimpose themselves on one another. More specifically, the disturbances of waves are superimposed when they come together—a phenomenon called **superposition**. Each disturbance corresponds to a force, and forces add. If the disturbances are along the same line, then the resulting wave is a simple addition of the disturbances of the individual waves—that is, their amplitudes add. Figure 2.3.6.2 and Figure 2.3.6.3 illustrate superposition in two special cases, both of which produce simple results.

Figure 2.3.6.2 shows two identical waves that arrive at the same point exactly in phase. The crests of the two waves are precisely aligned, as are the troughs. This superposition produces pure **constructive interference**. Because the disturbances add, pure constructive interference produces a wave that has twice the amplitude of the individual waves, but has the same wavelength.

Figure 2.3.6.3 shows two identical waves that arrive exactly out of phase—that is, precisely aligned crest to trough—producing pure **destructive interference**. Because the disturbances are in the opposite direction for this superposition, the resulting amplitude is zero for pure destructive interference—the waves completely cancel.

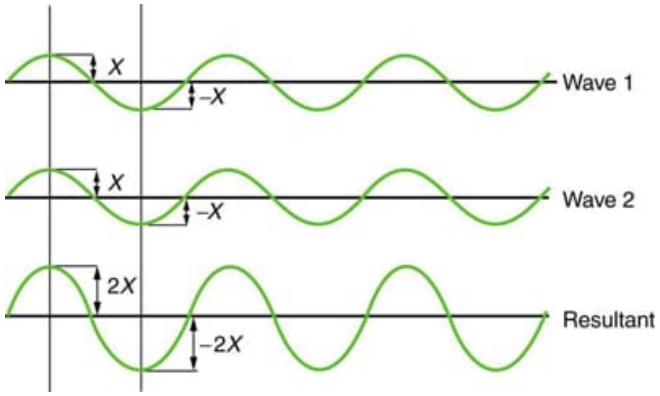


Figure 2.3.6.2: Pure constructive interference of two identical waves produces one with twice the amplitude, but the same wavelength.

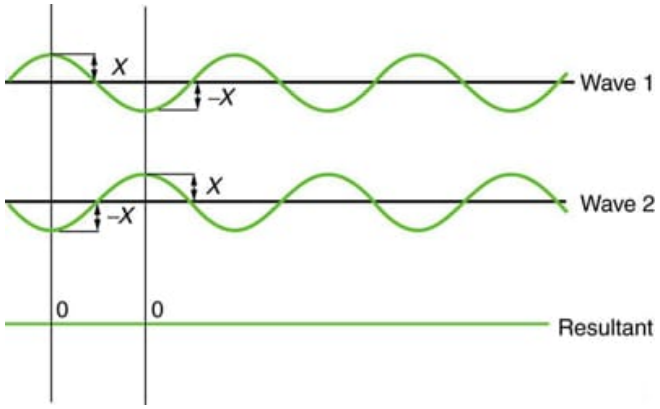


Figure 2.3.6.3: Pure destructive interference of two identical waves produces zero amplitude, or complete cancellation.

While pure constructive and pure destructive interference do occur, they require precisely aligned identical waves. The superposition of most waves produces a combination of constructive and destructive interference and can vary from place to place and time to time. Sound from a stereo, for example, can be loud in one spot and quiet in another. Varying loudness means the sound waves add partially constructively and partially destructively at different locations. A stereo has at least two speakers creating sound waves, and waves can reflect from walls. All these waves superimpose. An example of sounds that vary over time from constructive to destructive is found in the combined whine of airplane jets heard by a stationary passenger. The combined sound can fluctuate up and down in volume as the sound from the two engines varies in time from constructive to destructive. These examples are of waves that are similar.

An example of the superposition of two dissimilar waves is shown in Figure 2.3.6.4. Here again, the disturbances add and subtract, producing a more complicated looking wave.

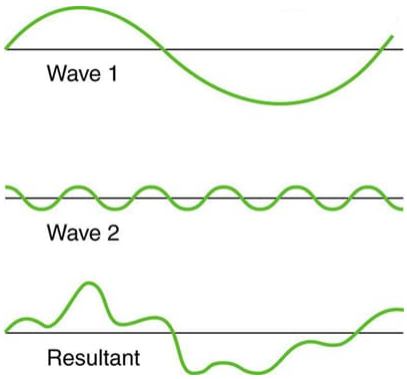


Figure 2.3.6.4: Superposition of non-identical waves exhibits both constructive and destructive interference.

Standing Waves

Sometimes waves do not seem to move; rather, they just vibrate in place. Unmoving waves can be seen on the surface of a glass of milk in a refrigerator, for example. Vibrations from the refrigerator motor create waves on the milk that oscillate up and down but do not seem to move across the surface. These waves are formed by the superposition of two or more moving waves, such as illustrated in Figure 2.3.6.5 for two identical waves moving in opposite directions. The waves move through each other with their disturbances adding as they go by. If the two waves have the same amplitude and wavelength, then they alternate between constructive and destructive interference. The resultant looks like a wave standing in place and, thus, is called a **standing wave**. Waves on the glass of milk are one example of standing waves. There are other standing waves, such as on guitar strings and in organ pipes. With the glass of milk, the two waves that produce standing waves may come from reflections from the side of the glass.

A closer look at earthquakes provides evidence for conditions appropriate for resonance, standing waves, and constructive and destructive interference. A building may be vibrated for several seconds with a driving frequency matching that of the natural frequency of vibration of the building—producing a resonance resulting in one building collapsing while neighboring buildings do not. Often buildings of a certain height are devastated while other taller buildings remain intact. The building height matches the condition for setting up a standing wave for that particular height. As the earthquake waves travel along the surface of Earth and reflect off denser rocks, constructive interference occurs at certain points. Often areas closer to the epicenter are not damaged while areas farther away are damaged.

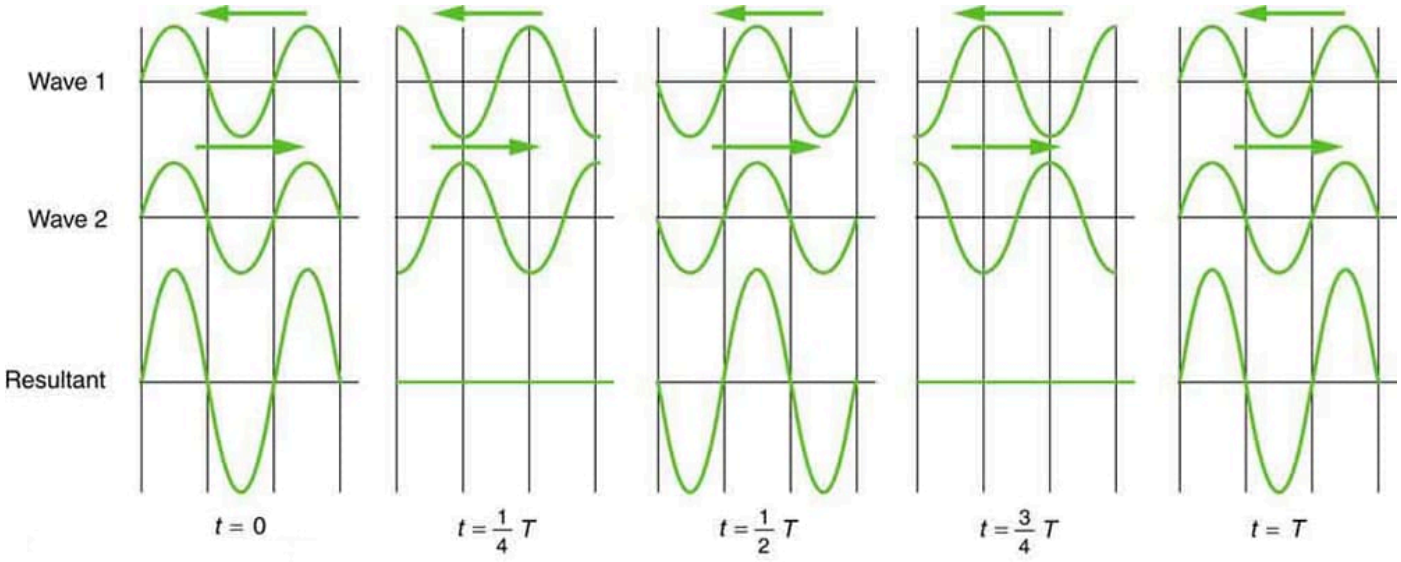


Figure 2.3.6.5: Standing wave created by the superposition of two identical waves moving in opposite directions. The oscillations are at fixed locations in space and result from alternately constructive and destructive interference.

Standing waves are also found on the strings of musical instruments and are due to reflections of waves from the ends of the string. Figure 2.3.6.6 and Figure 2.3.6.7 show three standing waves that can be created on a string that is fixed at both ends. **Nodes** are the points where the string does not move; more generally, nodes are where the wave disturbance is zero in a standing wave. The fixed ends of strings must be nodes, too, because the string cannot move there. The word **antinode** is used to denote the location of maximum amplitude in standing waves. Standing waves on strings have a frequency that is related to the propagation speed v_{w} of the disturbance on the string. The wavelength λ is determined by the distance between the points where the string is fixed in place.

The lowest frequency, called the **fundamental frequency**, is thus for the longest wavelength, which is seen to be $\lambda_1 = 2L$. Therefore, the fundamental frequency is $f_1 = v_{\mathrm{w}}/\lambda_1 = v_{\mathrm{w}}/2L$. In this case, the **overtones** or harmonics are multiples of the fundamental frequency. As seen in Figure 2.3.6.7, the first harmonic can easily be calculated since $\lambda_2 = L$. Thus, $f_2 = v_{\mathrm{w}}/\lambda_2 = v_{\mathrm{w}}/L = 2f_1$. Similarly, $f_3 = 3f_1$, and so on. All of these frequencies can be changed by adjusting the tension in the string. The greater the tension, the greater v_{w} is and the higher the frequencies. This observation is familiar to anyone who has ever observed a string instrument being tuned. We will see in later chapters that standing waves are crucial to many resonance phenomena, such as in sounding boxes on string instruments.

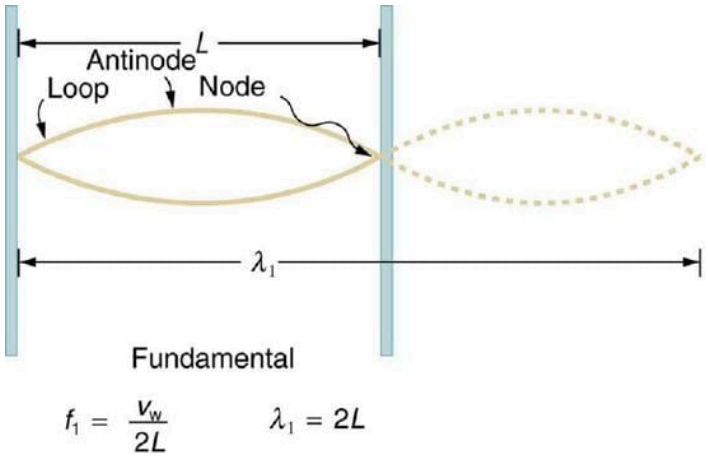


Figure 2.3.6.6: The figure shows a string oscillating at its fundamental frequency.

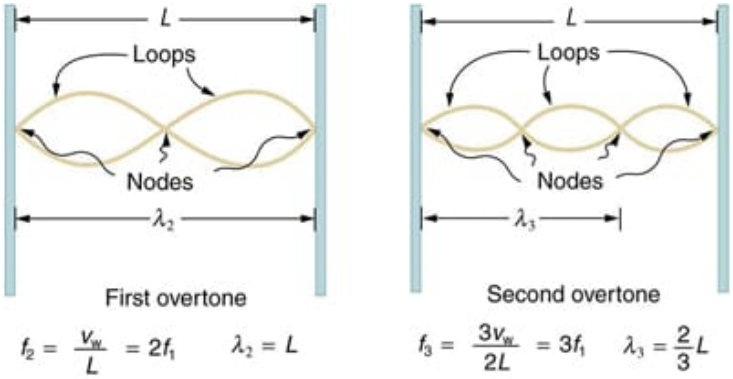


Figure 2.3.6.7: First and second harmonic frequencies are shown.

Beats

Striking two adjacent keys on a piano produces a warbling combination usually considered to be unpleasant. The superposition of two waves of similar but not identical frequencies is the culprit. Another example is often noticeable in jet aircraft, particularly the two-engine variety, while taxiing. The combined sound of the engines goes up and down in loudness. This varying loudness happens because the sound waves have similar but not identical frequencies. The discordant warbling of the piano and the fluctuating loudness of the jet engine noise are both due to alternately constructive and destructive interference as the two waves go in and out of phase. Figure 2.3.6.8 illustrates this graphically.

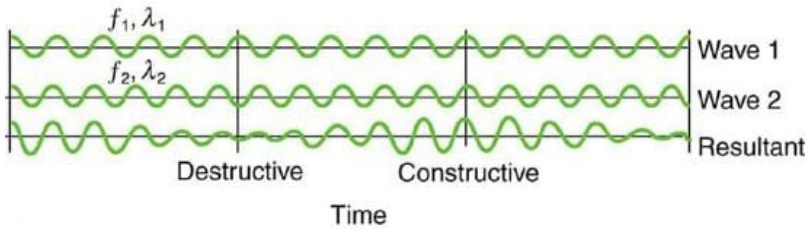


Figure 2.3.6.8: Beats are produced by the superposition of two waves of slightly different frequencies but identical amplitudes. The waves alternate in time between constructive interference and destructive interference, giving the resulting wave a time-varying amplitude.

The wave resulting from the superposition of two similar-frequency waves has a frequency that is the average of the two. This wave fluctuates in amplitude, or *beats*, with a frequency called the **beat frequency**. We can determine the beat frequency by adding two waves together mathematically. The result,

$$f_{\mathrm{B}} = |f_1 - f_2|$$

is the beat frequency. The resultant wave has the average frequency of the two superimposed waves, but it also fluctuates in

overall amplitude at the beat frequency f_{B}. This result is valid for all types of waves. However, if it is a sound wave, providing the two frequencies are similar, then what we hear is an average frequency that gets louder and softer (or warbles) at the beat frequency.

MAKING CAREER CONNECTIONS

Piano tuners use beats routinely in their work. When comparing a note with a tuning fork, they listen for beats and adjust the string until the beats go away (to zero frequency). For example, if the tuning fork has a $256\ \mathrm{Hz}$ frequency and two beats per second are heard, then the other frequency is either 254 or $258\ \mathrm{Hz}$. Most keys hit multiple strings, and these strings are actually adjusted until they have nearly the same frequency and give a slow beat for richness. Twelve-string guitars and mandolins are also tuned using beats.

While beats may sometimes be annoying in audible sounds, we will find that beats have many applications. Observing beats is a very useful way to compare similar frequencies. There are applications of beats as apparently disparate as in ultrasonic imaging and radar speed traps.

Exercise 2.3.6.1

Imagine you are holding one end of a jump rope, and your friend holds the other. If your friend holds her end still, you can move your end up and down, creating a transverse wave. If your friend then begins to move her end up and down, generating a wave in the opposite direction, what resultant wave forms would you expect to see in the jump rope?

Answer available at the end of this section.

Exercise 2.3.6.2

Define nodes and antinodes.

Answer available at the end of this section.

Answer available at the end of this section.

Exercise 2.3.6.3

You hook up a stereo system. When you test the system, you notice that in one corner of the room, the sounds seem dull. In another area, the sounds seem excessively loud. Describe how the sound moving about the room could result in these effects.

Answer available at the end of this section.

Section Summary

- Superposition is the combination of two waves at the same location.
- Constructive interference occurs when two waves are superimposed in phase (both disturbances in the same direction).
- Destructive interference occurs when two waves are superimposed exactly out of phase (two disturbances in opposite directions).
- A standing wave is one in which two waves superimpose to produce a wave that forms a standing pattern of nodes and antinodes, locations of no displacement and of maximum displacement.
- Standing waves form at specific resonance frequencies, with the lowest frequency being the fundamental frequency, and higher frequencies being called overtones or harmonics.
- Beats occur when waves of similar frequencies f_1 and f_2 are superimposed. The resulting amplitude oscillates with a beat frequency given by

$$f_{\mathrm{B}} = |f_1 - f_2| \, .$$

Glossary

antinode
: the location of maximum amplitude in standing waves

beat frequency
: the frequency of the amplitude fluctuations of a wave

constructive interference
: when two waves arrive at the same point exactly in phase; that is, the crests of the two waves are precisely aligned, as are the troughs

destructive interference
: when two identical waves arrive at the same point exactly out of phase; that is, precisely aligned crest to trough

fundamental frequency
: the lowest frequency of a periodic waveform

nodes
: the points where the string does not move; more generally, nodes are where the wave disturbance is zero in a standing wave

overtones
: multiples of the fundamental frequency of a sound

superposition
: the phenomenon that occurs when two or more waves arrive at the same point

Answers

Exercise 2.3.6.1

The rope would alternate between having waves with amplitudes two times the original amplitude and reaching equilibrium with no amplitude at all. The wavelengths will result in both constructive and destructive interference.

Exercise 2.3.6.2

Exercise 2.3.6.2

Nodes are areas of wave interference where there is no motion. Antinodes are areas of wave interference where the motion is at its maximum point.

Exercise 2.3.6.3

With multiple speakers putting out sounds into the room, and these sounds bouncing off walls, there is bound to be some wave interference. In the dull areas, the interference is probably mostly destructive. In the louder areas, the interference is probably mostly constructive.

This page titled 2.3.6: Wave Interference- Standing Waves and Beats is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.6: Wave Interference- Standing Waves and Beats by Andrew Park is licensed CC BY 4.0.

2.3.7: Sound

Learning Objectives

- Define sound and hearing.
- Describe sound as a longitudinal wave.
- Distinguish between audible sound, ultrasound, and infrasound.



Figure 2.3.7.1: This glass has been shattered by a high-intensity sound wave of the same frequency as the resonant frequency of the glass. While the sound is not visible, the effects of the sound prove its existence. (credit: ||read||, Flickr)

Sound can be used as a familiar illustration of waves. Because hearing is one of our most important senses, it is interesting to see how the physical properties of sound correspond to our perceptions of it. **Hearing** is the perception of sound, just as vision is the perception of visible light. But sound has important applications beyond hearing. Ultrasound, for example, is not heard but can be employed to form medical images and is also used in treatment.

The physical phenomenon of **sound** is defined to be a disturbance of matter that is transmitted from its source outward. Sound is a wave. On the atomic scale, it is a disturbance of atoms that is far more ordered than their thermal motions. In many instances, sound is a periodic wave, and the atoms undergo simple harmonic motion. In this text, we shall explore such periodic sound waves.

A vibrating string produces a sound wave as illustrated in Figure 2.3.7.2, Figure 2.3.7.3, and Figure 2.3.7.4. As the string oscillates back and forth, it transfers energy to the air, mostly as thermal energy created by turbulence. But a small part of the string's energy goes into compressing and expanding the surrounding air, creating slightly higher and lower local pressures. These compressions (high pressure regions) and rarefactions (low pressure regions) move out as longitudinal pressure waves having the same frequency as the string—they are the disturbance that is a sound wave. (Sound waves in air and most fluids are longitudinal, because fluids have almost no shear strength. In solids, sound waves can be both transverse and longitudinal.) Figure 2.3.7.4 shows a graph of gauge pressure versus distance from the vibrating string.

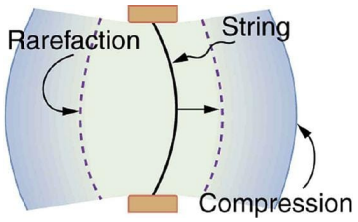


Figure 2.3.7.2: A vibrating string moving to the right compresses the air in front of it and expands the air behind it.

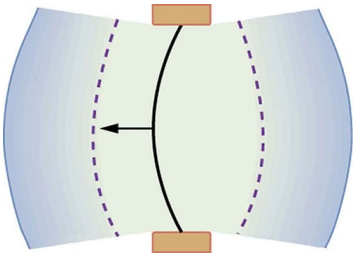


Figure 2.3.7.3: As the string moves to the left, it creates another compression and rarefaction as the ones on the right move away from the string.

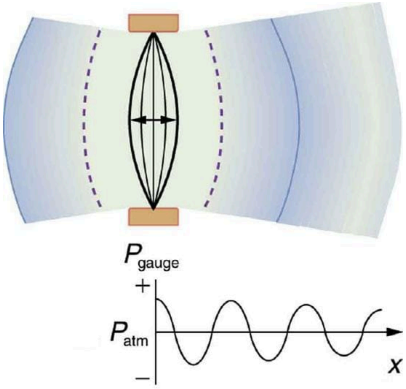


Figure 2.3.7.4: After many vibrations, there are a series of compressions and rarefactions moving out from the string as a sound wave. The graph shows gauge pressure versus distance from the source. Pressures vary only slightly from atmospheric for ordinary sounds.

The amplitude of a sound wave decreases with distance from its source, because the energy of the wave is spread over a larger and larger area. But it is also absorbed by objects, such as the eardrum in Figure 2.3.7.5, and converted to thermal energy by the viscosity of air. In addition, during each compression a little heat transfers to the air and during each rarefaction even less heat transfers from the air, so that the heat transfer reduces the organized disturbance into random thermal motions. Whether the heat transfer from compression to rarefaction is significant depends on how far apart they are—that is, it depends on wavelength. Wavelength, frequency, amplitude, and speed of propagation are important for sound, as they are for all waves.

When the frequency of the sound wave is in **audible range** of 20 to 20,000 Hz, the vibration of the eardrums at these frequencies can be sensed, which is what we hear. Sound waves of frequency above 20 kHz are called **ultrasound** waves. Some construction equipment produces **infrasound** waves of frequencies less than 20 Hz, which, although they can't be heard, can still affect the human body at large enough amplitudes.

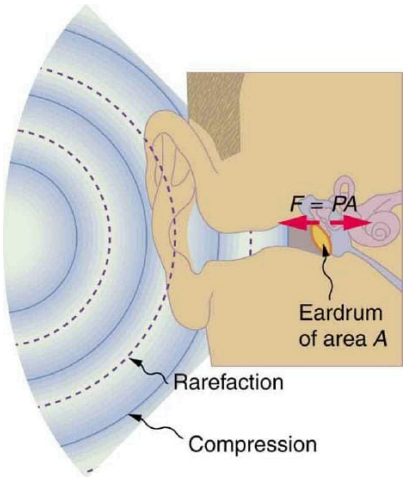


Figure 2.3.7.5: Sound wave compressions and rarefactions travel up the ear canal and force the eardrum to vibrate. There is a net force on the eardrum, since the sound wave pressures differ from the atmospheric pressure found behind the eardrum. A complicated mechanism converts the vibrations to nerve impulses, which are perceived by the person.

Section Summary

- Sound is a disturbance of matter that is transmitted from its source outward.
- Sound is one type of wave.
- Hearing is the perception of sound.

Glossary

audible range
: frequencies between 20 Hz and 20,000 Hz

infrasound
: sound waves below audible frequencies (less than 20 Hz)

hearing
: the perception of sound

sound
: a disturbance of matter that is transmitted from its source outward

ultrasound
: sound waves above audible frequencies (greater than 20,000 Hz)

This page titled 2.3.7: Sound is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.7: Sound by Andrew Park is licensed CC BY 4.0.

2.3.8: Speed of Sound, Frequency, and Wavelength

Learning Objectives

- Define pitch.
- Describe the relationship between the speed of sound, its frequency, and its wavelength.
- Describe the effects on the speed of sound as it travels through various media.
- Describe the effects of temperature on the speed of sound.



Figure 2.3.8.1: When a firework explodes, the light energy is perceived before the sound energy. Sound travels more slowly than light does. (credit: Dominic Alves, Flickr)

Sound, like all waves, travels at a certain speed and has the properties of frequency and wavelength. You can observe direct evidence of the speed of sound while watching a fireworks display. The flash of an explosion is seen well before its sound is heard, implying both that sound travels at a finite speed and that it is much slower than light. You can also directly sense the frequency of a sound. Perception of frequency is called **pitch**. The wavelength of sound is not directly sensed, but indirect evidence is found in the correlation of the size of musical instruments with their pitch. Small instruments, such as a piccolo, typically make high-pitch sounds, while large instruments, such as a tuba, typically make low-pitch sounds. High pitch means small wavelength, and the size of a musical instrument is directly related to the wavelengths of sound it produces. So a small instrument creates short-wavelength sounds. Similar arguments hold that a large instrument creates long-wavelength sounds.

The relationship of the speed of sound, its frequency, and wavelength is the same as for all waves:

$$v_{\mathrm{w}} = f\lambda,$$

where v_{w} is the speed of sound, f is its frequency, and λ is its wavelength. The wavelength of a sound is the distance between adjacent identical parts of a wave—for example, between adjacent compressions as illustrated in Figure 2.3.8.2. The frequency is the same as that of the source and is the number of waves that pass a point per unit time.

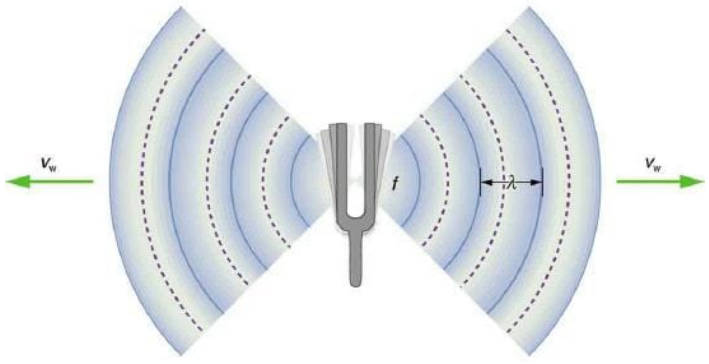


Figure 2.3.8.2: A sound wave emanates from a source vibrating at a frequency f, propagates at v_{w}, and has a wavelength λ.

Table 2.3.8.1 makes it apparent that the speed of sound varies greatly in different media. The speed of sound in a medium

is determined by a combination of the medium's rigidity (or compressibility in gases) and its density. The more rigid (or less compressible) the medium, the faster the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is directly proportional to the stiffness of the oscillating object. The greater the density of a medium, the slower the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is inversely proportional to the mass of the oscillating object. The speed of sound in air is low, because air is compressible. Because liquids and solids are relatively rigid and very difficult to compress, the speed of sound in such media is generally greater than in gases.

Table 2.3.8.1: Speed of Sound in Various Media

Medium	v_{w}(m/s)
Gases at 0°C	
Air	331
Carbon dioxide	259
Oxygen	316
Helium	965
Hydrogen	1290
Liquids at 20°C	
Ethanol	1160
Mercury	1450
Water, fresh	1480
Sea water	1540
Human tissue	1540
Solids (longitudinal or bulk)	
Vulcanized rubber	54
Polyethylene	920
Marble	3810
Glass, Pyrex	5640
Lead	1960
Aluminum	5120
Steel	5960

Earthquakes, essentially sound waves in Earth's crust, are an interesting example of how the speed of sound depends on the rigidity of the medium. Earthquakes have both longitudinal and transverse components, and these travel at different speeds. The bulk modulus of granite is greater than its shear modulus. For that reason, the speed of longitudinal or pressure waves (P-waves) in earthquakes in granite is significantly higher than the speed of transverse or shear waves (S-waves). Both components of earthquakes travel slower in less rigid material, such as sediments. P-waves have speeds of 4 to 7 km/s, and S-waves correspondingly range in speed from 2 to 5 km/s, both being faster in more rigid material. The P-wave gets progressively farther ahead of the S-wave as they travel through Earth's crust. The time between the P- and S-waves is routinely used to determine the distance to their source, the epicenter of the earthquake.

The speed of sound is affected by temperature in a given medium. For air at sea level, the speed of sound is given by

$$v_{\mathrm{w}} = (331\ \mathrm{m/s})\sqrt{\frac{T}{273\ \mathrm{K}}},$$

where the temperature (denoted as T) is in units of kelvin. While not negligible, this is not a strong dependence. At 0°C, the speed of sound is 331 m/s, whereas at 20.0°C it is 343 m/s, less than a 4% increase. Figure 2.3.8.3 shows a use of the speed of sound by a bat to sense distances. Echoes are also used in medical imaging.

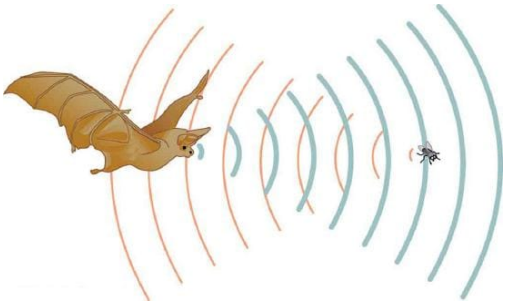


Figure 2.3.8.3: A bat uses sound echoes to find its way about and to catch prey. The time for the echo to return is directly proportional to the distance.

One of the more important properties of sound is that its speed is nearly independent of frequency. This independence is certainly true in open air for sounds in the audible range of 20 to 20,000 Hz. If this independence were not true, you would certainly notice it for music played by a marching band in a football stadium, for example. Suppose that high-frequency sounds traveled faster—then the farther you were from the band, the more the sound from the low-pitch instruments would lag that from the high-pitch ones. But the music from all instruments arrives in cadence independent of distance, and so all frequencies must travel at nearly the same speed. Recall that

$$v_{\text{w}} = f\lambda.$$

In a given medium under fixed conditions, v_{w} is constant, so that there is a relationship between f and λ; the higher the frequency, the smaller the wavelength. See Figure 2.3.8.4 and consider the following example.

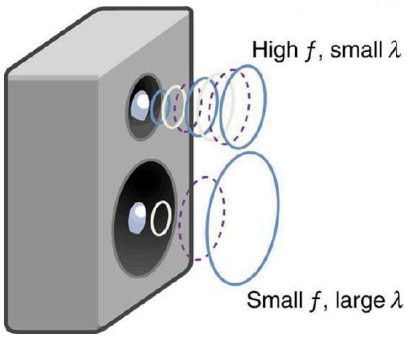


Figure 2.3.8.4: Because they travel at the same speed in a given medium, low-frequency sounds must have a greater wavelength than high-frequency sounds. Here, the lower-frequency sounds are emitted by the large speaker, called a woofer, while the higher-frequency sounds are emitted by the small speaker, called a tweeter.

Example 2.3.8.1: Calculating Wavelengths: What Are the Wavelengths of Audible Sounds?

Calculate the wavelengths of sounds at the extremes of the audible range, 20 and 20,000 Hz, in 30.0°C air. (Assume that the frequency values are accurate to two significant figures.)

Strategy

To find wavelength from frequency, we can use $v_{\text{w}} = f\lambda$.

Solution

1. Identify knowns. The value for v_{w}, is given by

$$v_{\text{w}} = (331\ \text{m/s})\sqrt{\frac{T}{273\ \text{K}}}.$$

2. Convert the temperature into kelvin and then enter the temperature into the equation

$$v_{\text{w}} = (331 \text{ m/s})\sqrt{\frac{303 \text{ K}}{273 \text{ K}}} = 348.7 \text{ m/s}.$$

3. Solve the relationship between speed and wavelength for λ:

$$\lambda = \frac{v_{\text{w}}}{f}.$$

4. Enter the speed and the minimum frequency to give the maximum wavelength:

$$\lambda_{\text{max}} = \frac{348.7 \text{ m/s}}{20 \text{ Hz}} = 17 \text{ m}.$$

5. Enter the speed and the maximum frequency to give the minimum wavelength:

$$\lambda_{\text{min}} = \frac{348.7 \text{ m/s}}{20,000 \text{ Hz}} = 0.017 \text{ m} = 1.7 \text{ cm}.$$

Discussion

Because the product of f multiplied by λ equals a constant, the smaller f is, the larger λ must be, and vice versa.

The speed of sound can change when sound travels from one medium to another. However, the frequency usually remains the same because it is like a driven oscillation and has the frequency of the original source. If v_{w} changes and f remains the same, then the wavelength λ must change. That is, because $v_{\text{w}} = f\lambda$, the higher the speed of a sound, the greater its wavelength for a given frequency.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—VOICE AS A SOUND WAVE

Suspend a sheet of paper so that the top edge of the paper is fixed and the bottom edge is free to move. You could tape the top edge of the paper to the edge of a table. Gently blow near the edge of the bottom of the sheet and note how the sheet moves. Speak softly and then louder such that the sounds hit the edge of the bottom of the paper, and note how the sheet moves. Explain the effects.

Exercise 2.3.8.1

Imagine you observe two fireworks explode. You hear the explosion of one as soon as you see it. However, you see the other firework for several milliseconds before you hear the explosion. Explain why this is so.

Answer available at the end of this section.

Answer available at the end of this section.

Exercise 2.3.8.2

You observe two musical instruments that you cannot identify. One plays high-pitch sounds and the other plays low-pitch sounds. How could you determine which is which without hearing either of them play?

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

The relationship of the speed of sound v_{w}, its frequency f, and its wavelength λ is given by

$$v_{\mathrm{w}} = f\lambda,$$

which is the same relationship given for all waves.

In air, the speed of sound is related to air temperature T by

$$v_{\mathrm{w}} = (331\ \mathrm{m/s})\sqrt{\frac{T}{273\ \mathrm{K}}}.$$

v_{w} is the same for all frequencies and wavelengths.

Glossary

pitch
: the perception of the frequency of a sound

Answers

Exercise 2.3.8.1

Exercise 2.3.8.1

Sound and light both travel at definite speeds. The speed of sound is slower than the speed of light. The first firework is probably very close by, so the speed difference is not noticeable. The second firework is farther away, so the light arrives at your eyes noticeably sooner than the sound wave arrives at your ears.

Exercise 2.3.8.2

Exercise 2.3.8.2

Compare their sizes. High-pitch instruments are generally smaller than low-pitch instruments because they generate a smaller wavelength.

Compare their sizes. High-pitch instruments are generally smaller than low-pitch instruments because they generate a smaller wavelength.

This page titled 2.3.8: Speed of Sound, Frequency, and Wavelength is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.8: Speed of Sound, Frequency, and Wavelength by Andrew Park is licensed CC BY 4.0.

2.3.9: Doppler Effect and Sonic Booms

Learning Objectives

- Define Doppler effect, Doppler shift, and sonic boom.
- Describe the sounds produced by objects moving faster than the speed of sound.

The characteristic sound of a motorcycle buzzing by is an example of the Doppler effect. The high-pitch scream shifts dramatically to a lower-pitch roar as the motorcycle passes by a stationary observer. The closer the motorcycle brushes by, the more abrupt the shift. The faster the motorcycle moves, the greater the shift. We also hear this characteristic shift in frequency for passing race cars, airplanes, and trains. It is so familiar that it is used to imply motion and children often mimic it in play.

The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer. Although less familiar, this effect is easily noticed for a stationary source and moving observer. For example, if you ride a train past a stationary warning bell, you will hear the bell's frequency shift from high to low as you pass by. This change in frequency due to *relative* motion of source and observer is called a **Doppler shift**. The Doppler effect is named for the Austrian physicist and mathematician Christian Johann Doppler (1803–1853), who did experiments with both moving sources and moving observers. Doppler, for example, had musicians play on a moving open train car and also play standing next to the train tracks as a train passed by. Their music was observed both on and off the train, and changes in frequency were measured.

What causes the Doppler shift? Figure 2.3.9.1, Figure 2.3.9.2, and Figure 2.3.9.3 compare sound waves emitted by stationary and moving sources in a stationary air mass. Each disturbance spreads out spherically from the point where the sound was emitted. If the source is stationary, then all of the spheres representing the air compressions in the sound wave centered on the same point, and the stationary observers on either side see the same wavelength and frequency as emitted by the source, as in Figure 2.3.9.1 . If the source is moving, as in Figure 2.3.9.2, then the situation is different. Each compression of the air moves out in a sphere from the point where it was emitted, but the point of emission moves. This moving emission point causes the air compressions to be closer together on one side and farther apart on the other. Thus, the wavelength is shorter in the direction the source is moving (on the right in Figure 2.3.9.2), and longer in the opposite direction (on the left in Figure 2.3.9.2). Finally, if the observers move, as in Figure 2.3.9.3, the frequency at which they receive the compressions changes. The observer moving toward the source receives them at a higher frequency, and the person moving away from the source receives them at a lower frequency.



Figure 2.3.9.1: Sounds emitted by a source spread out in spherical waves. Because the source, observers, and air are stationary, the wavelength and frequency are the same in all directions and to all observers.

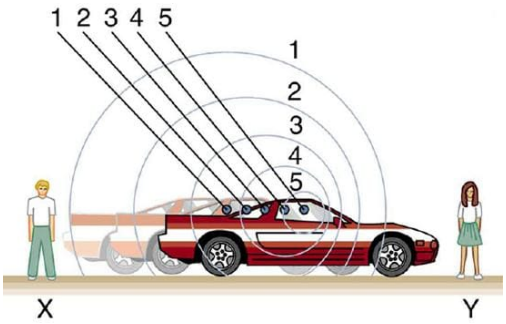


Figure 2.3.9.2: Sounds emitted by a source moving to the right spread out from the points at which they were emitted. The wavelength is reduced and, consequently, the frequency is increased in the direction of motion, so that the observer on the right hears a higher-pitch sound. The opposite is true for the observer on the left, where the wavelength is increased and the frequency is reduced.



Figure 2.3.9.3: The same effect is produced when the observers move relative to the source. Motion toward the source increases frequency as the observer on the right passes through more wave crests than she would if stationary. Motion away from the source decreases frequency as the observer on the left passes through fewer wave crests than he would if stationary.

We know that wavelength and frequency are related by $v_{\mathrm{w}} = f\lambda$, where v_{w} is the fixed speed of sound. The sound moves in a medium and has the same speed v_{w} in that medium whether the source is moving or not. Thus f multiplied by λ is a constant. Because the observer on the right in Figure 2.3.9.2 receives a shorter wavelength, the frequency she receives must be higher. Similarly, the observer on the left receives a longer wavelength, and hence he hears a lower frequency. The same thing happens in Figure 2.3.9.3. A higher frequency is received by the observer moving toward the source, and a lower frequency is received by an observer moving away from the source. In general, then, relative motion of source and observer toward one another increases the received frequency. Relative motion apart decreases frequency. The greater the relative speed is, the greater the effect.

THE DOPPLER EFFECT

The Doppler effect occurs not only for sound but for any wave when there is relative motion between the observer and the source. There are Doppler shifts in the frequency of sound, light, and water waves, for example. Doppler shifts can be used to determine velocity, such as when ultrasound is reflected from blood in a medical diagnostic. The recession of galaxies is determined by the observed frequency shift of light received from them measured against the frequency emitted by atoms in the laboratory. This implies much about the origins of the universe. Modern physics has been profoundly affected by observations of Doppler shifts.

Sonic Booms to Bow Wakes

What happens to the sound produced by a moving source, such as a jet airplane, that approaches or even exceeds the speed of sound? The answer to this question applies not only to sound but to all other waves as well.

Suppose a jet airplane is coming nearly straight at you, emitting a sound of frequency f_{s}. The greater the plane's speed v_{s}, the greater the Doppler shift and the greater the value observed for f_{obs}. Now, as v_{s} approaches the speed of sound, f_{obs} approaches infinity. At the speed of sound, this result means that in front of the source, each successive wave is superimposed on the previous one because the source moves forward at the speed of sound. The observer gets them all at the same instant, and so the frequency is infinite. (Before airplanes exceeded the speed of sound, some people argued it would be impossible because such constructive superposition would produce pressures great enough to destroy the airplane.) If the source exceeds the speed

of sound, no sound is received by the observer until the source has passed, so that the sounds from the approaching source are mixed with those from it when receding. This mixing appears messy, but something interesting happens—a sonic boom is created. (See Figure 2.3.9.4.)

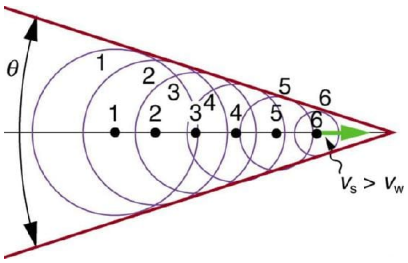


Figure 2.3.9.4: Sound waves from a source that moves faster than the speed of sound spread spherically from the point where they are emitted, but the source moves ahead of each. Constructive interference along the lines shown (actually a cone in three dimensions) creates a shock wave called a sonic boom. The faster the speed of the source, the smaller the angle θ.

There is constructive interference along the lines shown (a cone in three dimensions) from similar sound waves arriving there simultaneously. This superposition forms a disturbance called a **sonic boom**, a constructive interference of sound created by an object moving faster than sound. Inside the cone, the interference is mostly destructive, and so the sound intensity there is much less than on the shock wave. An aircraft creates two sonic booms, one from its nose and one from its tail. (See Figure 2.3.9.5.) During television coverage of space shuttle landings, two distinct booms could often be heard. These were separated by exactly the time it would take the shuttle to pass by a point. Observers on the ground often do not see the aircraft creating the sonic boom, because it has passed by before the shock wave reaches them, as seen in Figure 2.3.9.5. If the aircraft flies close by at low altitude, pressures in the sonic boom can be destructive and break windows as well as rattle nerves. Because of how destructive sonic booms can be, supersonic flights are banned over populated areas of the United States.

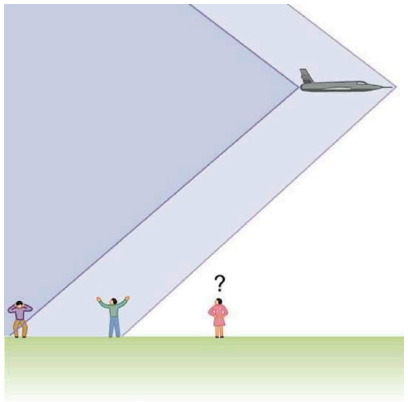


Figure 2.3.9.5: Two sonic booms, created by the nose and tail of an aircraft, are observed on the ground after the plane has passed by.

Sonic booms are one example of a broader phenomenon called bow wakes. A **bow wake**, such as the one in Figure 2.3.9.6, is created when the wave source moves faster than the wave propagation speed. Water waves spread out in circles from the point where created, and the bow wake is the familiar V-shaped wake trailing the source. A more exotic bow wake is created when a subatomic particle travels through a medium faster than the speed of light travels in that medium (the speed of light in vacuum is $c = 2.998 \times 10^8 \text{ m/s}$, and this is the maximum speed anything can move, but in water, for example, speed of light slows to $0.75c$, and other particles can move faster). If the particle creates light in its passage, that light spreads on a cone with an angle indicative of the speed of the particle, as illustrated in Figure 2.3.9.7. Such a bow wake is called Cherenkov radiation and is commonly observed in particle physics.



Figure 2.3.9.6: Bow wake created by a duck. Constructive interference produces the rather structured wake, while there is relatively little wave action inside the wake, where interference is mostly destructive. (credit: Horia Varlan, Flickr)



Figure 2.3.9.7: The blue glow in this research reactor pool is Cherenkov radiation caused by subatomic particles traveling faster than the speed of light in water. (credit: U.S. Nuclear Regulatory Commission)

Doppler shifts and sonic booms are interesting sound phenomena that occur in all types of waves. They can be of considerable use. For example, the Doppler shift in ultrasound can be used to measure blood velocity, while police use the Doppler shift in radar (a microwave) to measure car velocities. In meteorology, the Doppler shift is used to track the motion of storm clouds; such "Doppler Radar" can give velocity and direction and rain or snow potential of imposing weather fronts. In astronomy, we can examine the light emitted from distant galaxies and determine their speed relative to ours. As galaxies move away from us, their light is shifted to a lower frequency, and so to a longer wavelength—the so-called red shift. Such information from galaxies far, far away has allowed us to estimate the age of the universe (from the Big Bang) as about 14 billion years.

Exercise 2.3.9.1

Why did scientist Christian Doppler observe musicians both on a moving train and also from a stationary point not on the train?

Answer available at the end of this section.

Answer available at the end of this section.

Exercise 2.3.9.2

Describe a situation in your life when you might rely on the Doppler shift to help you either while driving a car or walking near traffic.

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- The Doppler shift is change in observed frequency of a wave (for example, a sound wave) due to relative motion of the source and the observer, with approaching source and/or observer increasing the observed frequency and receding source and/or observer decreasing the observed frequency.
- A sonic boom is constructive interference of sound created by an object moving faster than sound.
- A shock wave is a type of bow wake created when any wave source moves faster than the wave propagation speed.

Glossary

Doppler shift
: the change in wave frequency due to relative motion of source and observer

sonic boom
: a constructive interference of sound created by an object moving faster than sound

bow wake
: V-shaped disturbance created when the wave source moves faster than the wave propagation speed

Answers

Exercise 2.3.9.1

Exercise 2.3.9.1

Doppler needed to compare the perception of sound when the observer is stationary and the sound source moves, as well as when the sound source and the observer are both in motion.

Exercise 2.3.9.2

Exercise 2.3.9.2

If I am driving and I hear Doppler shift in an ambulance siren, I would be able to tell when it was getting closer and also if it has passed by. This would help me to know whether I needed to pull over and let the ambulance through

This page titled 2.3.9: Doppler Effect and Sonic Booms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.9: Doppler Effect and Sonic Booms by Andrew Park is licensed CC BY 4.0.

2.3.10: Oscillations and Waves (Exercise)

Conceptual Questions

5.1: Period and Frequency in Oscillations

1. If an oscillation occurs more rapidly, how does the period of oscillation change? How does the frequency change?

2. Give examples of natural phenomena that occur with constant frequency (or period) over a length of time.

5.2: Simple Harmonic Motion- A Special Periodic Motion

3. What conditions must be met to produce simple harmonic motion?

4. (a) If frequency is not constant for some oscillation, can the oscillation be simple harmonic motion?

(b) Can you think of any examples of harmonic motion where the frequency may depend on the amplitude?

5. Give an example of a simple harmonic oscillator, specifically noting how its frequency is independent of amplitude.

6. Explain why you expect an object made of a stiff material to vibrate at a higher frequency than a similar object made of a spongy material.

7. As you pass a freight truck with a trailer on a highway, you notice that its trailer is bouncing up and down slowly. Is it more likely that the trailer is heavily loaded or nearly empty? Explain your answer.

8. Some people modify cars to be much closer to the ground than when manufactured. Should they install stiffer springs? Explain your answer.

5.3: Forced Oscillations and Resonance

9. Give additional examples of resonance phenomenon from everyday life.

10. What "damps" the resonance of a forced oscillation? What are the dynamics of "damping" and what physical quantity is involved?

5.4: Waves

11. Give one example of a transverse wave and another of a longitudinal wave, being careful to note the relative directions of the disturbance and wave propagation in each.

12. What is the difference between propagation speed and the frequency of a wave? Does one or both affect wavelength? If so, how?

13. Given a wave of frequency ff traveling with wave speed v_{w} , what happens to the wave speed if the frequency doubles? What happens to the wavelength?

5.5: Wave Interference- Standing Waves and Beats

14. How does destructive interference occur? What do you see in a destructive interference?

15. Explain how noise-cancelling headphones work in terms of wave interference phenomena.

16. Explain how nodes and antinodes form on standing waves.

17. In the beat phenomenon, the loudness of the sound combined from two sound waves "wobbles" (it increases and decreases over time). Explain how the loudness of the combined sound becomes louder then quieter, even though individual sound waves remain at the same loudness, or amplitude.

5.7: Speed of Sound, Frequency, and Wavelength

18. How do sound vibrations of atoms differ from thermal motion?

19. When sound passes from one medium to another where its propagation speed is different, does its frequency or wavelength change? Explain your answer briefly.

20. For a high-fidelity speaker containing a woofer and a tweeter (see Figure 2.3.10.1 for example), why is the woofer larger than the tweeter?

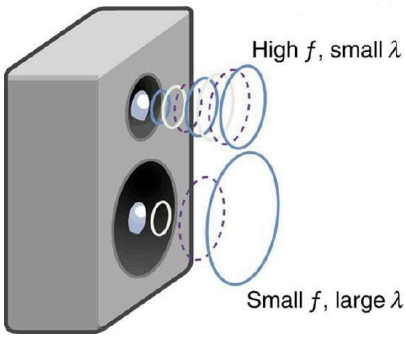


Figure 2.3.10.1: Because they travel at the same speed in a given medium, low-frequency sounds must have a greater wavelength than high-frequency sounds. Here, the lower-frequency sounds are emitted by the large speaker, called a woofer, while the higher-frequency sounds are emitted by the small speaker, called a tweeter.

5.8: Doppler Effect and Sonic Booms

21. Explain Doppler shift. How does the measured frequency increase or decrease?

22. Due to efficiency considerations related to its bow wake, the supersonic transport aircraft must maintain a cruising speed that is a constant ratio to the speed of sound (a constant Mach number). If the aircraft flies from warm air into colder air, should it increase or decrease its speed? Explain your answer.

23. When you hear a sonic boom, you often cannot see the plane that made it. Why is that?

This page titled 2.3.10: Oscillations and Waves (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 6.10: Oscillations and Waves (Exercise) by Andrew Park is licensed CC BY 4.0.

CHAPTER 3: CLASSICAL PHYSICS - THERMODYNAMICS, ELECTRICITY AND MAGNETISM, AND LIGHT

3.1: Electricity

3.1.1: Introduction to Electricity

3.1.2: Static Electricity and Charge- Conservation of Charge

3.1.3: Coulomb's Law

3.1.4: Electric Field- Concept of a Field Revisited

3.1.5: Electric Field Lines

3.1.6: Electric Potential and Potential Energy

3.1.7: Conductors and Applications of Electrostatics

3.1.8: Current

3.1.9: Ohm's Law- Resistance and Simple Circuits

3.1.10: Electric Power and Energy

3.1.11: Resistors in Series and Parallel

3.1.12: Electric Hazards and the Human Body

3.1.13: Electricity (Exercise)

3.2: Magnetism

3.2.1: Introduction to Magnetism

3.2.2: Magnets

3.2.3: Ferromagnets and Electromagnets

3.2.4: Magnetic Fields and Magnetic Field Lines

3.2.5: Magnetic Field Strength- Force on a Moving Charge in a Magnetic Field

3.2.6: Magnetic Force on a Current-Carrying Conductor

3.2.7: Motors and Meters

3.2.8: Magnetic Fields Produced by Currents- Ampere's Law

3.2.9: Induced Voltage and Magnetic Flux

3.2.10: Faraday's Law of Induction- Lenz's Law

3.2.11: Transformers

3.2.12: Alternating Current versus Direct Current

3.2.13: Magnetism (Exercise)

3.3: Light

3.3.1: Introduction to Light

3.3.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed

3.3.3: Production and Properties of Electromagnetic Waves

3.3.4: The Electromagnetic Spectrum- an Overview

3.3.5: The Electromagnetic Spectrum- Application Notes

3.3.6: Reflection

3.3.7: Refraction

3.3.8: Dispersion- The Rainbow and Prisms
3.3.9: Polarization
3.3.10: Image Formation by Lenses
3.3.11: Image Formation by Mirrors
3.3.12: Light (Exercises)

3.4: Thermal Physics

3.4.1: Introduction to Thermal Physics
3.4.2: Temperature
3.4.3: The Ideal Gas Law
3.4.4: Heat
3.4.5: Heat Transfer Methods
3.4.6: Temperature Change and Heat Capacity
3.4.7: Phase Change and Latent Heat
3.4.8: The First Law of Thermodynamics
3.4.9: The First Law of Thermodynamics and Heat Engine Processes
3.4.10: Introduction to the Second Law of Thermodynamics- Heat Engines and Their Efficiency
3.4.11: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated
3.4.12: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy
3.4.13: Statistical Interpretation of Entropy and the Second Law of Thermodynamics- The Underlying Explanation
3.4.14: Applications of Thermodynamics- Heat Pumps and Refrigerators
3.4.15: Thermal Physics (Exercises)

This page titled Chapter 3: Classical Physics - Thermodynamics, Electricity and Magnetism, and Light is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

3.1: ELECTRICITY

3.1.1: Introduction to Electricity

3.1.2: Static Electricity and Charge- Conservation of Charge

3.1.3: Coulomb's Law

3.1.4: Electric Field- Concept of a Field Revisited

3.1.5: Electric Field Lines

3.1.6: Electric Potential and Potential Energy

3.1.7: Conductors and Applications of Electrostatics

3.1.8: Current

3.1.9: Ohm's Law- Resistance and Simple Circuits

3.1.10: Electric Power and Energy

3.1.11: Resistors in Series and Parallel

3.1.12: Electric Hazards and the Human Body

3.1.13: Electricity (Exercise)

This page titled 3.1: Electricity is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

3.1.1: Introduction to Electricity



Figure 3.1.1.1: Static electricity from this plastic slide causes the child's hair to stand on end. The sliding motion stripped electrons away from the child's body, leaving an excess of positive charges, which repel each other along each strand of hair. (credit: Ken Bosma/Wikimedia Commons)

The image of American politician and scientist Benjamin Franklin (1706–1790) flying a kite in a thunderstorm is familiar to every schoolchild. (See Figure.) In this experiment, Franklin demonstrated a connection between lightning and **static electricity**. Sparks were drawn from a key hung on a kite string during an electrical storm. These sparks were like those produced by static electricity, such as the spark that jumps from your finger to a metal doorknob after you walk across a wool carpet. What Franklin demonstrated in his dangerous experiment was a connection between phenomena on two different scales: one the grand power of an electrical storm, the other an effect of more human proportions. Connections like this one reveal the underlying unity of the laws of nature, an aspect we humans find particularly appealing.

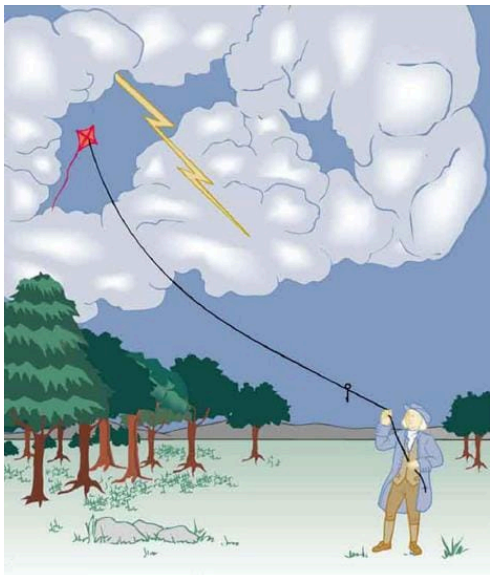


Figure 3.1.1.2: When Benjamin Franklin demonstrated that lightning was related to static electricity, he made a connection that is now part of the evidence that all directly experienced forces except the gravitational force are manifestations of the electromagnetic force.

Our understanding of electricity and the range of phenomena that are electrical in nature has vastly expanded since those early days. For example, atomic and molecular interactions, such as the forces of friction, cohesion, and adhesion, are now known to

be manifestations of the **electromagnetic force**. Static electricity is just one aspect of the electromagnetic force, which also includes moving electricity and magnetism. Furthermore, All the macroscopic forces that we experience directly, such as the sensations of touch and the tension in a rope, are due to the electromagnetic force, one of the four fundamental forces in nature. The gravitational force, another fundamental force, is actually sensed through the electromagnetic interaction of molecules, such as between those in our feet and those on the top of a bathroom scale. (The other two fundamental forces, the strong nuclear force and the weak nuclear force, cannot be sensed on the human scale.)

In this chapter, we will develop the concepts of electric fields and electric potential (also known as *voltage*), which are used to describe electric force and electrical energy. Consider, for example, great amounts of electrical energy stored in batteries or transmitted cross-country through power lines, or consider electrical signals sent in our nervous systems at molecular levels, with *ions* crossing cell membranes and transferring information, or the role of electricity in our household appliances and devices, including your laptop and smartphones. We will look at the laws that help us describe electric circuits and begin to explore some of the many applications of electricity.

Glossary

static electricity
: a buildup of electric charge on the surface of an object

electromagnetic force
: one of the four fundamental forces of nature; the electromagnetic force consists of static electricity, moving electricity and magnetism

This page titled 3.1.1: Introduction to Electricity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.1: Introduction to Electricity by Andrew Park is licensed CC BY 4.0.

3.1.2: Static Electricity and Charge- Conservation of Charge

Learning Objectives

- Define electric charge, and describe how the two types of charge interact.
- Describe three common situations that generate static electricity.
- State the law of conservation of charge.



Figure 3.1.2.1: Borneo amber was mined in Sabah, Malaysia, from shale-sandstone-mudstone veins. When a piece of amber is rubbed with a piece of silk, the amber gains more electrons, giving it a net negative charge. At the same time, the silk, having lost electrons, becomes positively charged. (credit: Sebakoamber, Wikimedia Commons)

What makes plastic wrap cling? Static electricity. Not only are applications of static electricity common these days, its existence has been known since ancient times. The first record of its effects dates to ancient Greeks who noted more than 500 years B.C. that polishing amber temporarily enabled it to attract bits of straw (see Figure 3.1.2.1). The very word *electric* derives from the Greek word for amber (*electron*).

Many of the characteristics of static electricity can be explored by rubbing things together. Rubbing creates the spark you get from walking across a wool carpet, for example. Static cling generated in a clothes dryer and the attraction of straw to recently polished amber also result from rubbing. Similarly, lightning results from air movements under certain weather conditions. You can also rub a balloon on your hair, and the static electricity created can then make the balloon cling to a wall. We also have to be cautious of static electricity, especially in dry climates. When we pump gasoline, we are warned to discharge ourselves (after sliding across the seat) on a metal surface before grabbing the gas nozzle. Attendants in hospital operating rooms must wear booties with aluminum foil on the bottoms to avoid creating sparks which may ignite the oxygen being used.

Some of the most basic characteristics of static electricity include:

- The effects of static electricity are explained by a physical quantity not previously introduced, called electric charge.
- There are only two types of charge, one called positive and the other called negative.
- Like charges repel, whereas unlike charges attract.
- The force between charges decreases with distance.

How do we know there are two types of **electric charge**? When various materials are rubbed together in controlled ways, certain combinations of materials always produce one type of charge on one material and the opposite type on the other. By convention, we call one type of charge “positive”, and the other type “negative.” For example, when glass is rubbed with silk, the glass becomes positively charged and the silk negatively charged. Since the glass and silk have opposite charges, they attract one another like clothes that have rubbed together in a dryer. Two glass rods rubbed with silk in this manner will repel one another, since each rod has positive charge on it. Similarly, two silk cloths so rubbed will repel, since both cloths have negative charge. Figure 3.1.2.2 shows how these simple materials can be used to explore the nature of the force between charges.

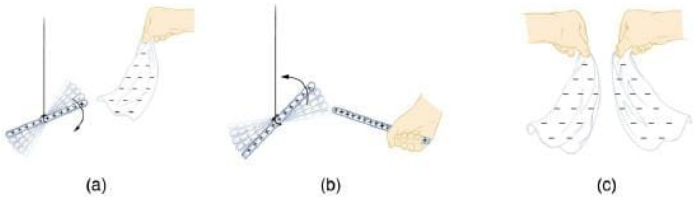


Figure 3.1.2.2: A glass rod becomes positively charged when rubbed with silk, while the silk becomes negatively charged. (a) The glass rod is attracted to the silk because their charges are opposite. (b) Two similarly charged glass rods repel. (c) Two similarly charged silk cloths repel.

More sophisticated questions arise. Where do these charges come from? Can you create or destroy charge? Is there a smallest unit of charge? Exactly how does the force depend on the amount of charge and the distance between charges? Such questions obviously occurred to Benjamin Franklin and other early researchers, and they interest us even today.

Charge Carried by Electrons and Protons

Franklin wrote in his letters and books that he could see the effects of electric charge but did not understand what caused the phenomenon. Today we have the advantage of knowing that normal matter is made of atoms, and that atoms contain positive and negative charges, usually in equal amounts.

Figure 3.1.2.3 shows a simple model of an atom with negative **electrons** orbiting its positive nucleus. The nucleus is positive due to the presence of positively charged **protons**. Nearly all charge in nature is due to electrons and protons, which are two of the three building blocks of most matter. (The third is the neutron, which is neutral, carrying no charge.) Other charge-carrying particles are observed in cosmic rays and nuclear decay, and are created in particle accelerators. All but the electron and proton survive only a short time and are quite rare by comparison.

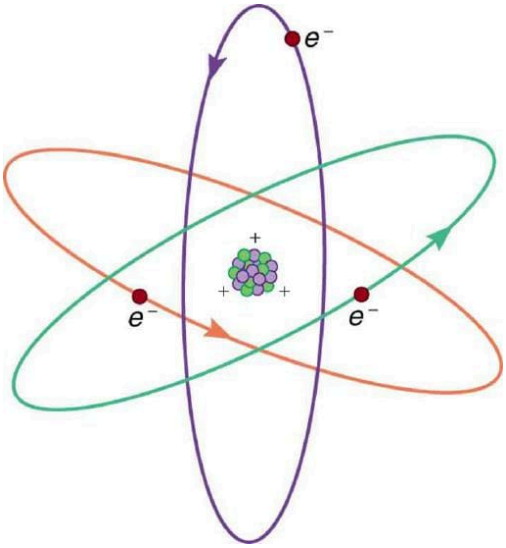


Figure 3.1.2.3: This simplified (and not to scale) view of an atom is called the planetary model of the atom. Negative electrons orbit a much heavier positive nucleus, as the planets orbit the much heavier sun. There the similarity ends, because forces in the atom are electromagnetic, whereas those in the planetary system are gravitational. Normal macroscopic amounts of matter contain immense numbers of atoms and molecules and, hence, even greater numbers of individual negative and positive charges.

The charges of electrons and protons are identical in magnitude but opposite in sign. Furthermore, all charged objects in nature are integral multiples of this basic quantity of charge, meaning that all charges are made of combinations of a basic unit of charge. Usually, charges are formed by combinations of electrons and protons. The magnitude of this basic charge is

$$|q_e| = 1.60 \times 10^{-19}\text{C}.$$

The symbol q is commonly used for charge and the subscript e indicates the charge of a single electron (or proton).

The SI unit of charge is the coulomb (C). The number of protons needed to make a charge of 1.00 C is

$$1.00\text{C} \times \frac{1 \text{ proton}}{1.60 \times 10^{-19}\text{C}} = 6.25 \times 10^{18} \text{ protons.}$$

Similarly, 6.25×10^{18} electrons have a combined charge of –1.00 coulomb. Just as there is a smallest bit of an element (an atom), there is a smallest bit of charge. There is no directly observed charge smaller than $|q_e|$, and all observed charges are integral multiples of $|q_e|$.

THINGS GREAT AND SMALL: THE SUBMICROSCOPIC ORIGIN OF CHARGE

With the exception of exotic, short-lived particles, all charge in nature is carried by electrons and protons. Electrons carry the charge we have named negative. Protons carry an equal-magnitude charge that we call positive. (See Figure 3.1.2.4.) Electron and proton charges are considered fundamental building blocks, since all other charges are integral multiples of those carried by electrons and protons. Electrons and protons are also two of the three fundamental building blocks of ordinary matter. The neutron is the third and has zero total charge.

Figure 3.1.2.4 shows a person touching a Van de Graaff generator and receiving excess positive charge. The expanded view of a hair shows the existence of both types of charges but an excess of positive. The repulsion of these positive like charges causes the strands of hair to repel other strands of hair and to stand up. The further blowup shows an artist's conception of an electron and a proton perhaps found in an atom in a strand of hair.

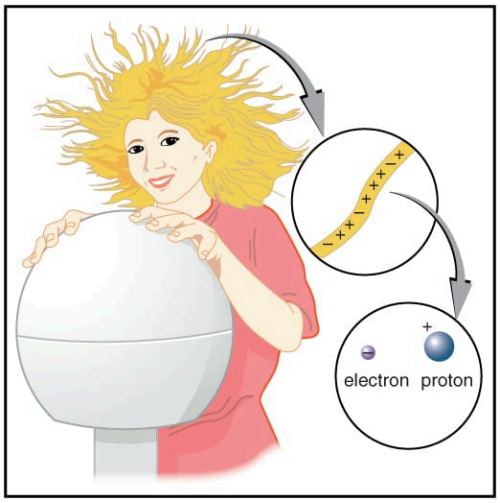


Figure 3.1.2.4: When this person touches a Van de Graaff generator, she receives an excess of positive charge, causing her hair to stand on end. The charges in one hair are shown. An artist's conception of an electron and a proton illustrate the particles carrying the negative and positive charges. We cannot really see these particles with visible light because they are so small (the electron seems to be an infinitesimal point), but we know a great deal about their measurable properties, such as the charges they carry.

The electron seems to have no substructure; in contrast, when the substructure of protons is explored by scattering extremely energetic electrons from them, it appears that there are point-like particles inside the proton. These sub-particles, named quarks, have never been directly observed, but they are believed to carry fractional charges as seen in Figure 3.1.2.5. Charges on electrons and protons and all other directly observable particles are unitary, but these quark substructures carry charges of either $-\frac{1}{3}$ or $+\frac{2}{3}$. There are continuing attempts to observe fractional charge directly and to learn of the properties of quarks, which are perhaps the ultimate substructure of matter.

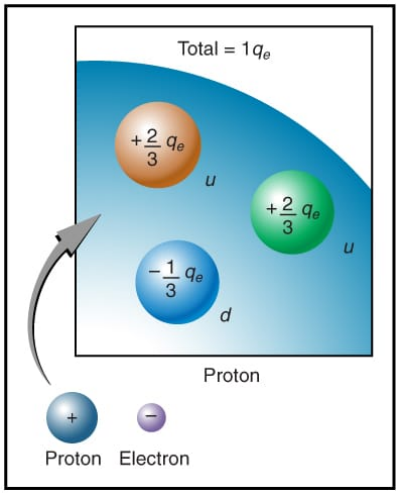


Figure 3.1.2.5: Artist's conception of fractional quark charges inside a proton. A group of three quark charges add up to the single positive charge on the proton: $-\frac{1}{3}q_e + \frac{2}{3}q_e + \frac{2}{3}q_e = +1q_e$.

Separation of Charge in Atoms

Charges in atoms and molecules can be separated—for example, by rubbing materials together. Some atoms and molecules have a greater affinity for electrons than others and will become negatively charged by close contact in rubbing, leaving the other material positively charged. (See Figure 3.1.2.6.) Positive charge can similarly be induced by rubbing. Methods other than rubbing can also separate charges. Batteries, for example, use combinations of substances that interact in such a way as to separate charges. Chemical interactions may transfer negative charge from one substance to the other, making one battery terminal negative and leaving the first one positive.

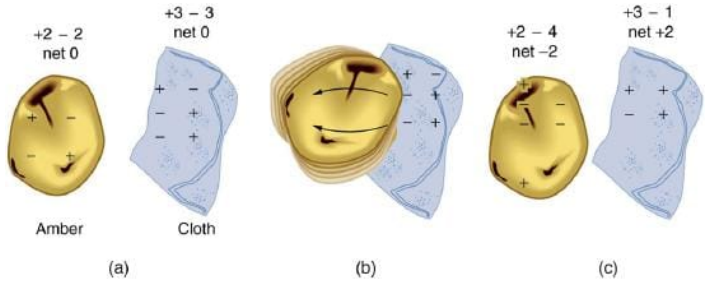


Figure 3.1.2.6: When materials are rubbed together, charges can be separated, particularly if one material has a greater affinity for electrons than another. (a) Both the amber and cloth are originally neutral, with equal positive and negative charges. Only a tiny fraction of the charges are involved, and only a few of them are shown here. (b) When rubbed together, some negative charge is transferred to the amber, leaving the cloth with a net positive charge. (c) When separated, the amber and cloth now have net charges, but the absolute value of the net positive and negative charges will be equal.

No charge is actually created or destroyed when charges are separated as we have been discussing. Rather, existing charges are moved about. In fact, in all situations the total amount of charge is always constant. This universally obeyed law of nature is called the **law of conservation of charge**.

LAW OF CONSERVATION OF CHARGE

Total charge is constant in any process.

In more exotic situations, such as in particle accelerators, mass, Δm, can be created from energy in the amount $\Delta m = \frac{E}{c^2}$. Sometimes, the created mass is charged, such as when an electron is created. Whenever a charged particle is created, another having an opposite charge is always created along with it, so that the total charge created is zero. Usually, the two particles are "matter-antimatter" counterparts. For example, an antielectron would usually be created at the same time as an electron. The antielectron has a positive charge (it is called a positron), and so the total charge created is zero. (See Figure 3.1.2.7.) All

particles have antimatter counterparts with opposite signs. When matter and antimatter counterparts are brought together, they completely annihilate one another. By annihilate, we mean that the mass of the two particles is converted to energy E, again obeying the relationship $\Delta m = \frac{E}{c^2}$. Since the two particles have equal and opposite charge, the total charge is zero before and after the annihilation; thus, total charge is conserved.

MAKING CONNECTIONS: CONSERVATION LAWS

Only a limited number of physical quantities are universally conserved. Charge is one—energy, momentum, and angular momentum are others. Because they are conserved, these physical quantities are used to explain more phenomena and form more connections than other, less basic quantities. We find that conserved quantities give us great insight into the rules followed by nature and hints to the organization of nature. Discoveries of conservation laws have led to further discoveries, such as the weak nuclear force and the quark substructure of protons and other particles.

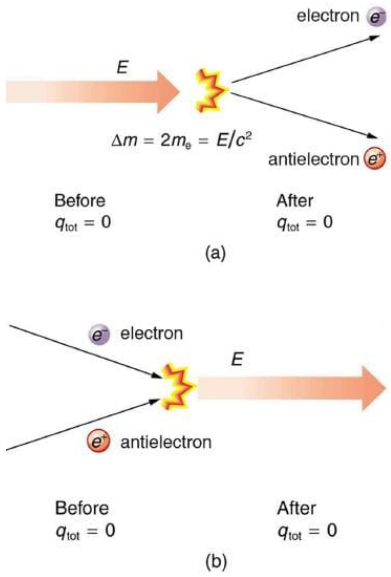


Figure 3.1.2.7: (a) When enough energy is present, it can be converted into matter. Here the matter created is an electron–antielectron pair. (m_e is the electron's mass.) The total charge before and after this event is zero. (b) When matter and antimatter collide, they annihilate each other; the total charge is conserved at zero before and after the annihilation.

The law of conservation of charge is absolute—it has never been observed to be violated. Charge, then, is a special physical quantity, joining a very short list of other quantities in nature that are always conserved. Other conserved quantities include energy, momentum, and angular momentum.

Section Summary

- There are only two types of charge, which we call positive and negative.
- Like charges repel, unlike charges attract, and the force between charges decreases with the square of the distance.
- The vast majority of positive charge in nature is carried by protons, while the vast majority of negative charge is carried by electrons.
- The electric charge of one electron is equal in magnitude and opposite in sign to the charge of one proton.
- An ion is an atom or molecule that has nonzero total charge due to having unequal numbers of electrons and protons.
- The SI unit for charge is the coulomb (C), with protons and electrons having charges of opposite sign but equal magnitude; the magnitude of this basic charge $|q_e|$ is

$$|q_e| = 1.60 \times 10^{-19}\text{C}.$$

- Whenever charge is created or destroyed, equal amounts of positive and negative are involved.
- Most often, existing charges are separated from neutral objects to obtain some net charge.
- Both positive and negative charges exist in neutral objects and can be separated by rubbing one object with another. For macroscopic objects, negatively charged means an excess of electrons and positively charged means a depletion of

electrons.
- The law of conservation of charge ensures that whenever a charge is created, an equal charge of the opposite sign is created at the same time.

Glossary

electric charge
: a physical property of an object that causes it to be attracted toward or repelled from another charged object; each charged object generates and is influenced by a force called an electromagnetic force

law of conservation of charge
: states that whenever a charge is created, an equal amount of charge with the opposite sign is created simultaneously

electron
: a particle orbiting the nucleus of an atom and carrying the smallest unit of negative charge

proton
: a particle in the nucleus of an atom and carrying a positive charge equal in magnitude and opposite in sign to the amount of negative charge carried by an electron

This page titled 3.1.2: Static Electricity and Charge- Conservation of Charge is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.2: Static Electricity and Charge- Conservation of Charge by Andrew Park is licensed CC BY 4.0.

3.1.3: Coulomb's Law

Learning Objectives

- State Coulomb's law in terms of how the electrostatic force changes with the distance between two objects.
- Calculate the electrostatic force between two charged point forces, such as electrons or protons.
- Compare the electrostatic force to the gravitational attraction for a proton and an electron; for a human and the Earth.



Figure 3.1.3.1: This NASA image of Arp 87 shows the result of a strong gravitational attraction between two galaxies. In contrast, at the subatomic level, the electrostatic attraction between two objects, such as an electron and a proton, is far greater than their mutual attraction due to gravity. (credit: NASA/HST)

Through the work of scientists in the late 18th century, the main features of the **electrostatic force**—the existence of two types of charge, the observation that like charges repel, unlike charges attract, and the decrease of force with distance—were eventually refined, and expressed as a mathematical formula. The mathematical formula for the electrostatic force is called **Coulomb's law** after the French physicist Charles Coulomb (1736–1806), who performed experiments and first proposed a formula to calculate it.

Definition: COULOMB'S LAW

$$F = k\frac{|q_1 q_2|}{r^2}.$$

Coulomb's law calculates the magnitude of the force F between two point charges, q_1 and q_2, separated by a distance r. In SI units, the constant k is equal to

$$k = 8.988 \times 10^9 \frac{\mathrm{N \cdot m^2}}{\mathrm{C^2}} \approx 8.99 \times 10^9 \frac{\mathrm{N \cdot m^2}}{\mathrm{C^2}}.$$

The electrostatic force is a vector quantity and is expressed in units of newtons. The force is understood to be along the line joining the two charges. (See Figure 3.1.3.2.)

Although the formula for Coulomb's law is simple, it was no mean task to prove it. The experiments Coulomb did, with the primitive equipment then available, were difficult. Modern experiments have verified Coulomb's law to great precision. For example, it has been shown that the force is inversely proportional to distance between two objects squared ($F \propto 1/r^2$) to an accuracy of 1 part in 10^{16}. No exceptions have ever been found, even at the small distances within the atom.

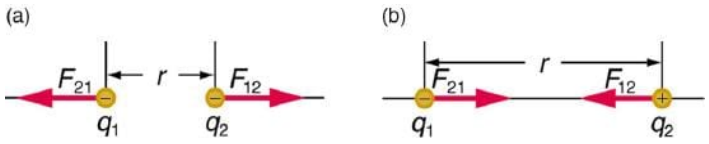


Figure 3.1.3.2: The magnitude of the electrostatic force F between point charges q_1 and q_2 separated by a distance r is given by Coulomb's law. Note that Newton's third law (every force exerted creates an equal and opposite force) applies as usual—the force on q_1 is equal in magnitude and opposite in direction to the force it exerts on q_2. (a) Like charges. (b) Unlike charges.

Example 3.1.3.1: How Strong is the Coulomb Force Relative to the Gravitational Force?

Compare the electrostatic force between an electron and proton separated by 0.530×10^{-10} m with the gravitational force between them. This distance is their average separation in a hydrogen atom.

Strategy

To compare the two forces, we first compute the electrostatic force using Coulomb's law, $F = k\frac{|q_1q_2|}{r^2}$. We then calculate the gravitational force using Newton's universal law of gravitation. Finally, we take a ratio to see how the forces compare in magnitude.

Solution

Entering the given and known information about the charges and separation of the electron and proton into the expression of Coulomb's law yields

$$F = k\frac{|q_1q_2|}{r^2}$$
$$= \left(8.99\times10^{9}\ \mathrm{N\cdot m^2/C^2}\right)\times\frac{\left(1.60\times10^{-19}\mathrm{C}\right)\left(1.60\times10^{-19}\mathrm{C}\right)}{\left(0.530\times10^{-10}\ \mathrm{m}\right)^2}$$

Thus the Coulomb force is

$$F = 8.19\times10^{-8}\ \mathrm{N}.$$

The charges are opposite in sign, so this is an attractive force. This is a very large force for an electron—it would cause an acceleration of 8.99×10^{22} m/s^2(verification is left as an end-of-section problem).The gravitational force is given by Newton's law of gravitation as:

$$F_G = G\frac{mM}{r^2},$$

where $G = 6.67\times10^{-11}\ \mathrm{N\cdot m^2/kg^2}$. Here m and M represent the electron and proton masses, which can be found in the appendices. Entering values for the knowns yields

$$F_G = \left(6.67\times10^{-11}\ \mathrm{N\cdot m^2/kg^2}\right)\times\frac{\left(9.11\times10^{-31}\ \mathrm{kg}\right)\left(1.67\times10^{-27}\ \mathrm{kg}\right)}{\left(0.530\times10^{-10}\ \mathrm{m}\right)^2} = 3.61\times10^{-47}\ \mathrm{N}$$

This is also an attractive force, although it is traditionally shown as positive since gravitational force is always attractive. The ratio of the magnitude of the electrostatic force to gravitational force in this case is, thus,

$$\frac{F}{F_G} = 2.27\times10^{39}.$$

Discussion

> This is a remarkably large ratio! Note that this will be the ratio of electrostatic force to gravitational force for an electron and a proton at any distance (taking the ratio before entering numerical values shows that the distance cancels). This ratio gives some indication of just how much larger the Coulomb force is than the gravitational force between two of the most common particles in nature.

As the example implies, gravitational force is completely negligible on a small scale, where the interactions of individual charged particles are important. On a large scale, such as between the Earth and a person, the reverse is true. Most objects are nearly electrically neutral, and so attractive and repulsive **Coulomb forces** nearly cancel. Gravitational force on a large scale dominates interactions between large objects because it is always attractive, while Coulomb forces tend to cancel.

Section Summary

- Frenchman Charles Coulomb was the first to publish the mathematical equation that describes the electrostatic force between two objects.
- Coulomb's law gives the magnitude of the force between point charges. It is

$$F = k\frac{|q_1 q_2|}{r^2},$$

  where q_1 and q_1 are two point charges separated by a distance r, and $k \approx 8.99 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2$

- This Coulomb force is extremely basic, since most charges are due to point-like particles. It is responsible for all electrostatic effects and underlies most macroscopic forces.
- The Coulomb force is extraordinarily strong compared with the gravitational force, another basic force—but unlike gravitational force it can cancel, since it can be either attractive or repulsive.
- The electrostatic force between two subatomic particles is far greater than the gravitational force between the same two particles.

Glossary

Coulomb's law
: the mathematical equation calculating the electrostatic force vector between two charged particles

Coulomb force
: another term for the electrostatic force

electrostatic force
: the amount and direction of attraction or repulsion between two charged bodies

This page titled 3.1.3: Coulomb's Law is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.3: Coulomb's Law by Andrew Park is licensed CC BY 4.0.

3.1.4: Electric Field- Concept of a Field Revisited

Learning Objectives

- Describe a force field and calculate the strength of an electric field due to a point charge.
- Calculate the force exerted on a test charge by an electric field.
- Explain the relationship between electrical force (F) on a test charge and electrical field strength (E).

Contact forces, such as between a baseball and a bat, are explained on the small scale by the interaction of the charges in atoms and molecules in close proximity. They interact through forces that include the **Coulomb force**. Action at a distance is a force between objects that are not close enough for their atoms to "touch." That is, they are separated by more than a few atomic diameters.

For example, a charged rubber comb attracts neutral bits of paper from a distance via the Coulomb force. It is very useful to think of an object being surrounded in space by a **force field**. The force field carries the force to another object (called a test object) some distance away.

Concept of a Field

A field is a way of conceptualizing and mapping the force that surrounds any object and acts on another object at a distance without apparent physical connection. For example, the gravitational field surrounding the earth (and all other masses) represents the gravitational force that would be experienced if another mass were placed at a given point within the field.

In the same way, the Coulomb force field surrounding any charge extends throughout space. Using Coulomb's law, $F = k\left|q_1 q_2\right| / r^2$, its magnitude is given by the equation $F = k|qQ|/r^2$, for a **point charge** (a particle having a charge Q) acting on a **test charge** q at a distance r (see Figure 3.1.4.1). Both the magnitude and direction of the Coulomb force field depend on Q and the test charge q.

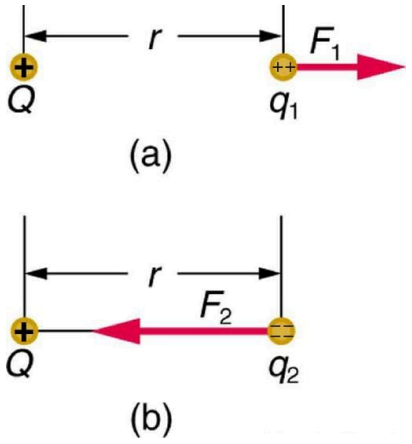


Figure 3.1.4.1: The Coulomb force field due to a positive charge Q is shown acting on two different charges. Both charges are the same distance from Q. (a) Since q_1 is positive, the force F_1 acting on it is repulsive. (b) The charge q_2 is negative and greater in magnitude than q_1, and so the force F_2 acting on it is attractive and stronger than F_1. The Coulomb force field is thus not unique at any point in space, because it depends on the test charges q_1 and q_2 as well as the charge Q.

To simplify things, we would prefer to have a field that depends only on Q and not on the test charge q. The electric field is defined in such a manner that it represents only the charge creating it and is unique at every point in space. Specifically, the electric field E is defined to be the ratio of the Coulomb force to the test charge:

$$\mathbf{E} = \frac{\mathbf{F}}{q},$$

where $\mathbf{F}$ is the electrostatic force (or Coulomb force) exerted on a positive test charge q. It is understood that $\mathbf{E}$ is in the same direction as $\mathbf{F}$. It is also assumed that q is so small that it does not alter the charge distribution creating the electric field. The units of electric field are newtons per coulomb (N/C). If the electric field is known, then the electrostatic force on any charge q is simply obtained by multiplying charge times electric field, or $\mathbf{F} = q\mathbf{E}$. Consider the electric field due to a point charge Q. According to Coulomb's law, the force it exerts on a test charge q is $F = k|qQ|/r^2$. Thus the magnitude of the electric field, E, for a point charge is

$$E = \left|\frac{F}{q}\right| = k\left|\frac{qQ}{qr^2}\right| = k\frac{|Q|}{r^2}.$$

Since the test charge cancels, we see that

$$E = k\frac{|Q|}{r^2}.$$

The electric field is thus seen to depend only on the charge Q and the distance r; it is completely independent of the test charge q.

Example 3.1.4.1: Calculating the Electric Field of a Point Charge

Calculate the strength and direction of the electric field E due to a point charge of 2.00 nC (nano-Coulombs) at a distance of 5.00 mm from the charge.

Strategy

We can find the electric field created by a point charge by using the equation $E = kQ/r^2$.

Solution

Here $Q = 2.00 \times 10^{-9}$ C and $r = 5.00 \times 10^{-3}$ m. Entering those values into the above equation gives

$$\begin{aligned} E &= k\frac{Q}{r^2} \\ &= \left(8.99 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2\right) \times \frac{\left(2.00 \times 10^{-9}\text{C}\right)}{\left(5.00 \times 10^{-3} \text{ m}\right)^2} \\ &= 7.19 \times 10^5 \text{ N/C}. \end{aligned}$$

Discussion

This **electric field strength** is the same at any point 5.00 mm away from the charge Q that creates the field. It is positive, meaning that it has a direction pointing away from the charge Q.

Example 3.1.4.2: Calculating the Force Exerted on a Point Charge by an Electric Field

What force does the electric field found in the previous example exert on a point charge of $-0.250\ \mu$C?

Strategy

Since we know the electric field strength and the charge in the field, the force on that charge can be calculated using the definition of electric field $\mathbf{E} = \mathbf{F}/q$ rearranged to $\mathbf{F} = q\mathbf{E}$.

Solution

The magnitude of the force on a charge $q = -0.250\ \mu$C exerted by a field of strength $E = 7.20 \times 10^5$ N/C is thus,

$$\begin{aligned} F &= -qE \\ &= \left(0.250 \times 10^{-6}\text{C}\right)\left(7.20 \times 10^{5}\ \text{N/C}\right) \\ &= 0.180\ \text{N}. \end{aligned}$$

Because q is negative, the force is directed opposite to the direction of the field.

Discussion

The force is attractive, as expected for unlike charges. (The field was created by a positive charge and here acts on a negative charge.) The charges in this example are typical of common static electricity, and the modest attractive force obtained is similar to forces experienced in static cling and similar situations.

Section Summary

- The electrostatic force field surrounding a charged object extends out into space in all directions.
- The electrostatic force exerted by a point charge on a test charge at a distance r depends on the charge of both charges, as well as the distance between the two.
- The electric field $\mathbf{E}$ is defined to be

$$\mathbf{E} = \frac{\mathbf{F}}{q},$$

where $\mathbf{F}$ is the Coulomb or electrostatic force exerted on a small positive test charge q. $\mathbf{E}$ has units of N/C.

- The magnitude of the electric field $\mathbf{E}$ created by a point charge Q is

$$\mathbf{E} = k\frac{|Q|}{r^2}.$$

where r is the distance from Q. The electric field $\mathbf{E}$ is a vector and fields due to multiple charges add like vectors.

Glossary

field
: a map of the amount and direction of a force acting on other objects, extending out into space

point charge
: A charged particle, designated Q, generating an electric field

test charge
: A particle (designated q) with either a positive or negative charge set down within an electric field generated by a point charge

This page titled 3.1.4: Electric Field- Concept of a Field Revisited is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.4: Electric Field- Concept of a Field Revisited by Andrew Park is licensed CC BY 4.0.

3.1.5: Electric Field Lines

Learning Objectives

- Calculate the total force (magnitude and direction) exerted on a test charge from more than one charge
- Describe an electric field diagram of a positive point charge; of a negative point charge with twice the magnitude of positive charge
- Draw the electric field lines between two points of the same charge; between two points of opposite charge.

Drawings using lines to represent **electric fields** around charged objects are very useful in visualizing field strength and direction. Since the electric field has both magnitude and direction, it is a vector. Like all **vectors**, the electric field can be represented by an arrow that has length proportional to its magnitude and that points in the correct direction. (We have used arrows extensively to represent force vectors, for example.)

Figure 3.1.5.1 shows two pictorial representations of the same electric field created by a positive point charge Q. Figure 3.1.5.1 (b) shows the standard representation using continuous lines. Figure 3.1.5.1(a) shows numerous individual arrows with each arrow representing the force on a test charge q.

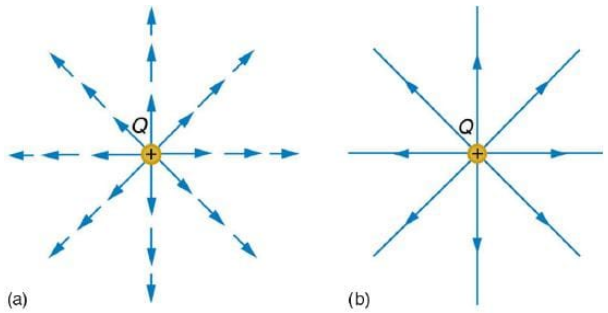


Figure 3.1.5.1: Two equivalent representations of the electric field due to a positive charge Q. (a) Arrows representing the electric field's magnitude and direction. (b) In the standard representation, the arrows are replaced by continuous field lines having the same direction at any point as the electric field. The closeness of the lines is directly related to the strength of the electric field. A test charge placed anywhere will feel a force in the direction of the field line; this force will have a strength proportional to the density of the lines (being greater near the charge, for example).

Note that the electric field is defined for a positive test charge q, so that the field lines point away from a positive charge and toward a negative charge. (See Figure 3.1.5.2.) The electric field strength is exactly proportional to the number of field lines per unit area, since the magnitude of the electric field for a point charge is $E = k|Q|/r^2$ and area is proportional to r^2. This pictorial representation, in which field lines represent the direction and their closeness (that is, their areal density or the number of lines crossing a unit area) represents strength, is used for all fields: electrostatic, gravitational, magnetic, and others.

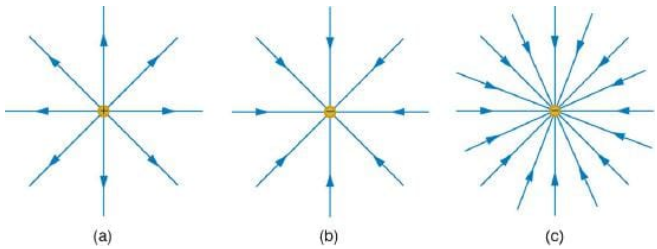


Figure 3.1.5.2: The electric field surrounding three different point charges. (a) A positive charge. (b) A negative charge of equal magnitude. (c) A larger negative charge.

In many situations, there are multiple charges. The total electric field created by multiple charges is the vector sum of the individual fields created by each charge.

Figure 3.1.5.3 shows how the electric field from two point charges can be drawn by finding the total field at representative points

and drawing electric field lines consistent with those points. While the electric fields from multiple charges are more complex than those of single charges, some simple features are easily noticed.

For example, the field is weaker between like charges, as shown by the lines being farther apart in that region. (This is because the fields from each charge exert opposing forces on any charge placed between them.) (See Figure 3.1.5.3 and Figure 3.1.5.4(a).) Furthermore, at a great distance from two like charges, the field becomes identical to the field from a single, larger charge.

Figure 3.1.5.4(b) shows the electric field of two unlike charges. The field is stronger between the charges. In that region, the fields from each charge are in the same direction, and so their strengths add. The field of two unlike charges is weak at large distances, because the fields of the individual charges are in opposite directions and so their strengths subtract. At very large distances, the field of two unlike charges looks like that of a smaller single charge.

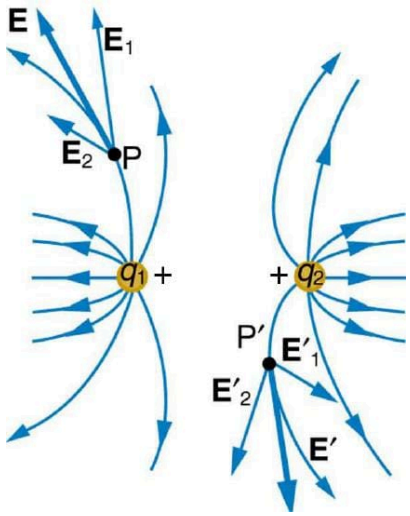


Figure 3.1.5.3: Two positive point charges q_1 and q_2 produce the resultant electric field shown. The field is calculated at representative points and then smooth field lines drawn following the rules outlined in the text.

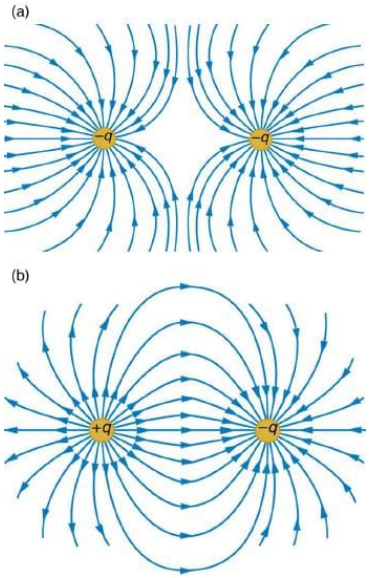


Figure 3.1.5.4: (a) Two negative charges produce the fields shown. It is very similar to the field produced by two positive charges, except that the directions are reversed. The field is clearly weaker between the charges. The individual forces on a test charge in that region are in opposite directions. (b) Two opposite charges produce the field shown, which is stronger in the region between the charges.

We use electric field lines to visualize and analyze electric fields (the lines are a pictorial tool, not a physical entity in themselves). The properties of electric field lines for any charge distribution can be summarized as follows:

1. Field lines must begin on positive charges and terminate on negative charges, or at infinity in the hypothetical case of isolated charges.

2. The number of field lines leaving a positive charge or entering a negative charge is proportional to the magnitude of the charge.
3. The strength of the field is proportional to the closeness of the field lines—more precisely, it is proportional to the number of lines per unit area perpendicular to the lines.
4. The direction of the electric field is tangent to the field line at any point in space.
5. Field lines can never cross.

The last property means that the field is unique at any point. The field line represents the direction of the field; so if they crossed, the field would have two directions at that location (an impossibility if the field is unique).

Section Summary

- Drawings of electric field lines are useful visual tools. The properties of electric field lines for any charge distribution are that:
- Field lines must begin on positive charges and terminate on negative charges, or at infinity in the hypothetical case of isolated charges.
- The number of field lines leaving a positive charge or entering a negative charge is proportional to the magnitude of the charge.
- The strength of the field is proportional to the closeness of the field lines—more precisely, it is proportional to the number of lines per unit area perpendicular to the lines.
- The direction of the electric field is tangent to the field line at any point in space.
- Field lines can never cross.

Glossary

electric field
: a three-dimensional map of the electric force extended out into space from a point charge

electric field lines
: a series of lines drawn from a point charge representing the magnitude and direction of force exerted by that charge

vector
: a quantity with both magnitude and direction

vector addition
: mathematical combination of two or more vectors, including their magnitudes, directions, and positions

This page titled 3.1.5: Electric Field Lines is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.5: Electric Field Lines by Andrew Park is licensed CC BY 4.0.

3.1.6: Electric Potential and Potential Energy

Learning Objectives

- Define electric potential and electric potential energy.
- Describe the relationship between electric potential difference and electric field.
- Describe the relationship between electric potential and electrical potential energy.
- Explain electron volt and its usage in submicroscopic process.

When a free positive charge q is accelerated by an electric field, such as shown in Figure 3.1.6.1, it is given kinetic energy. The process is analogous to an object being accelerated by a gravitational field. It is as if the charge is going down an electrical hill where its electric potential energy is converted to kinetic energy. Let us explore the work done on a charge q by the electric field in this process, so that we may develop a definition of electric potential energy.

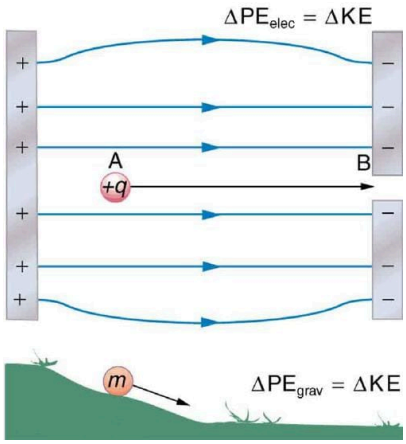


Figure 3.1.6.1: A charge accelerated by an electric field is analogous to a mass going down a hill. In both cases potential energy is converted to another form. Work is done by a force, but since this force is conservative, we can write $W = -\Delta\mathrm{PE}$.

The electrostatic or Coulomb force is conservative, which means that the work done on q is independent of the path taken. This is exactly analogous to the gravitational force in the absence of dissipative forces such as friction. When a force is conservative, it is possible to define a potential energy associated with the force, and it is usually easier to deal with the potential energy (because it depends only on position) than to calculate the work done directly from force ($W = \langle \boldsymbol{F}, \boldsymbol{d} \rangle$, where d is displacement and F is the force).

We use the letters PE to denote electric potential energy, which has units of joules (J). The change in potential energy, $\triangle\mathrm{PE}$, is crucial, since the work done by a conservative force is the negative of the change in potential energy; that is, $W = -\Delta\mathrm{PE}$. For example, work W done to accelerate a positive charge from rest is positive and results from a loss in PE, or a negative $\triangle\mathrm{PE}$. There must be a minus sign in front of $\triangle\mathrm{PE}$ to make W positive. PE can be found at any point by taking one point as a reference and calculating the work needed to move a charge to the other point.

Definition: POTENTIAL ENERGY

$W = -\Delta\mathrm{PE}$. For example, work W done to accelerate a positive charge from rest is positive and results from a loss in PE, or a negative $\Delta\mathrm{PE}$. There must be a minus sign in front of $\Delta\mathrm{PE}$ to make W positive. PE can be found at any point by taking one point as a reference and calculating the work needed to move a charge to the other point.

Gravitational potential energy and electric potential energy are quite analogous. Potential energy accounts for work done by a conservative force and gives added insight regarding energy and energy transformation without the necessity of dealing with the force directly. It is much more common, for example, to use the concept of voltage (related to electric potential energy) than to deal with the electric field (related to Coulomb force) directly.

Given some conservative force F and displacement d under the force, the work done and the change in potential energy can be calculated as, $W = \langle \boldsymbol{F}, \boldsymbol{d} \rangle$ and $\Delta \mathrm{PE} = -\mathrm{W} = \langle -\mathbf{F}, \boldsymbol{d} \rangle$. For electric force, the force is given by the product of electric charge and the electric field, $\boldsymbol{F} = q\boldsymbol{E}$, where q is the charge experiencing the force and E is the electric field at the location of the charge. So the potential energy change due to work done by electric force is $\Delta \mathrm{PE} = q(\langle -\mathbf{E}, \boldsymbol{d} \rangle)$. If we define change in **electric potential** V as $\Delta V = \langle -\mathbf{E}, \boldsymbol{d} \rangle$, then the electric potential energy PE is simply expressed in terms of electric potential, $\mathrm{PE} = qV$, or,

$$V = \frac{\mathrm{PE}}{q},$$

electric potential energy per charge.

Definition: ELECTRIC POTENTIAL

Electric potential is the electric potential energy per unit charge.

$$V = \frac{\mathrm{PE}}{q}$$

With potential energy, the case often is that its value at a single point has no significant meaning but what is important is the difference in potential energy. From the difference in potential energy, we are able to calculate other quantities, such as change in kinetic energy (if no force other than the conservative force acts) or work needing to be done by other forces (if other forces act). So likewise, rather than the electric potential itself, we are often interested in difference in electric potential ΔV between two points, where,

$$\Delta V = V_{\mathrm{B}} - V_{\mathrm{A}} = \frac{\Delta \mathrm{PE}}{q}.$$

The **potential difference** between points A and B, $V_{\mathrm{B}} - V_{\mathrm{A}}$, is thus defined to be the change in potential energy of a charge q moved from A to B, divided by the charge. Units of potential difference are joules per coulomb, given the name volt (V) after Alessandro Volta.

$$1 \text{ V} = 1\frac{\mathrm{J}}{\mathrm{C}}$$

Definition: POTENTIAL DIFFERENCE

The potential difference between points A and B, $V_{\mathrm{B}} - V_{\mathrm{A}}$, is defined to be the change in potential energy of a charge q moved from A to B, divided by the charge. Units of potential difference are joules per coulomb, given the name volt (V) after Alessandro Volta.

$$1 \text{ V} = 1\frac{\mathrm{J}}{\mathrm{C}}$$

The familiar term **voltage** is the common name for potential difference. Keep in mind that whenever a voltage is quoted, it is understood to be the potential difference between two points. For example, every battery has two terminals, and its voltage is the potential difference between them. More fundamentally, the point you choose to be zero volts is arbitrary. This is analogous to the fact that gravitational potential energy has an arbitrary zero, such as sea level or perhaps a lecture hall floor.

In summary, the relationship between potential difference (or voltage) and electrical potential energy is given by

$$\Delta V = \frac{\Delta \mathrm{PE}}{q} \text{ and } \Delta \mathrm{PE} = q\Delta V.$$

POTENTIAL DIFFERENCE AND ELECTRICAL POTENTIAL ENERGY

The relationship between potential difference (or voltage) and electrical potential energy is given by

$$\Delta V = \frac{\Delta \text{PE}}{q} \text{ and } \Delta \text{PE} = q\Delta V.$$

The second equation is equivalent to the first.

Voltage is not the same as energy. Voltage is the energy per unit charge. Thus a motorcycle battery and a car battery can both have the same voltage (more precisely, the same potential difference between battery terminals), yet one stores much more energy than the other since $\Delta \text{PE} = q\Delta V$. The car battery can move more charge than the motorcycle battery, although both are 12 V batteries.

Example 3.1.6.1: Calculating Energy

Suppose you have a 12.0 V motorcycle battery that can move 5000 C of charge, and a 12.0 V car battery that can move 60,000 C of charge. How much energy does each deliver? (Assume that the numerical value of each charge is accurate to three significant figures.)

Strategy

To say we have a 12.0 V battery means that its terminals have a 12.0 V potential difference. When such a battery moves charge, it puts the charge through a potential difference of 12.0 V, and the charge is given a change in potential energy equal to $\Delta \text{PE} = q\Delta V$.

So to find the energy output, we multiply the charge moved by the potential difference.

Solution

For the motorcycle battery, $q = 5000 \text{ C}$ and $\Delta V = 12.0 \text{ V}$. The total energy delivered by the motorcycle battery is

$$\begin{aligned} \Delta \text{PE}_{\text{cycle}} &= (5000\text{C})(12.0 \text{ V}) \\ &= (5000\text{C})(12.0 \text{ J/C}) \\ &= 6.00 \times 10^4 \text{ J}. \end{aligned}$$

Similarly, for the car battery, $q = 60,000 \text{ C}$ and

$$\begin{aligned} \Delta \text{PE}_{\text{car}} &= (60,000\text{C})(12.0 \text{ V}) \\ &= 7.20 \times 10^5 \text{ J}. \end{aligned}$$

Discussion

While voltage and energy are related, they are not the same thing. The voltages of the batteries are identical, but the energy supplied by each is quite different. Note also that as a battery is discharged, some of its energy is used internally and its terminal voltage drops, such as when headlights dim because of a low car battery. The energy supplied by the battery is still calculated as in this example, but not all of the energy is available for external use.

Note that the energies calculated in the previous example are absolute values. The change in potential energy for the battery is negative, since it loses energy. These batteries, like many electrical systems, actually move negative charge—electrons in particular. The batteries repel electrons from their negative terminals (A) through whatever circuitry is involved and attract them to their positive terminals (B) as shown in Figure 3.1.6.2. The change in potential is $\Delta V = V_{\text{B}} - V_{\text{A}} = +12 \text{ V}$ and the charge q is negative, so that $\Delta \text{PE} = q\Delta V$ is negative, meaning the potential energy of the battery has decreased when q has moved from A to B.

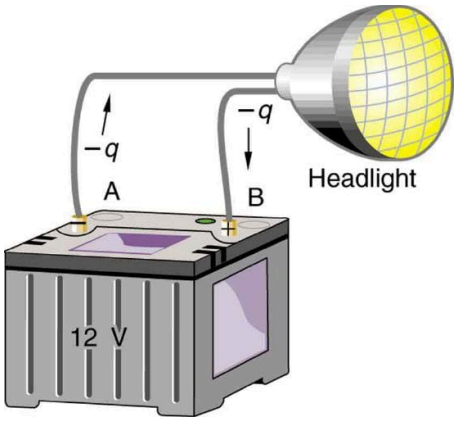


Figure 3.1.6.2: A battery moves negative charge from its negative terminal through a headlight to its positive terminal. Appropriate combinations of chemicals in the battery separate charges so that the negative terminal has an excess of negative charge, which is repelled by it and attracted to the excess positive charge on the other terminal. In terms of potential, the positive terminal is at a higher voltage than the negative. Inside the battery, both positive and negative charges move.

Example 3.1.6.2: How Many Electrons Move through a Headlight Each Second?

When a 12.0 V car battery runs a single 30.0 W headlight, how many electrons pass through it each second?

Strategy

To find the number of electrons, we must first find the charge that moved in 1.00 s. The charge moved is related to voltage and energy through the equation $\Delta \mathrm{PE} = q\Delta V$. A 30.0 W lamp uses 30.0 joules per second. Since the battery loses energy, we have $\Delta \mathrm{PE} = -30.0\ \mathrm{J}$ and, since the electrons are going from the negative terminal to the positive, we see that $\Delta V = +12.0\ \mathrm{V}$.

Solution

To find the charge q moved, we solve the equation $\Delta \mathrm{PE} = q\Delta V$:

$$q = \frac{\Delta \mathrm{PE}}{\Delta V}.$$

Entering the values for $\triangle \mathrm{PE}$ and ΔV, we get

$$q = \frac{-30.0\ \mathrm{J}}{+12.0\ \mathrm{V}} = \frac{-30.0\ \mathrm{J}}{+12.0\ \mathrm{J/C}} = -2.50\mathrm{C}.$$

The number of electrons $\mathrm{n_e}$ is the total charge divided by the charge per electron. That is,

$$\mathrm{n_e} = \frac{-2.50\mathrm{C}}{-1.60 \times 10^{-19}\mathrm{C/e^-}} = 1.56 \times 10^{19}\ \text{electrons}.$$

Discussion

This is a very large number. It is no wonder that we do not ordinarily observe individual electrons with so many being present in ordinary systems. In fact, electricity had been in use for many decades before it was determined that the moving charges in many circumstances were negative. Positive charge moving in the opposite direction of negative charge often produces identical effects; this makes it difficult to determine which is moving or whether both are moving.

The Electron Volt

The energy per electron is very small in macroscopic situations like that in the previous example—a tiny fraction of a joule. But on a submicroscopic scale, such energy per particle (electron, proton, or ion) can be of great importance. For example, even a tiny fraction of a joule can be great enough for these particles to destroy organic molecules and harm living tissue. The particle may

do its damage by direct collision, or it may create harmful X-rays, which can also inflict damage. It is useful to have an energy unit related to submicroscopic effects. Figure 3.1.6.3 shows a situation related to the definition of such an energy unit. An electron is accelerated between two charged metal plates as it might be in an old-model television tube or oscilloscope. The electron is given kinetic energy that is later converted to another form—light in the television tube, for example. (Note that downhill for the electron is uphill for a positive charge.) Since energy is related to voltage by $\Delta \mathrm{PE} = q\Delta V$, we can think of the joule as a coulomb-volt.

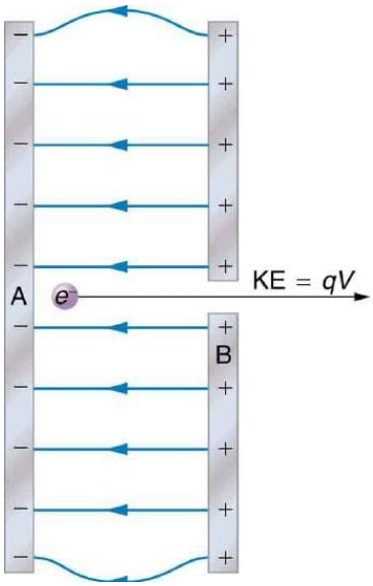


Figure 3.1.6.3: A typical electron gun accelerates electrons using a potential difference between two metal plates. The energy of the electron in electron volts is numerically the same as the voltage between the plates. For example, a 5000 V potential difference produces 5000 eV electrons.

Definition: ELECTRON VOLT

On the submicroscopic scale, it is more convenient to define an energy unit called the electron volt (eV), which is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,

$$\begin{aligned} 1\mathrm{eV} &= \left(1.60 \times 10^{-19}\mathrm{C}\right)(1\ \mathrm{V}) = \left(1.60 \times 10^{-19}\mathrm{C}\right)(1\ \mathrm{J/C}) \\ &= 1.60 \times 10^{-19}\ \mathrm{J}. \end{aligned}$$

An electron accelerated through a potential difference of 1 V is given an energy of 1 eV. It follows that an electron accelerated through 50 V is given 50 eV. A potential difference of 100,000 V (100 kV) will give an electron an energy of 100,000 eV (100 keV), and so on. Similarly, an ion with a double positive charge accelerated through 100 V will be given 200 eV of energy. These simple relationships between accelerating voltage and particle charges make the electron volt a simple and convenient energy unit in such circumstances.

CONNECTIONS: ENERGY UNITS

The electron volt (eV) is the most common energy unit for submicroscopic processes. This will be particularly noticeable in the chapters on modern physics. Energy is so important to so many subjects that there is a tendency to define a special energy unit for each major topic. There are, for example, calories for food energy, and kilowatt-hours for electrical energy.

The electron volt is commonly employed in submicroscopic processes—chemical valence energies and molecular and nuclear binding energies are among the quantities often expressed in electron volts. For example, about 5 eV of energy is required to break up certain organic molecules. If a proton is accelerated from rest through a potential difference of 30 kV, it is given an energy of 30 keV (30,000 eV) and it can break up as many as 6000 of these molecules ($30,000\ \mathrm{eV} \div 5\mathrm{eV\ per\ molecule} = 6000\ \mathrm{molecules}$). Nuclear decay energies are on the order of 1 MeV (1,000,000 eV) per event and can, thus, produce significant biological damage.

Conservation of Energy

The total energy of a system is conserved if there is no net addition (or subtraction) of work or heat transfer. For conservative forces, such as the electrostatic force, conservation of energy states that mechanical energy is a constant.

Mechanical energy is the sum of the kinetic energy and potential energy of a system; that is, $\mathrm{KE} + \mathrm{PE} = \text{ constant}$. A loss of PE of a charged particle becomes an increase in its KE. Here PE is the electric potential energy. Conservation of energy is stated in equation form as

$$\mathrm{KE} + \mathrm{PE} = \text{ constant}$$

or

$$\mathrm{KE_i} + \mathrm{PE_i} = \mathrm{KE_f} + \mathrm{PE_f},$$

where i and f stand for initial and final conditions. As we have found many times before, considering energy can give us insights and facilitate problem solving.

Example 3.1.6.3: Electrical Potential Energy Converted to Kinetic Energy

Calculate the final speed of a free electron accelerated from rest through a potential difference of 100 V. (Assume that this numerical value is accurate to three significant figures.)

Strategy

We have a system with only conservative forces. Assuming the electron is accelerated in a vacuum, and neglecting the gravitational force (we will check on this assumption later), all of the electrical potential energy is converted into kinetic energy. We can identify the initial and final forms of energy to be $\mathrm{KE_i} = 0, \mathrm{KE_f} = 1/2mv^2, \mathrm{PE_i} = qV, \text{ and } \mathrm{PE_f} = 0$.

Solution

Conservation of energy states that

$$\mathrm{KE_i} + \mathrm{PE_i} = \mathrm{KE_f} + \mathrm{PE_f}.$$

Entering the forms identified above, we obtain

$$qV = \frac{mv^2}{2}.$$

We solve this for v:

$$v = \sqrt{\frac{2qV}{m}}.$$

Entering values for q,V, and m gives

$$\begin{aligned} v &= \sqrt{\frac{2\left(-1.60 \times 10^{-19}\mathrm{C}\right)\left(-100\ \mathrm{J/C}\right)}{9.11 \times 10^{-31}\ \mathrm{kg}}} \\ &= 5.93 \times 10^{6}\ \mathrm{m/s}. \end{aligned}$$

Discussion

Note that both the charge and the initial voltage are negative, as in Figure 3.1.6.3. We know that electrostatic forces on small particles are generally very large compared with the gravitational force. The large final speed confirms that the gravitational force is indeed negligible here. The large speed also indicates how easy it is to accelerate electrons with small voltages because of their very small mass. Voltages much higher than the 100 V in this problem are typically used in electron guns.

Section Summary

- Electric potential is potential energy per unit charge.
- The potential difference between points A and B, $V_{\mathrm{B}} - V_{\mathrm{A}}$, defined to be the change in potential energy of a charge q moved from A to B, is equal to the change in potential energy divided by the charge, Potential difference is commonly called voltage, represented by the symbol ΔV.

$$.\Delta V = \frac{\Delta \mathrm{PE}}{q} \text{ and } \Delta \mathrm{PE} = q\Delta V$$

- An electron volt is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,

$$\begin{aligned} 1\mathrm{eV} &= \left(1.60 \times 10^{-19}\mathrm{C}\right)(1\ \mathrm{V}) = \left(1.60 \times 10^{-19}\mathrm{C}\right)(1\ \mathrm{J/C}) \\ &= 1.60 \times 10^{-19}\ \mathrm{J}. \end{aligned}$$

- Mechanical energy is the sum of the kinetic energy and potential energy of a system, that is, $\mathrm{KE} + \mathrm{PE}$ This sum is a constant.

Glossary

electric potential
: potential energy per unit charge

potential difference (or voltage)
: change in potential energy of a charge moved from one point to another, divided by the charge; units of potential difference are joules per coulomb, known as volt

electron volt
: the energy given to a fundamental charge accelerated through a potential difference of one volt

mechanical energy
: sum of the kinetic energy and potential energy of a system; this sum is a constant

This page titled 3.1.6: Electric Potential and Potential Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.6: Electric Potential and Potential Energy by Andrew Park is licensed CC BY 4.0.

3.1.7: Conductors and Applications of Electrostatics

Learning Objectives

- List the three properties of a conductor in electrostatic equilibrium.
- Explain why the electric field is zero inside a conductor in electrostatic equilibrium.
- Describe how a lightning rod works.
- Name several real-world applications of the study of electrostatics.

Conductors contain **free charges** that move easily. When excess charge is placed on a conductor or the conductor is put into a static electric field, charges in the conductor quickly respond to reach a steady state called **electrostatic equilibrium**.

Figure 3.1.7.1 shows the effect of an electric field on free charges in a conductor. The free charges move until the field is perpendicular to the conductor's surface. There can be no component of the field parallel to the surface in electrostatic equilibrium, since, if there were, it would produce further movement of charge.

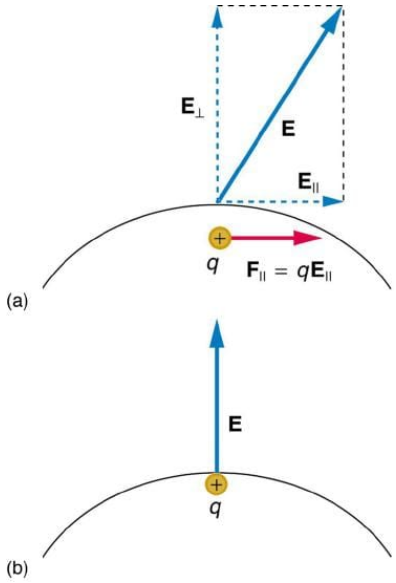


Figure 3.1.7.1: When an electric field $\mathbf{E}$ is applied to a conductor, free charges inside the conductor move until the field is perpendicular to the surface. (a) The electric field is a vector quantity, with both parallel and perpendicular components. The parallel component ($\mathbf{E}_{\|}$) exerts a force ($\mathbf{F}_{\|}$) on the free charge q, which moves the charge until $\mathbf{F}_{\|} = 0$. (b) The resulting field is perpendicular to the surface. The free charge has been brought to the conductor's surface, leaving electrostatic forces in equilibrium. A positive free charge is shown in the figure for the example, but the result holds similarly for a negative free charge.

A conductor placed in an **electric field** will be **polarized**. Figure 3.1.7.2 shows the result of placing a neutral conductor in an originally uniform electric field. The field becomes stronger near the conductor but entirely disappears inside it.

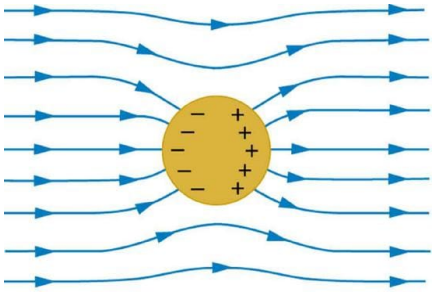


Figure 3.1.7.2: This illustration shows a spherical conductor in static equilibrium with an originally uniform electric field. Free charges move within the conductor, polarizing it, until the electric field lines are perpendicular to the surface. The field lines end on excess negative charge on one section of the surface and begin again on excess positive charge on the opposite side. No electric field exists inside the conductor, since free charges in the conductor would continue moving in response to any field until it was neutralized.

MISCONCEPTION ALERT: ELECTRIC FIELD INSIDE A CONDUCTOR

Excess charges placed on a spherical conductor repel and move until they are evenly distributed, as shown in Figure 3.1.7.3. Excess charge is forced to the surface until the field inside the conductor is zero. Outside the conductor, the field is exactly the same as if the conductor were replaced by a point charge at its center equal to the excess charge.

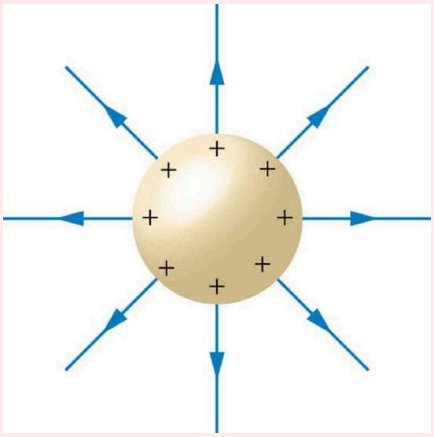


Figure 3.1.7.3: The mutual repulsion of excess positive charges on a spherical conductor distributes them uniformly on its surface. The resulting electric field is perpendicular to the surface and zero inside. Outside the conductor, the field is identical to that of a point charge at the center equal to the excess charge.

PROPERTIES OF A CONDUCTOR IN ELECTROSTATIC EQUILIBRIUM

1. The electric field is zero inside a conductor.
2. Just outside a conductor, the electric field lines are perpendicular to its surface, ending or beginning on charges on the surface.
3. Any excess charge resides entirely on the surface or surfaces of a conductor.

The properties of a conductor are consistent with the situations already discussed and can be used to analyze any conductor in electrostatic equilibrium. This can lead to some interesting new insights, such as described below.

How can a very uniform electric field be created? Consider a system of two metal plates with opposite charges on them, as shown in Figure 3.1.7.4. The properties of conductors in electrostatic equilibrium indicate that the electric field between the plates will be uniform in strength and direction. Except near the edges (which can be ignored when plates are close to each other), the excess charges distribute themselves uniformly, producing field lines that are uniformly spaced (hence uniform in strength) and perpendicular to the surfaces (hence uniform in direction, since the plates are flat).

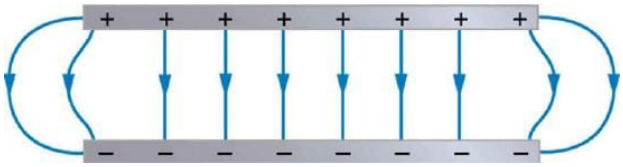


Figure 3.1.7.4: Two metal plates with equal, but opposite, excess charges. The field between them is uniform in strength and direction except near the edges. One use of such a field is to produce uniform acceleration of charges between the plates, such as in the electron gun of a TV tube.

In fact, this is a particular example of electric circuit elements called **capacitors**. This particular example is called a **parallel-plate capacitor** and can be used to analyze properties of capacitors in general, such as how many charges can be stored using two nearby conductors set at a certain electric potential difference (voltage) apart. Capacitors find wide use in electric circuits alongside the resistors and inductors that you will learn about later.

Electric Fields on Uneven Surfaces

So far we have considered excess charges on a smooth, symmetrical conductor surface. What happens if a conductor has sharp corners or is pointed? Excess charges on a nonuniform conductor become concentrated at the sharpest points. Additionally, excess charge may move on or off the conductor at the sharpest points. To see how and why this happens, consider the charged conductor in Figure 3.1.7.5. The electrostatic repulsion of like charges is most effective in moving them apart on the flattest surface, and so they become least concentrated there. The same effect is produced on a conductor by an externally applied electric field, as seen in Figure 3.1.7.5(c). Since the field lines must be perpendicular to the surface, more of them are concentrated on the most curved parts.

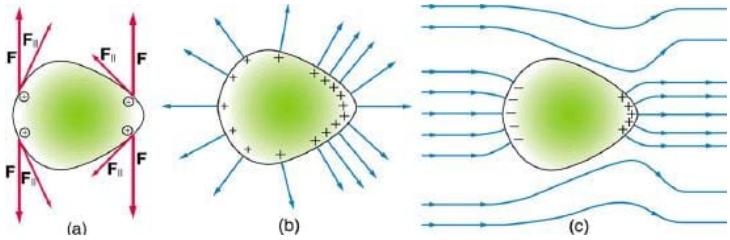


Figure 3.1.7.5: Excess charge on a nonuniform conductor becomes most concentrated at the location of greatest curvature. (a) The forces between identical pairs of charges at either end of the conductor are identical, but the components of the forces parallel to the surface are different. It is $\mathbf{F}_{\parallel}$ that moves the charges apart once they have reached the surface. (b) $\mathbf{F}_{\parallel}$ is smallest at the more pointed end, the charges are left closer together, producing the electric field shown. (c) An uncharged conductor in an originally uniform electric field is polarized, with the most concentrated charge at its most pointed end.

Applications of Conductors

On a very sharply curved surface, such as shown in Figure 3.1.7.6, the charges are so concentrated at the point that the resulting electric field can be great enough to remove them from the surface. This can be useful.

Lightning rods work best when they are most pointed. The large charges created in storm clouds induce an opposite charge on a building that can result in a lightning bolt hitting the building. The induced charge is bled away continually by a lightning rod, preventing the more dramatic lightning strike.

Of course, we sometimes wish to prevent the transfer of charge rather than to facilitate it. In that case, the conductor should be very smooth and have as large a radius of curvature as possible (see Figure 3.1.7.7, which shows a Van de Graaff generator). Smooth surfaces are used on high-voltage transmission lines, for example, to avoid leakage of charge into the air.

Another device that makes use of some of these principles is a **Faraday cage**. This is a metal shield that encloses a volume. All electrical charges will reside on the outside surface of this shield, and there will be no electrical field inside. A Faraday cage is used to prohibit stray electrical fields in the environment from interfering with sensitive measurements, such as the electrical signals inside a nerve cell.

During electrical storms if you are driving a car, it is best to stay inside the car as its metal body acts as a Faraday cage with zero electrical field inside. If in the vicinity of a lightning strike, its effect is felt on the outside of the car and the inside is unaffected,

provided you remain totally inside. This is also true if an active (“hot”) electrical wire was broken (in a storm or an accident) and fell on your car.

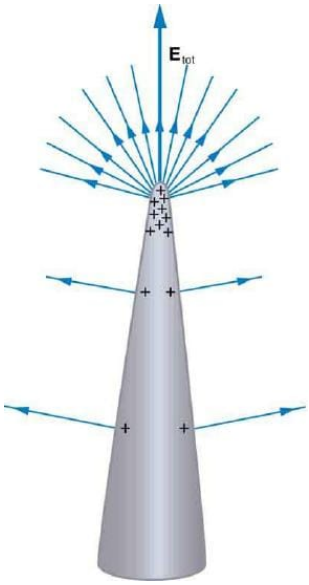


Figure 3.1.7.6: A very pointed conductor has a large charge concentration at the point. The electric field is very strong at the point and can exert a force large enough to transfer charge on or off the conductor. Lightning rods are used to prevent the buildup of large excess charges on structures and, thus, are pointed.



Figure 3.1.7.7: (a) A lightning rod is pointed to facilitate the transfer of charge. (credit: Romaine, Wikimedia Commons) (b) This Van de Graaff generator has a smooth surface with a large radius of curvature to prevent the transfer of charge and allow a large voltage to be generated. The mutual repulsion of like charges is evident in the person’s hair while touching the metal sphere. (credit: Jon ‘ShakataGaNai’ Davis/ Wikimedia Commons).

Below are additional applications of **electrostatics**, the study of electricity in electrostatic equilibrium, many of them using conductors for their properties just discussed above.

The Van de Graaff Generator

Van de Graaff generators (or Van de Graaffs) are not only spectacular devices used to demonstrate high voltage due to static electricity—they are also used for serious research. The first was built by Robert Van de Graaff in 1931 (based on original suggestions by Lord Kelvin) for use in nuclear physics research. Figure 3.1.7.8 shows a schematic of a large research version. Van de Graaffs utilize both smooth and pointed surfaces, and conductors and insulators to generate large static charges and, hence, large voltages.

A very large excess charge can be deposited on the sphere, because it moves quickly to the outer surface. Practical limits arise because the large electric fields polarize and eventually ionize surrounding materials, creating free charges that neutralize excess charge or allow it to escape. Nevertheless, voltages of 15 million volts are well within practical limits.

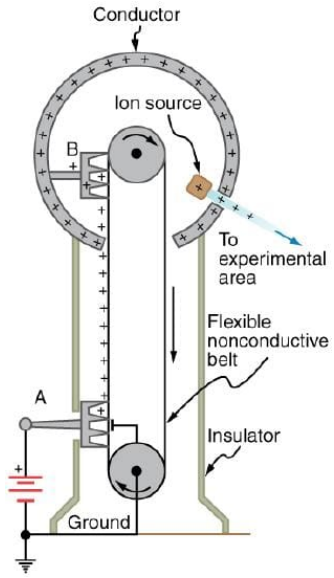


Figure 3.1.7.8: Schematic of Van de Graaff generator. A battery (A) supplies excess positive charge to a pointed conductor, the points of which spray the charge onto a moving insulating belt near the bottom. The pointed conductor (B) on top in the large sphere picks up the charge. (The induced electric field at the points is so large that it removes the charge from the belt.) This can be done because the charge does not remain inside the conducting sphere but moves to its outside surface. An ion source inside the sphere produces positive ions, which are accelerated away from the positive sphere to high velocities.

TAKE-HOME EXPERIMENT: ELECTROSTATICS AND HUMIDITY

Rub a comb through your hair and use it to lift pieces of paper. It may help to tear the pieces of paper rather than cut them neatly. Repeat the exercise in your bathroom after you have had a long shower and the air in the bathroom is moist. Is it easier to get electrostatic effects in dry or moist air? Why would torn paper be more attractive to the comb than cut paper? Explain your observations.

Xerography

Most copy machines use an electrostatic process called **xerography**—a word coined from the Greek words *xeros* for dry and *graphos* for writing. The heart of the process is shown in simplified form in Figure 3.1.7.9.

A selenium-coated aluminum drum is sprayed with positive charge from points on a device called a corotron. Selenium is a substance with an interesting property—it is a **photoconductor**. That is, selenium is an insulator when in the dark and a conductor when exposed to light.

In the first stage of the xerography process, the conducting aluminum drum is **grounded** so that a negative charge is induced under the thin layer of uniformly positively charged selenium. In the second stage, the surface of the drum is exposed to the image of whatever is to be copied. Where the image is light, the selenium becomes conducting, and the positive charge is neutralized. In dark areas, the positive charge remains, and so the image has been transferred to the drum.

The third stage takes a dry black powder, called toner, and sprays it with a negative charge so that it will be attracted to the positive regions of the drum. Next, a blank piece of paper is given a greater positive charge than on the drum so that it will pull the toner from the drum. Finally, the paper and electrostatically held toner are passed through heated pressure rollers, which melt and permanently adhere the toner within the fibers of the paper.

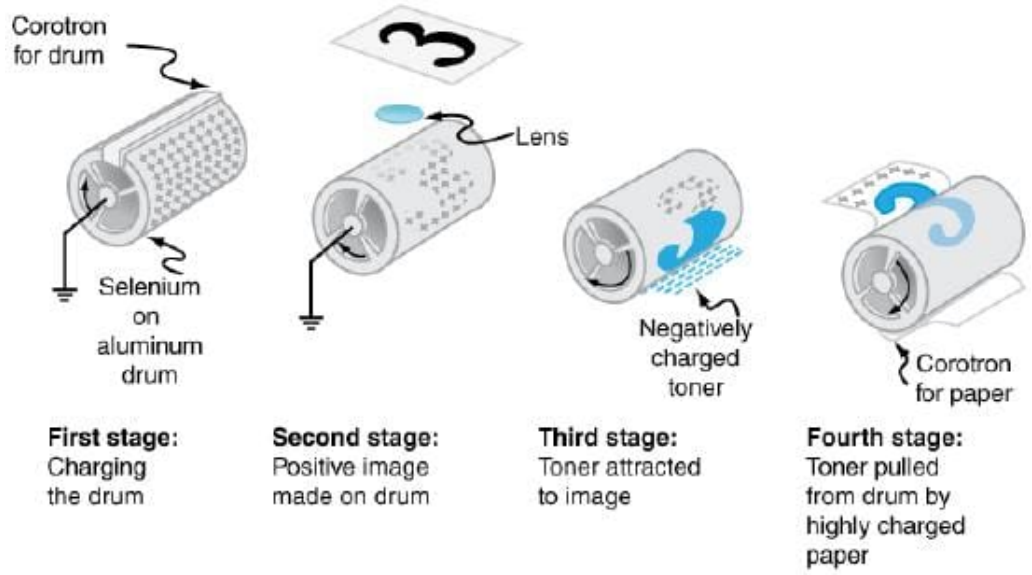


Figure 3.1.7.9: Xerography is a dry copying process based on electrostatics. The major steps in the process are the charging of the photoconducting drum, transfer of an image creating a positive charge duplicate, attraction of toner to the charged parts of the drum, and transfer of toner to the paper. Not shown are heat treatment of the paper and cleansing of the drum for the next copy.

Laser Printers

Laser printers use the xerographic process to make high-quality images on paper, employing a laser to produce an image on the photoconducting drum as shown in Figure 3.1.7.10. In its most common application, the laser printer receives output from a computer, and it can achieve high-quality output because of the precision with which laser light can be controlled. Many laser printers do significant information processing, such as making sophisticated letters or fonts, and may contain a computer more powerful than the one giving them the raw data to be printed.

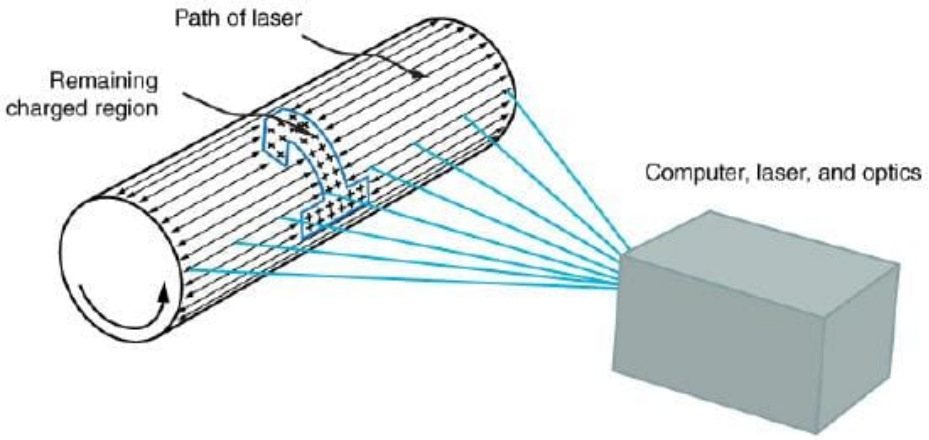


Figure 3.1.7.10: In a laser printer, a laser beam is scanned across a photoconducting drum, leaving a positive charge image. The other steps for charging the drum and transferring the image to paper are the same as in xerography. Laser light can be very precisely controlled, enabling laser printers to produce high-quality images.

Ink Jet Printers and Electrostatic Painting

The **ink jet printer**, commonly used to print computer-generated text and graphics, also employs electrostatics. A nozzle makes a fine spray of tiny ink droplets, which are then given an electrostatic charge. (See Figure 3.1.7.11.)

Once charged, the droplets can be directed, using pairs of charged plates, with great precision to form letters and images on paper. Ink jet printers can produce color images by using a black jet and three other jets with primary colors, usually cyan, magenta, and yellow, much as a color television produces color. (This is more difficult with xerography, requiring multiple drums

and toners.)

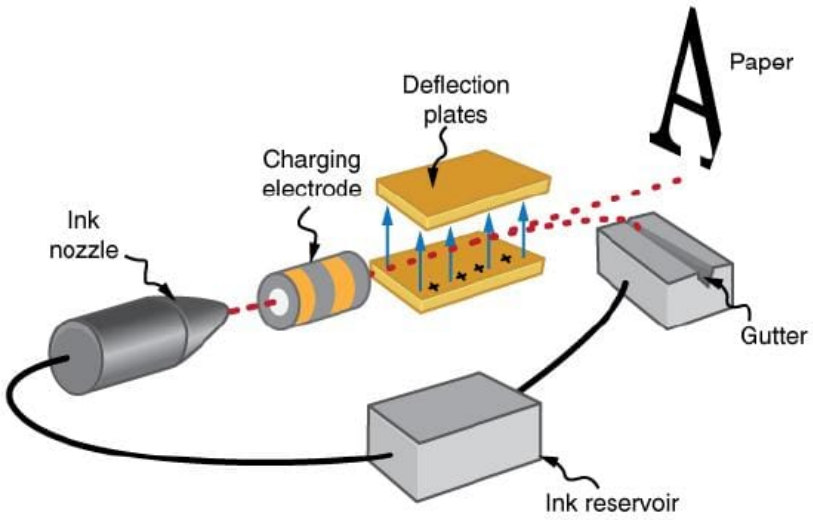


Figure 3.1.7.11: The nozzle of an ink-jet printer produces small ink droplets, which are sprayed with electrostatic charge. Various computer-driven devices are then used to direct the droplets to the correct positions on a page.

Electrostatic painting employs electrostatic charge to spray paint onto odd-shaped surfaces. Mutual repulsion of like charges causes the paint to fly away from its source. Surface tension forms drops, which are then attracted by unlike charges to the surface to be painted. Electrostatic painting can reach those hard-to-get at places, applying an even coat in a controlled manner. If the object is a conductor, the electric field is perpendicular to the surface, tending to bring the drops in perpendicularly. Corners and points on conductors will receive extra paint. Felt can similarly be applied.

Smoke Precipitators and Electrostatic Air Cleaning

Another important application of electrostatics is found in air cleaners, both large and small. The electrostatic part of the process places excess (usually positive) charge on smoke, dust, pollen, and other particles in the air and then passes the air through an oppositely charged grid that attracts and retains the charged particles. (See Figure 3.1.7.12.)

Large **electrostatic precipitators** are used industrially to remove over 99% of the particles from stack gas emissions associated with the burning of coal and oil. Home precipitators, often in conjunction with the home heating and air conditioning system, are very effective in removing polluting particles, irritants, and allergens.

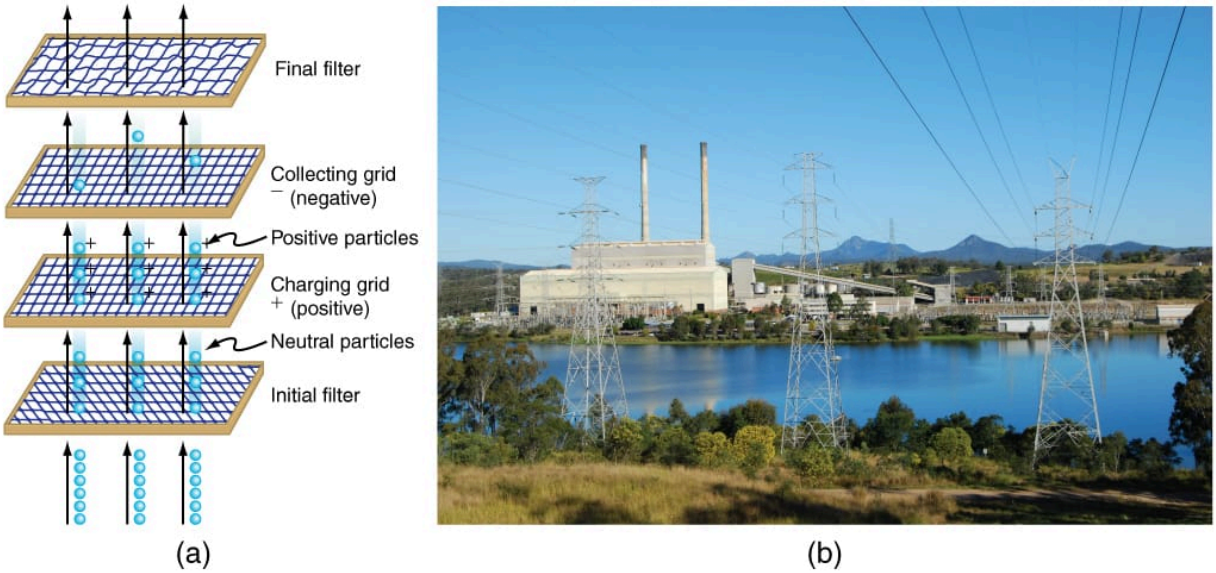


Figure 3.1.7.12: (a) Schematic of an electrostatic precipitator. Air is passed through grids of opposite charge. The first grid charges airborne particles, while the second attracts and collects them. (b) The dramatic effect of electrostatic precipitators is seen by the absence of smoke from this power plant. (credit: Cmdalgleish, Wikimedia Commons)

Section Summary

- The electrical forces around a conductor will cause free charges to move around inside the conductor until static equilibrium is reached.
- Any excess charge will collect along the surface of a conductor.
- Conductors with sharp corners or points will collect more charge at those points.
- A lightning rod is a conductor with sharply pointed ends that collect excess charge on the building caused by an electrical storm and allow it to dissipate back into the air.
- A Faraday cage acts like a shield around an object, preventing electric charge from penetrating inside.
- In addition to research using equipment such as a Van de Graaff generator, many practical applications of electrostatics exist, including photocopiers, laser printers, ink-jet printers and electrostatic air filters.

Glossary

conductor
an object with properties that allow charges to move about freely within it

free charge
an electrical charge (either positive or negative) which can move about separately from its base molecule

electrostatic equilibrium
an electrostatically balanced state in which all free electrical charges have stopped moving about

polarized
a state in which the positive and negative charges within an object have collected in separate locations

Faraday cage
a metal shield which prevents electric charge from penetrating its surface

capacitor
an arrangement of conductors designed to store charge using voltage difference

parallel-plate capacitor
an example of capacitor using arrangement of two parallel conducting plates placed near each other

Van de Graaff generator
: a machine that produces a large amount of excess charge, used for experiments with high voltage

electrostatics
: the study of electricity in electrostatic equilibrium

photoconductor
: a substance that is an insulator until it is exposed to light, when it becomes a conductor

xerography
: a dry copying process based on electrostatics

grounded
: connected to the ground with a conductor, so that charge flows freely to and from the Earth to the grounded object

laser printer
: uses a laser to create a photoconductive image on a drum, which attracts dry ink particles that are then rolled onto a sheet of paper to print a high-quality copy of the image

ink-jet printer
: small ink droplets sprayed with an electric charge are controlled by electrostatic plates to create images on paper

electrostatic precipitators
: filters that apply charges to particles in the air, then attract those charges to a filter, removing them from the airstream

This page titled 3.1.7: Conductors and Applications of Electrostatics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.7: Conductors and Applications of Electrostatics by Andrew Park is licensed CC BY 4.0.

3.1.8: Current

Learning Objectives

- Define electric current, ampere, and drift velocity
- Describe the direction of charge flow in conventional current.

Electric Current

Electric current is defined to be the rate at which charge flows. A large current, such as that used to start a truck engine, moves a large amount of charge in a small time, whereas a small current, such as that used to operate a hand-held calculator, moves a small amount of charge over a long period of time. In equation form, **electric current** I is defined to be

$$I = \frac{\Delta Q}{\Delta t},$$

where ΔQ is the amount of charge passing through a given area in time Δt. (As in previous chapters, initial time is often taken to be zero, in which case $\Delta t = t$.) (See Figure 3.1.8.1.) The SI unit for current is the **ampere** (A), named for the French physicist André-Marie Ampère (1775–1836). Since $I = \Delta Q/\Delta t$, we see that an ampere is one coulomb per second:

$$1 \text{ A} = 1 \text{ C/s}$$

Not only are fuses and circuit breakers rated in amperes (or amps), so are many electrical appliances.

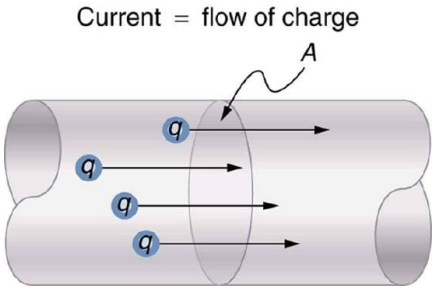


Figure 3.1.8.1: The rate of flow of charge is current. An ampere is the flow of one coulomb through an area in one second.

Example 3.1.8.1: Calculating Currents: Current in a Truck Battery and a Handheld Calculator

(a) What is the current involved when a truck battery sets in motion 720 C of charge in 4.00 s while starting an engine? (b) How long does it take 1.00 C of charge to flow through a handheld calculator if a 0.300-mA current is flowing?

Strategy

We can use the definition of current in the equation $I = \Delta Q/\Delta t$ to find the current in part (a), since charge and time are given. In part (b), we rearrange the definition of current and use the given values of charge and current to find the time required.

Solution for (a)

Entering the given values for charge and time into the definition of current gives

$$I = \frac{\Delta Q}{\Delta t} = \frac{720\text{C}}{4.00 \text{ s}} = 180\text{C/s} = 180 \text{ A}.$$

Discussion for (a)

This large value for current illustrates the fact that a large charge is moved in a small amount of time. The currents in these "starter motors" are fairly large because large frictional forces need to be overcome when setting something in motion.

Solution for (b)

Solving the relationship $I = \Delta Q/\Delta t$ for time Δt, and entering the known values for charge and current gives

$$\Delta t = \frac{\Delta Q}{I} = \frac{1.00\text{C}}{0.300 \times 10^{-3}\text{C/s}}$$
$$= 3.33 \times 10^{3}\ \text{s}.$$

Discussion for (b)

This time is slightly less than an hour. The small current used by the hand-held calculator takes a much longer time to move a smaller charge than the large current of the truck starter. So why can we operate our calculators only seconds after turning them on? It's because calculators require very little energy. Such small current and energy demands allow handheld calculators to operate from solar cells or to get many hours of use out of small batteries. Remember, calculators do not have moving parts in the same way that a truck engine has with cylinders and pistons, so the technology requires smaller currents.

Figure 3.1.8.2 shows a simple circuit and the standard schematic representation of a battery, conducting path, and load (a resistor). Schematics are very useful in visualizing the main features of a circuit. A single schematic can represent a wide variety of situations. The schematic in Figure 3.1.8.2(b), for example, can represent anything from a truck battery connected to a headlight lighting the street in front of the truck to a small battery connected to a penlight lighting a keyhole in a door. Such schematics are useful because the analysis is the same for a wide variety of situations. We need to understand a few schematics to apply the concepts and analysis to many more situations.

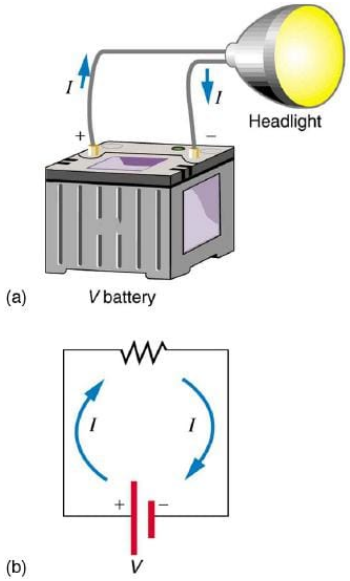


Figure 3.1.8.2: (a) A simple electric circuit. A closed path for current to flow through is supplied by conducting wires connecting a load to the terminals of a battery. (b) In this schematic, the battery is represented by the two parallel red lines, conducting wires are shown as straight lines, and the zigzag represents the load. The schematic represents a wide variety of similar circuits.

Note that the direction of current flow in Figure 3.1.8.2 is from positive to negative. *The direction of conventional current is the direction that positive charge would flow*. Depending on the situation, positive charges, negative charges, or both may move. In metal wires, for example, current is carried by electrons—that is, negative charges move. In ionic solutions, such as salt water, both positive and negative charges move. This is also true in nerve cells. A Van de Graaff generator used for nuclear research can produce a current of pure positive charges, such as protons. Figure 3.1.8.3 illustrates the movement of charged particles that compose a current. The fact that conventional current is taken to be in the direction that positive charge would flow can be traced back to American politician and scientist Benjamin Franklin in the 1700s. He named the type of charge associated with electrons

negative, long before they were known to carry current in so many situations. Franklin, in fact, was totally unaware of the small-scale structure of electricity.

It is important to realize that there is an electric field in conductors responsible for producing the current, as illustrated in Figure 3.1.8.3. Unlike static electricity, where a conductor in equilibrium cannot have an electric field in it, conductors carrying a current have an electric field and are not in static equilibrium. An electric field is needed to supply energy to move the charges.

MAKING CONNECTIONS: TAKE-HOME INVESTIGATION—ELECTRIC CURRENT ILLUSTRATION

Find a straw and little peas that can move freely in the straw. Place the straw flat on a table and fill the straw with peas. When you pop one pea in at one end, a different pea should pop out the other end. This demonstration is an analogy for an electric current. Identify what compares to the electrons and what compares to the supply of energy. What other analogies can you find for an electric current?

Note that the flow of peas is based on the peas physically bumping into each other; electrons flow due to mutually repulsive electrostatic forces.

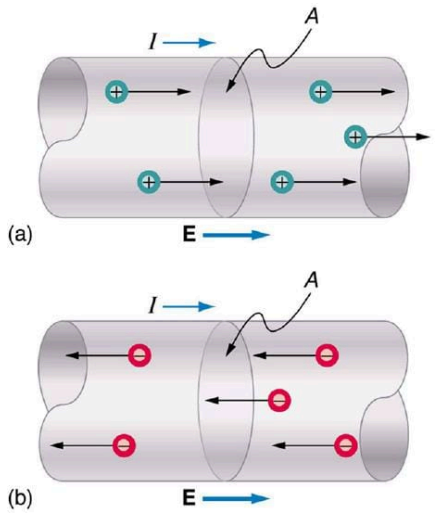


Figure 3.1.8.3: Current I is the rate at which charge moves through an area A, such as the cross-section of a wire. Conventional current is defined to move in the direction of the electric field. (a) Positive charges move in the direction of the electric field and the same direction as conventional current. (b) Negative charges move in the direction opposite to the electric field. Conventional current is in the direction opposite to the movement of negative charge. The flow of electrons is sometimes referred to as electronic flow.

Example 3.1.8.2: Calculating the Number of Electrons that Move through a Calculator

If the 0.300-mA current through the calculator mentioned in the Example 3.1.8.1 is carried by electrons, how many electrons per second pass through it?

Strategy

The current calculated in the previous example was defined for the flow of positive charge. For electrons, the magnitude is the same, but the sign is opposite, $I_{\text{electrons}} = -0.300 \times 10^{-3}\text{C/s}$. Since each electron (e^-) has a charge of -1.60×10^{-19} C, we can convert the current in coulombs per second to electrons per second.

Solution

Starting with the definition of current, we have

$$I_{\text{electrons}} = \frac{\Delta Q_{\text{electrons}}}{\Delta t} = \frac{-0.300 \times 10^{-3}\text{C}}{\text{s}}.$$

We divide this by the charge per electron, so that

$$\begin{aligned}\frac{e^-}{\text{s}} &= \frac{-0.300 \times 10^{-3}\text{C}}{\text{s}} \times \frac{1e^-}{-1.60 \times 10^{-19}\text{C}} \\ &= 1.88 \times 10^{15}\frac{e^-}{\text{s}}.\end{aligned}$$

Discussion

There are so many charged particles moving, even in small currents, that individual charges are not noticed, just as individual water molecules are not noticed in water flow. Even more amazing is that they do not always keep moving forward like soldiers in a parade. Rather they are like a crowd of people with movement in different directions but a general trend to move forward. There are lots of collisions with atoms in the metal wire and, of course, with other electrons.

Section Summary

- Electric current I is the rate at which charge flows, given by

$$I = \frac{\Delta Q}{\Delta t},$$

  where ΔQ is the amount of charge passing through an area in time Δt.
- The direction of conventional current is taken as the direction in which positive charge moves.
- The SI unit for current is the ampere (A), where $1\ \text{A} = 1\ \text{C/s}$.
- Current is the flow of free charges, such as electrons and ions.

Glossary

electric current
: the rate at which charge flows, $I = \Delta Q/\Delta t$

ampere
: (amp) the SI unit for current; 1 A = 1 C/s

This page titled 3.1.8: Current is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.8: Current by Andrew Park is licensed CC BY 4.0.

3.1.9: Ohm's Law- Resistance and Simple Circuits

Learning Objectives

- Explain the origin of Ohm's law.
- Calculate voltages, currents, or resistances with Ohm's law.
- Explain what an ohmic material is.
- Describe a simple circuit.

What drives current? We can think of various devices—such as batteries, generators, wall outlets, and so on—which are necessary to maintain a current. All such devices create a potential difference and are loosely referred to as voltage sources. When a voltage source is connected to a conductor, it applies a potential difference V that creates an electric field. The electric field in turn exerts force on charges, causing current.

Ohm's Law

The current that flows through most substances is directly proportional to the voltage V applied to it. The German physicist Georg Simon Ohm (1787–1854) was the first to demonstrate experimentally that the current in a metal wire is *directly proportional to the voltage applied*:

$$I \propto V.$$

This important relationship is known as **Ohm's law**. It can be viewed as a cause-and-effect relationship, with voltage the cause and current the effect. This is an empirical law like that for friction—an experimentally observed phenomenon. Such a linear relationship doesn't always occur.

Resistance and Simple Circuits

If voltage drives current, what impedes it? The electric property that impedes current (crudely similar to friction and air resistance) is called **resistance** R. Collisions of moving charges with atoms and molecules in a substance transfer energy to the substance and limit current. Resistance is defined as inversely proportional to current, or

$$I \propto \frac{1}{R}.$$

Thus, for example, current is cut in half if resistance doubles. Combining the relationships of current to voltage and current to resistance gives

$$I = \frac{V}{R}.$$

This relationship is also called Ohm's law. Ohm's law in this form really defines resistance for certain materials. Ohm's law (like Hooke's law) is not universally valid. The many substances for which Ohm's law holds are called **ohmic**. These include good conductors like copper and aluminum, and some poor conductors under certain circumstances. Ohmic materials have a resistance R that is independent of voltage V and current I. An object that has simple resistance is called a *resistor*, even if its resistance is small. The unit for resistance is an **ohm** and is given the symbol Ω (upper case Greek omega). Rearranging $I = V/R$ gives $R = V/I$, and so the units of resistance are 1 ohm = 1 volt per ampere:

$$1\ \Omega = 1\frac{V}{A}.$$

Figure 3.1.9.1 shows the schematic for a simple circuit. A **simple circuit** has a single voltage source and a single resistor. The wires connecting the voltage source to the resistor can be assumed to have negligible resistance, or their resistance can be included in R.

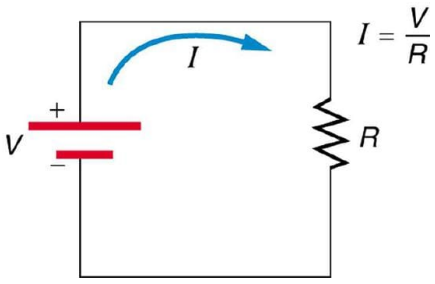


Figure 3.1.9.1: A simple electric circuit in which a closed path for current to flow is supplied by conductors (usually metal wires) connecting a load to the terminals of a battery, represented by the red parallel lines. The zigzag symbol represents the single resistor and includes any resistance in the connections to the voltage source.

Example 3.1.9.1: Calculating Resistance: An Automobile Headlight

What is the resistance of an automobile headlight through which 2.50 A flows when 12.0 V is applied to it?

Strategy

We can rearrange Ohm's law as stated by $I = V/R$ and use it to find the resistance.

Solution

Rearranging $I = V/R$ and substituting known values gives

$$R = \frac{V}{I} = \frac{12.0\ \text{V}}{2.50\ \text{A}} = 4.80\Omega.$$

Discussion

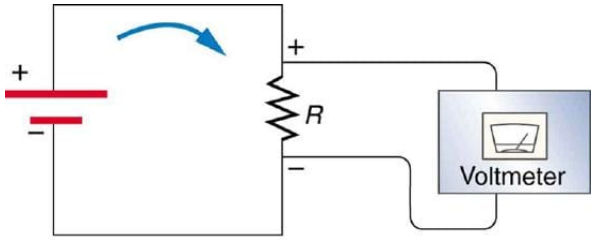
This is a relatively small resistance, but it is larger than the cold resistance of the headlight. Resistance usually increases with temperature, and so the bulb has a lower resistance when it is first switched on and will draw considerably more current during its brief warm-up period.

Resistances range over many orders of magnitude. Some ceramic insulators, such as those used to support power lines, have resistances of $10^{12}\ \Omega$ or more. A dry person may have a hand-to-foot resistance of $10^{5}\ \Omega$, whereas the resistance of the human heart is about $10^{3}\ \Omega$. A meter-long piece of large-diameter copper wire may have a resistance of $10^{-5}\ \Omega$, and superconductors have no resistance at all (they are non-ohmic). Resistance is related to the shape of an object and the material of which it is composed.

Additional insight is gained by solving $I = V/R$ for V yielding

$$V = IR.$$

This expression for V can be interpreted as the *voltage drop across a resistor produced by the flow of current* I. The phrase IR *drop* is often used for this voltage. For instance, the headlight in Example 3.1.9.1 has an IR drop of 12.0 V. If voltage is measured at various points in a circuit, it will be seen to increase at the voltage source and decrease at the resistor. Voltage is similar to fluid pressure. The voltage source is like a pump, creating a pressure difference, causing current—the flow of charge. The resistor is like a pipe that reduces pressure and limits flow because of its resistance. Conservation of energy has important consequences here. The voltage source supplies energy (causing an electric field and a current), and the resistor converts it to another form (such as thermal energy). In a simple circuit (one with a single simple resistor), the voltage supplied by the source equals the voltage drop across the resistor, since $\text{PE} = q\Delta V$, and the same q flows through each. Thus the energy supplied by the voltage source and the energy converted by the resistor are equal. (See Figure 3.1.9.2.)



$V = IR = 18$ V

Figure 3.1.9.2: The voltage drop across a resistor in a simple circuit equals the voltage output of the battery.

MAKING CONNECTIONS: CONSERVATION OF ENERGY

In a simple electrical circuit, the sole resistor converts energy supplied by the source into another form. Conservation of energy is evidenced here by the fact that all of the energy supplied by the source is converted to another form by the resistor alone. We will find that conservation of energy has other important applications in circuits and is a powerful tool in circuit analysis.

Section Summary

- A simple circuit *is* one in which there is a single voltage source and a single resistance.
- One statement of Ohm's law gives the relationship between current I, voltage V, and resistance R in a simple circuit to be $I = \frac{V}{R}$
- Resistance has units of ohms (Ω), related to volts and amperes by $1\ \Omega = 1\ \mathrm{V/A}$.
- There is a voltage or IR drop across a resistor, caused by the current flowing through it, given by $V = IR$.

Glossary

Ohm's law
: an empirical relation stating that the current *I* is proportional to the potential difference *V*. It is often written as *I* = *V*/*R*, where *R* is the resistance

resistance
: the electric property that impedes current; for ohmic materials, it is the ratio of voltage to current, *R* = *V*/*I*

ohm
: the unit of resistance, given by $1\ \Omega = 1\ \mathrm{V/A}$

ohmic
: a type of a material for which Ohm's law is valid

simple circuit
: a circuit with a single voltage source and a single resistor

This page titled 3.1.9: Ohm's Law- Resistance and Simple Circuits is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.9: Ohm's Law- Resistance and Simple Circuits by Andrew Park is licensed CC BY 4.0.

3.1.10: Electric Power and Energy

Learning Objectives

- Calculate the power dissipated by a resistor and power supplied by a power supply.
- Calculate the cost of electricity under various circumstances.

Power in Electric Circuits

Power is associated by many people with electricity. Knowing that power is the rate of energy use or energy conversion, what is the expression for **electric power**? Power transmission lines might come to mind. We also think of lightbulbs in terms of their power ratings in watts. Let us compare a 25-W bulb with a 60-W bulb. (See Figure 3.1.10.1(a).) Since both operate on the same voltage, the 60-W bulb must draw more current to have a greater power rating. Thus the 60-W bulb's resistance must be lower than that of a 25-W bulb. If we increase voltage, we also increase power. For example, when a 25-W bulb that is designed to operate on 120 V is connected to 240 V, it briefly glows very brightly and then burns out. Precisely how are voltage, current, and resistance related to electric power?

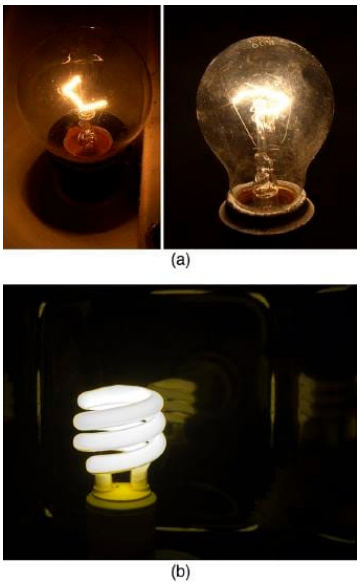


Figure 3.1.10.1: (a) Which of these lightbulbs, the 25-W bulb (upper left) or the 60-W bulb (upper right), has the higher resistance? Which draws more current? Which uses the most energy? Can you tell from the color that the 25-W filament is cooler? Is the brighter bulb a different color and if so why? (credits: Dickbauch, Wikimedia Commons; Greg Westfall, Flickr) (b) This compact fluorescent light (CFL) puts out the same intensity of light as the 60-W bulb, but at 1/4 to 1/10 the input power. (credit: dbgg1979, Flickr)

Electric energy depends on both the voltage involved and the charge moved. This is expressed most simply as $\mathrm{PE} = qV$, where q is the charge moved and V is the voltage (or more precisely, the potential difference the charge moves through). Power is the rate at which energy is moved, and so electric power is

$$P = \frac{PE}{t} = \frac{qV}{t}.$$

Recognizing that current is $I = q/t$ (note that $\Delta t = t$ here), the expression for power becomes

$$P = IV.$$

Electric power (P) is simply the product of current times voltage. Power has familiar units of watts. Since the SI unit for potential energy (PE) is the joule, power has units of joules per second, or watts. Thus, $1\ \mathrm{A} \cdot \mathrm{V} = 1\ \mathrm{W}$. For example, cars often have one or more auxiliary power outlets with which you can charge a cell phone or other electronic devices. These outlets may be rated at 20

A, so that the circuit can deliver a maximum power $P = IV = (20\ \text{A})(12\ \text{V}) = 240\ \text{W}$. In some applications, electric power may be expressed as volt-amperes or even kilovolt-amperes ($1\text{kA} \cdot \text{V} = 1\ \text{kW}$).

To see the relationship of power to resistance, we combine Ohm's law with $P = IV$. Substituting $I = V/R$ gives $P = (V/R)V = V^2/R$. Similarly, substituting $V = IR$ gives $P = I(IR) = I^2R$. Three expressions for electric power are listed together here for convenience:

$$P = IV$$

$$P = \frac{V^2}{R}$$

$$P = I^2R.$$

Note that the first equation is always valid, whereas the other two can be used only for resistors. In a simple circuit, with one voltage source and a single resistor, the power supplied by the voltage source and that dissipated by the resistor are identical. (In more complicated circuits, P can be the power dissipated by a single device and not the total power in the circuit.)

Different insights can be gained from the three different expressions for electric power. For example, $P = V^2/R$ implies that the lower the resistance connected to a given voltage source, the greater the power delivered. Furthermore, since voltage is squared in $P = V^2/R$, the effect of applying a higher voltage is perhaps greater than expected. Thus, when the voltage is doubled to a 25-W bulb, its power nearly quadruples to about 100 W, burning it out. If the bulb's resistance remained constant, its power would be exactly 100 W, but at the higher temperature its resistance is higher, too.

Example 3.1.10.1: Calculating Power Dissipation and Current

Consider the example given in "Ohm's Law: Resistance and Simple Circuits." Then find the power dissipated by the car headlight.

Strategy

For the headlight, we know voltage and current, so we can use $P = IV$ to find the power.

Solution

Entering the known values of current and voltage for the hot headlight, we obtain

$$P = IV = (2.50\ \text{A})(12.0\ \text{V}) = 30.0\ \text{W}.$$

Discussion

The 30 W dissipated by the hot headlight is typical.

The Cost of Electricity

The more electric appliances you use and the longer they are left on, the higher your electric bill. This familiar fact is based on the relationship between energy and power. You pay for the energy used. Since $P = E/t$, we see that

$$E = Pt$$

is the energy used by a device using power P for a time interval t. For example, the more lightbulbs burning, the greater P used; the longer they are on, the greater t is. The energy unit on electric bills is the kilowatt-hour ($\text{kW} \cdot \text{h}$), consistent with the relationship $E = Pt$. It is easy to estimate the cost of operating electric appliances if you have some idea of their power consumption rate in watts or kilowatts, the time they are on in hours, and the cost per kilowatt-hour for your electric utility. Kilowatt-hours, like all other specialized energy units such as food calories, can be converted to joules. You can prove to yourself that $1\ \text{kW} \cdot \text{h} = 3.6 \times 10^6\ \text{J}$.

The electrical energy (E) used can be reduced either by reducing the time of use or by reducing the power consumption of that

appliance or fixture. This will not only reduce the cost, but it will also result in a reduced impact on the environment. Improvements to lighting are some of the fastest ways to reduce the electrical energy used in a home or business. About 20% of a home's use of energy goes to lighting, while the number for commercial establishments is closer to 40%. Fluorescent lights are about four times more efficient than incandescent lights—this is true for both the long tubes and the compact fluorescent lights (CFL). (See Figure 3.1.10.1(b).) Thus, a 60-W incandescent bulb can be replaced by a 15-W CFL, which has the same brightness and color. CFLs have a bent tube inside a globe or a spiral-shaped tube, all connected to a standard screw-in base that fits standard incandescent light sockets. (Original problems with color, flicker, shape, and high initial investment for CFLs have been addressed in recent years.) The heat transfer from these CFLs is less, and they last up to 10 times longer. The significance of an investment in such bulbs is addressed in the next example. New white LED lights (which are clusters of small LED bulbs) are even more efficient (twice that of CFLs) and last 5 times longer than CFLs. However, their cost is still high.

MAKING CONNECTIONS: ENERGY, POWER, AND TIME

The relationship $E = Pt$ is one that you will find useful in many different contexts. The energy your body uses in exercise is related to the power level and duration of your activity, for example. The amount of heating by a power source is related to the power level and time it is applied. Even the radiation dose of an X-ray image is related to the power and time of exposure.

Example 3.1.10.2: Calculating the Cost Effectiveness of Compact Fluorescent Lights (CFL)

If the cost of electricity in your area is 12 cents per kWh, what is the total cost (capital plus operation) of using a 60-W incandescent bulb for 1000 hours (the lifetime of that bulb) if the bulb cost 25 cents? (b) If we replace this bulb with a compact fluorescent light that provides the same light output, but at one-quarter the wattage, and which costs $1.50 but lasts 10 times longer (10,000 hours), what will that total cost be?

Strategy

To find the operating cost, we first find the energy used in kilowatt-hours and then multiply by the cost per kilowatt-hour.

Solution for (a)

The energy used in kilowatt-hours is found by entering the power and time into the expression for energy:

$$E = Pt = (60\ \text{W})(1000\ \text{h}) = 60,000\ \text{W} \cdot \text{h}.$$

In kilowatt-hours, this is

$$E = 60.0\ \text{kW} \cdot \text{h}.$$

Now the electricity cost is

$$\text{cost} = (60.0\ \text{kW} \cdot \text{h})(\$0.12/\text{kW} \cdot \text{h}) = \$7.20.$$

The total cost will be $7.20 for 1000 hours (about one-half year at 5 hours per day).

Solution for (b)

Since the CFL uses only 15 W and not 60 W, the electricity cost will be $7.20/4 = $1.80. The CFL will last 10 times longer than the incandescent, so that the investment cost will be 1/10 of the bulb cost for that time period of use, or 0.1($1.50) = $0.15. Therefore, the total cost will be $1.95 for 1000 hours.

Discussion

Therefore, it is much cheaper to use the CFLs, even though the initial investment is higher. The increased cost of labor that a business must include for replacing the incandescent bulbs more often has not been figured in here.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—ELECTRICAL ENERGY USE INVENTORY

1) Make a list of the power ratings on a range of appliances in your home or room. Explain why something like a toaster has a higher rating than a digital clock. Estimate the energy consumed by these appliances in an average day (by estimating their time of use). Some appliances might only state the operating current. If the household voltage is 120 V, then use $P = IV$. 2) Check out the total wattage used in the rest rooms of your school's floor or building. (You might need to assume the long fluorescent lights in use are rated at 32 W.) Suppose that the building was closed all weekend and that these lights were left on from 6 p.m. Friday until 8 a.m. Monday. What would this oversight cost? How about for an entire year of weekends?

Section Summary

- Electric power P is the rate (in watts) that energy is supplied by a source or dissipated by a device.
- Three expressions for electrical power are

$$P = IV,$$

$$P = \frac{V^2}{R},$$

and

$$P = I^2R.$$

- The energy used by a device with a power P over a time t is $E = Pt$.

Glossary

electric power
: the rate at which electrical energy is supplied by a source or dissipated by a device; it is the product of current times voltage

This page titled 3.1.10: Electric Power and Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.10: Electric Power and Energy by Andrew Park is licensed CC BY 4.0.

3.1.11: Resistors in Series and Parallel

Learning Objectives

- Draw a circuit with resistors in parallel and in series.
- Calculate the voltage drop of a current across a resistor using Ohm's law.
- Contrast the way total resistance is calculated for resistors in series and in parallel.
- Explain why total resistance of a parallel circuit is less than the smallest resistance of any of the resistors in that circuit.
- Calculate total resistance of a circuit that contains a mixture of resistors connected in series and in parallel.

Most circuits have more than one component, called a **resistor** that limits the flow of charge in the circuit. A measure of this limit on charge flow is called **resistance**. The simplest combinations of resistors are the series and parallel connections illustrated in Figure 3.1.11.1. The total resistance of a combination of resistors depends on both their individual values and how they are connected.

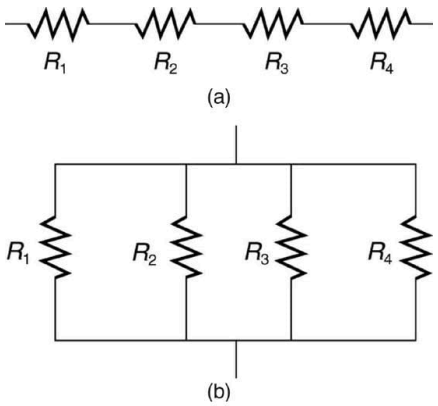


Figure 3.1.11.1: (a) A series connection of resistors. (b) A parallel connection of resistors.

Resistors in Series

When are resistors in **series**? Resistors are in series whenever the flow of charge, called the **current**, must flow through devices sequentially. For example, if current flows through a person holding a screwdriver and into the Earth, then R_1 in Figure 3.1.11.1 (a) could be the resistance of the screwdriver's shaft, R_2 the resistance of its handle, R_3 the person's body resistance, and R_4 the resistance of her shoes.

Figure 3.1.11.2 shows resistors in series connected to a **voltage** source. It seems reasonable that the total resistance is the sum of the individual resistances, considering that the current has to pass through each resistor in sequence. (This fact would be an advantage to a person wishing to avoid an electrical shock, who could reduce the current by wearing high-resistance rubber-soled shoes. It could be a disadvantage if one of the resistances were a faulty high-resistance cord to an appliance that would reduce the operating current.)

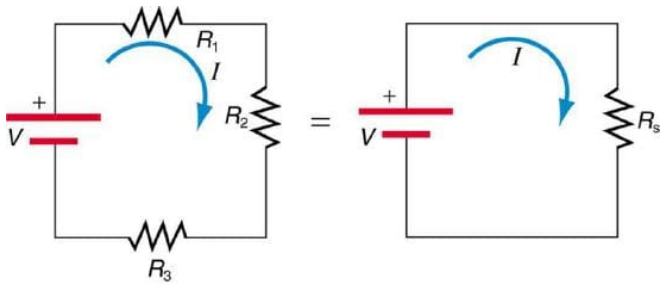


Figure 3.1.11.2: Three resistors connected in series to a battery (left) and the equivalent single or series resistance (right).

To verify that resistances in series do indeed add, let us consider the loss of electrical power, called a **voltage drop**, in each resistor in Figure 3.1.11.2.

According to **Ohm's law**, the voltage drop, V, across a resistor when a current flows through it is calculated using the equation $V = IR$, where I equals the current in amps (A) and R is the resistance in ohms (Ω). Another way to think of this is that V is the voltage necessary to make a current I flow through a resistance R.

So the voltage drop across R_1 is $V_1 = IR_1$, that across R_2 is $V_2 = IR_2$, and that across R_3 is $V_3 = IR_3$. The sum of these voltages equals the voltage output of the source; that is,

$$V = V_1 + V_2 + V_3.$$

This equation is based on the conservation of energy and conservation of charge. Electrical potential energy can be described by the equation $PE = qV$, where q is the electric charge and V is the voltage. Thus the energy supplied by the source is qV, while that dissipated by the resistors is

$$qV_1 + qV_2 + qV_3.$$

CONNECTIONS: CONSERVATION LAWS

The derivations of the expressions for series and parallel resistance are based on the laws of conservation of energy and conservation of charge, which state that total charge and total energy are constant in any process. These two laws are directly involved in all electrical phenomena and will be invoked repeatedly to explain both specific effects and the general behavior of electricity.

These energies must be equal, because there is no other source and no other destination for energy in the circuit. Thus, $qV = qV_1 + qV_2 + qV_3$. The charge q cancels, yielding $V = V_1 + V_2 + V_3$, as stated. (Note that the same amount of charge passes through the battery and each resistor in a given amount of time, since there is no capacitance to store charge, there is no place for charge to leak, and charge is conserved.)

Now substituting the values for the individual voltages gives

$$V = IR_1 + IR_2 + IR_3 = I\left(R_1 + R_2 + R_3\right).$$

Note that for the equivalent single series resistance R_s, we have

$$V = IR_\text{s}.$$

This implies that the total or equivalent series resistance R_s of three resistors is $R_\text{s} = R_1 + R_2 + R_3$.

This logic is valid in general for any number of resistors in series; thus, the total resistance R_s of a series connection is

$$R_\text{s} = R_1 + R_2 + R_3 + \ldots$$

as proposed. Since all of the current must pass through each resistor, it experiences the resistance of each, and resistances in series simply add up.

Example 3.1.11.1: Calculating Resistance, Current, Voltage Drop, and Power Dissipation: Analysis of a Series Circuit

Suppose the voltage output of the battery in Figure 3.1.11.2 is $12.0\ \text{V}$, and the resistances are $R_1 = 1.00\ \Omega$, $R_2 = 6.00\ \Omega$, and $R_3 = 13.0\ \Omega$. (a) What is the total resistance? (b) Find the current. (c) Calculate the voltage drop in each resistor, and show these add to equal the voltage output of the source. (d) Calculate the power dissipated by each resistor. (e) Find the power output of the source, and show that it equals the total power dissipated by the resistors.

Strategy and Solution for (a)

The total resistance is simply the sum of the individual resistances, as given by this equation:

$$\begin{aligned} R_{\text{s}} &= R_1 + R_2 + R_3 \\ &= 1.00\ \Omega + 6.00\ \Omega + 13.0\ \Omega \\ &= 20.0\ \Omega. \end{aligned}$$

Strategy and Solution for (b)

The current is found using Ohm's law, $V = IR$. Entering the value of the applied voltage and the total resistance yields the current for the circuit:

$$I = \frac{V}{R_{\text{s}}} = \frac{12.0\ \text{V}}{20.0\ \Omega} = 0.600\ \text{A}.$$

Strategy and Solution for (c)

The voltage—or IR drop—in a resistor is given by Ohm's law. Entering the current and the value of the first resistance yields

$$V_1 = IR_1 = (0.600\ \text{A})(1.0\ \Omega) = 0.600\ \text{V}.$$

Similarly,

$$V_2 = IR_2 = (0.600\ \text{A})(6.0\ \Omega) = 3.60\ \text{V}$$

and

$$V_3 = IR_3 = (0.600\ \text{A})(13.0\ \Omega) = 7.80\ \text{V}.$$

Discussion for (c)

The three IR drops add to $12.0\ \text{V}$, as predicted:

$$V_1 + V_2 + V_3 = (0.600 + 3.60 + 7.80)\text{V} = 12.0\ \text{V}.$$

Strategy and Solution for (d)

The easiest way to calculate power in watts (W) dissipated by a resistor in a DC circuit is to use **Joule's law**, $P = IV$, where P is electric power. In this case, each resistor has the same full current flowing through it. By substituting Ohm's law $V = IR$ into Joule's law, we get the power dissipated by the first resistor as

$$P_1 = I^2R_1 = (0.600\ \text{A})^2(1.00\ \Omega) = 0.360\ \text{W}.$$

Similarly,

$$P_2 = I^2R_2 = (0.600\ \text{A})^2(6.00\ \Omega) = 2.16\ \text{W}$$

and

$$P_3 = I^2R_3 = (0.600\ \text{A})^2(13.0\ \Omega) = 4.68\ \text{W}.$$

Discussion for (d)

Power can also be calculated using either $P = IV$ or $P = \frac{V^2}{R}$, where V is the voltage drop across the resistor (not the full voltage of the source). The same values will be obtained.

Strategy and Solution for (e)

The easiest way to calculate power output of the source is to use $P = IV$, where V is the source voltage. This gives

$$P = (0.600\ \text{A})(12.0\ \text{V}) = 7.20\ \text{W}.$$

Discussion for (e)

Note, coincidentally, that the total power dissipated by the resistors is also 7.20 W, the same as the power put out by the source. That is,

$$P_1 + P_2 + P_3 = (0.360 + 2.16 + 4.68)\text{W} = 7.20\ \text{W}.$$

Power is energy per unit time (watts), and so conservation of energy requires the power output of the source to be equal to the total power dissipated by the resistors.

MAJOR FEATURES OF RESISTORS IN SERIES

1. Series resistances add: $R_\text{s} = R_1 + R_2 + R_3 + \ldots$
2. The same current flows through each resistor in series.
3. Individual resistors in series do not get the total source voltage, but divide it.

Resistors in Parallel

Figure 3.1.11.3 shows resistors in **parallel**, wired to a voltage source. Resistors are in parallel when each resistor is connected directly to the voltage source by connecting wires having negligible resistance. Each resistor thus has the full voltage of the source applied to it.

Each resistor draws the same current it would if it alone were connected to the voltage source (provided the voltage source is not overloaded). For example, an automobile's headlights, radio, and so on, are wired in parallel, so that they utilize the full voltage of the source and can operate completely independently. The same is true in your house, or any building. (See Figure 3.1.11.3(b).)

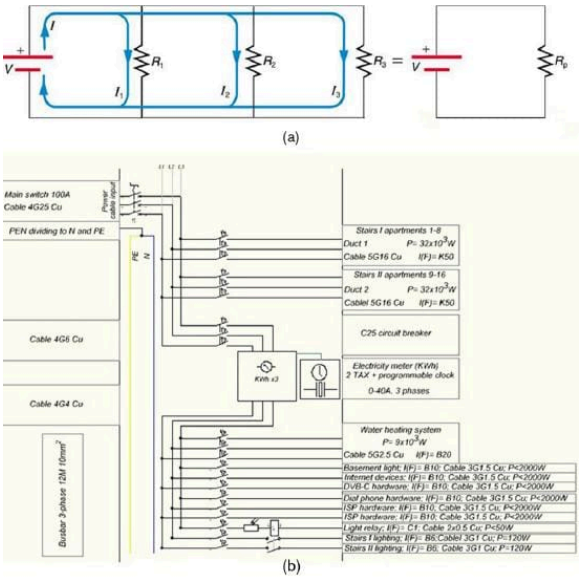


Figure 3.1.11.3: (a) Three resistors connected in parallel to a battery and the equivalent single or parallel resistance. (b) Electrical power setup in a house. (credit: Dmitry G, Wikimedia Commons)

To find an expression for the equivalent parallel resistance R_p, let us consider the currents that flow and how they are related to resistance. Since each resistor in the circuit has the full voltage, the currents flowing through the individual resistors are $I_1 = \frac{V}{R_1}$, $I_2 = \frac{V}{R_2}$, and $I_3 = \frac{V}{R_3}$. Conservation of charge implies that the total current I produced by the source is the sum of these currents:

$$I = I_1 + I_2 + I_3.$$

Substituting the expressions for the individual currents gives

$$I = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3} = V\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right).$$

Note that Ohm's law for the equivalent single resistance gives

$$I = \frac{V}{R_\text{p}} = V\left(\frac{1}{R_\text{p}}\right).$$

The terms inside the parentheses in the last two equations must be equal. Generalizing to any number of resistors, the total resistance R_p of a parallel connection is related to the individual resistances by

$$\frac{1}{R_\text{p}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots$$

This relationship results in a total resistance R_p that is less than the smallest of the individual resistances. (This is seen in the next example.) When resistors are connected in parallel, more current flows from the source than would flow for any of them individually, and so the total resistance is lower.

Example 3.1.11.2: Calculating Resistance, Current, Power Dissipation, and Power Output: Analysis of a Parallel Circuit

Let the voltage output of the battery and resistances in the parallel connection in Figure 3.1.11.3 be the same as the previously considered series connection: $V = 12.0\ \text{V}$, $R_1 = 1.00\ \Omega$, $R_2 = 6.00\ \Omega$, and $R_3 = 13.0\ \Omega$. (a) What is the total resistance? (b) Find the total current. (c) Calculate the currents in each resistor, and show these add to equal the total current output of the source. (d) Calculate the power dissipated by each resistor. (e) Find the power output of the source, and show that it equals the total power dissipated by the resistors.

Strategy and Solution for (a)

The total resistance for a parallel combination of resistors is found using the equation below. Entering known values gives

$$\frac{1}{R_\text{p}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} = \frac{1}{1.00\ \Omega} + \frac{1}{6.00\ \Omega} + \frac{1}{13.0\ \Omega}.$$

Thus,

$$\frac{1}{R_\text{p}} = \frac{1.00}{\Omega} + \frac{0.1667}{\Omega} + \frac{0.07692}{\Omega} = \frac{1.2436}{\Omega}.$$

(Note that in these calculations, each intermediate answer is shown with an extra digit.)

We must invert this to find the total resistance R_p. This yields

$$R_\text{p} = \frac{1}{1.2436}\ \Omega = 0.8041\ \Omega.$$

The total resistance with the correct number of significant digits is $R_\text{P} = 0.804\ \Omega$

Discussion for (a)

R_p is, as predicted, less than the smallest individual resistance.

Strategy and Solution for (b)

The total current can be found from Ohm's law, substituting R_p for the total resistance. This gives

$$I = \frac{V}{R_\text{p}} = \frac{12.0\ \text{V}}{0.8041\ \Omega} = 14.92\ \text{A}.$$

Discussion for (b)

Current I for each device is much larger than for the same devices connected in series (see the previous example). A circuit with parallel connections has a smaller total resistance than the resistors connected in series.

Strategy and Solution for (c)

The individual currents are easily calculated from Ohm's law, since each resistor gets the full voltage. Thus,

$$I_1 = \frac{V}{R_1} = \frac{12.0\ \text{V}}{1.00\ \Omega} = 12.0\ \text{A}.$$

Similarly,

$$I_2 = \frac{V}{R_2} = \frac{12.0\ \text{V}}{6.00\ \Omega} = 2.00\ \text{A}$$

and

$$I_3 = \frac{V}{R_3} = \frac{12.0\ \text{V}}{13.0\ \Omega} = 0.92\ \text{A}$$

Discussion for (c)

The total current is the sum of the individual currents:

$$I_1 + I_2 + I_3 = 14.92\ \text{A}.$$

This is consistent with conservation of charge.

Strategy and Solution for (d)

The power dissipated by each resistor can be found using any of the equations relating power to current, voltage, and resistance, since all three are known. Let us use $P = \frac{V^2}{R}$, since each resistor gets full voltage. Thus,

$$P_1 = \frac{V^2}{R_1} = \frac{(12.0\ \text{V})^2}{1.00\ \Omega} = 144\ \text{W}.$$

Similarly,

$$P_2 = \frac{V^2}{R_2} = \frac{(12.0\ \text{V})^2}{6.00\ \Omega} = 24.0\ \text{W}$$

and

$$P_3 = \frac{V^2}{R_3} = \frac{(12.0\ \text{V})^2}{13.0\ \Omega} = 11.1\ \text{W}$$

Discussion for (d)

The power dissipated by each resistor is considerably higher in parallel than when connected in series to the same voltage source.

Strategy and Solution for (e)

The total power can also be calculated in several ways. Choosing $P = IV$, and entering the total current, yields

$$P = IV = (14.92\ \text{A})(12.0\ \text{V}) = 179\ \text{W}.$$

Discussion for (e)

Total power dissipated by the resistors is also 179 W:

$$P_1 + P_2 + P_3 = 144\ \text{W} + 24.0\ \text{W} + 11.1\ \text{W} = 179\ \text{W}.$$

This is consistent with the law of conservation of energy.

Overall Discussion

Note that both the currents and powers in parallel connections are greater than for the same devices in series.

MAJOR FEATURES OF RESISTORS IN PARALLEL

1. Parallel resistance is found from $\frac{1}{R_\text{p}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots$, and it is smaller than any individual resistance in the combination.
2. Each resistor in parallel has the same full voltage of the source applied to it. (Power distribution systems most often use parallel connections to supply the myriad devices served with the same voltage and to allow them to operate independently.)
3. Parallel resistors do not each get the total current; they divide it.

Combinations of Series and Parallel

More complex connections of resistors are sometimes just combinations of series and parallel. These are commonly encountered, especially when wire resistance is considered. In that case, wire resistance is in series with other resistances that are in parallel.

Combinations of series and parallel can be reduced to a single equivalent resistance using the technique illustrated in Figure 3.1.11.4. Various parts are identified as either series or parallel, reduced to their equivalents, and further reduced until a single resistance is left. The process is more time consuming than difficult.

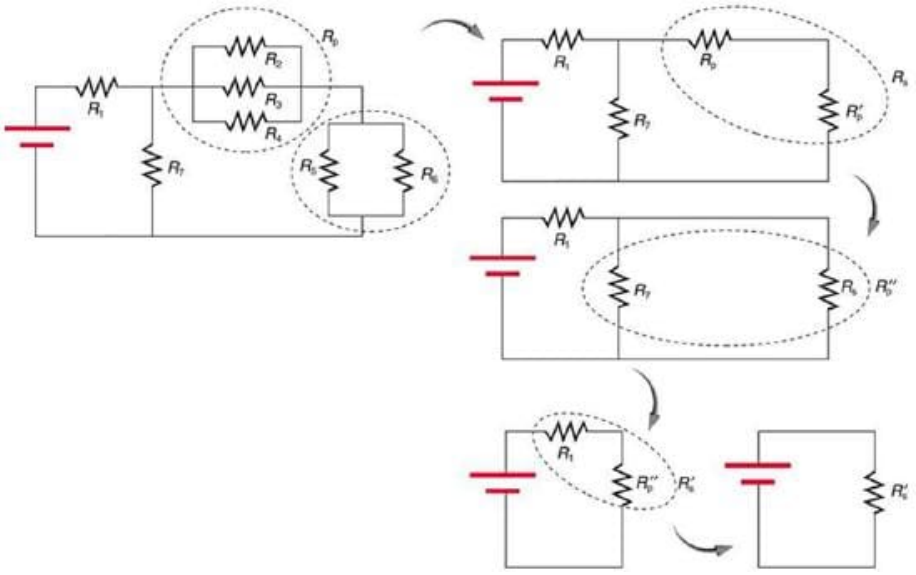


Figure 3.1.11.4: This combination of seven resistors has both series and parallel parts. Each is identified and reduced to an equivalent resistance, and these are further reduced until a single equivalent resistance is reached.

The simplest combination of series and parallel resistance, shown in Figure 3.1.11.5, is also the most instructive, since it is found in many applications. For example, R_1 could be the resistance of wires from a car battery to its electrical devices, which are in parallel. R_2 and R_3 could be the starter motor and a passenger compartment light. We have previously assumed that wire resistance is negligible, but, when it is not, it has important effects, as the next example indicates.

Example 3.1.11.3: Calculating Resistance, IR Drop, Current, and Power Dissipation: Combining Series and Parallel Circuits

Figure 3.1.11.5 shows the resistors from the previous two examples wired in a different way—a combination of series and parallel. We can consider R_1 to be the resistance of wires leading to R_2 and R_3. (a) Find the total resistance. (b) What is the IR drop in R_1? (c) Find the current I_2 through R_2. (d) What power is dissipated by R_2?

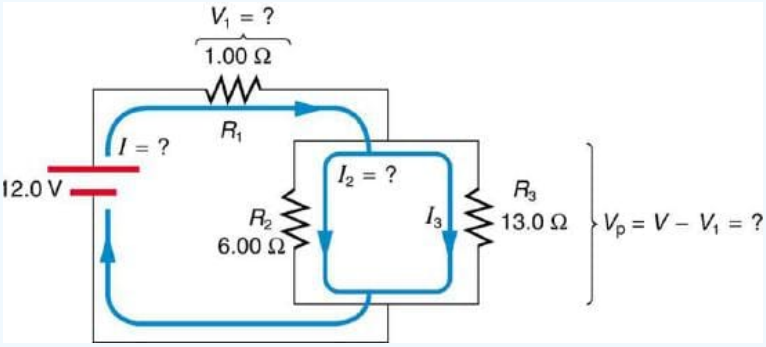


Figure 3.1.11.5: These three resistors are connected to a voltage source so that R_2 and R_3 are in parallel with one another and that combination is in series with R_1.

Strategy and Solution for (a)

To find the total resistance, we note that R_2 and R_3 are in parallel and their combination R_p is in series with R_1. Thus the total (equivalent) resistance of this combination is

$$R_\text{tot} = R_1 + R_\text{p}.$$

First, we find R_p using the equation for resistors in parallel and entering known values:

$$\frac{1}{R_\text{p}} = \frac{1}{R_2} + \frac{1}{R_3} = \frac{1}{6.00\ \Omega} + \frac{1}{13.0\ \Omega} = \frac{0.2436}{\Omega}.$$

Inverting gives

$$R_\text{p} = \frac{1}{0.2436}\ \Omega = 4.11\ \Omega.$$

So the total resistance is

$$R_\text{tot} = R_1 + R_\text{p} = 1.00\ \Omega + 4.11\ \Omega = 5.11\ \Omega.$$

Discussion for (a)

The total resistance of this combination is intermediate between the pure series and pure parallel values ($20.0\ \Omega$ and $0.804\ \Omega$, respectively) found for the same resistors in the two previous examples.

Strategy and Solution for (b)

To find the IR drop in R_1, we note that the full current I flows through R_1. Thus its IR drop is

$$V_1 = IR_1.$$

We must find I before we can calculate V_1. The total current I is found using Ohm's law for the circuit. That is,

$$I = \frac{V}{R_\text{tot}} = \frac{12.0\ \text{V}}{5.11\ \Omega} = 2.35\ \text{A}.$$

Entering this into the expression above, we get

$$V_1 = IR_1 = (2.35\ \text{A})(1.00\ \Omega) = 2.35\ \text{V}.$$

Discussion for (b)

The voltage applied to R_2 and R_3 is less than the total voltage by an amount V_1. When wire resistance is large, it can significantly affect the operation of the devices represented by R_2 and R_3.

Strategy and Solution for (c)

To find the current through R_2, we must first find the voltage applied to it. We call this voltage V_p, because it is applied to a parallel combination of resistors. The voltage applied to both R_2 and R_3 is reduced by the amount V_1, and so it is

$$V_\text{p} = V - V_1 = 12.0\ \text{V} - 2.35\ \text{V} = 9.65\ \text{V}.$$

Now the current I_2 through resistance R_2 is found using Ohm's law:

$$I_2 = \frac{V_\text{P}}{R_2} = \frac{9.65\ \text{V}}{6.00\ \Omega} = 1.61\ \text{A}.$$

Discussion for (c)

The current is less than the 2.00 A that flowed through R_2 when it was connected in parallel to the battery in the previous parallel circuit example.

Strategy and Solution for (d)

The power dissipated by R_2 is given by

$$P_2 = (I_2)^2 R_2 = (1.61\ \text{A})^2(6.00\ \Omega) = 15.5\ \text{W}.$$

Discussion for (d)

The power is less than the 24.0 W this resistor dissipated when connected in parallel to the 12.0-V source.

Practical Implications

One implication of this last example is that resistance in wires reduces the current and power delivered to a resistor. If wire resistance is relatively large, as in a worn (or a very long) extension cord, then this loss can be significant. If a large current is drawn, the IR drop in the wires can also be significant.

For example, when you are rummaging in the refrigerator and the motor comes on, the refrigerator light dims momentarily. Similarly, you can see the passenger compartment light dim when you start the engine of your car (although this may be due to resistance inside the battery itself).

What is happening in these high-current situations is illustrated in Figure 3.1.11.6. The device represented by R_3 has a very low resistance, and so when it is switched on, a large current flows. This increased current causes a larger IR drop in the wires represented by R_1, reducing the voltage across the light bulb (which is R_2), which then dims noticeably.

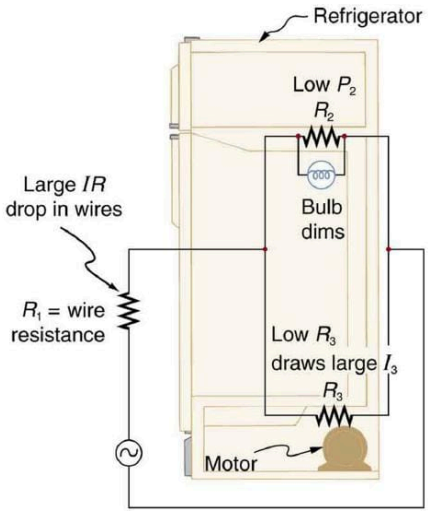


Figure 3.1.11.6: Why do lights dim when a large appliance is switched on? The answer is that the large current the appliance motor draws causes a significant IR drop in the wires and reduces the voltage across the light.

Exercise 3.1.11.1

Can any arbitrary combination of resistors be broken down into series and parallel combinations? See if you can draw a circuit diagram of resistors that cannot be broken down into combinations of series and parallel.

Answer available at the end of this section.

Answer available at the end of this section.

PROBLEM-SOLVING STRATEGIES FOR SERIES AND PARALLEL RESISTORS

1. Draw a clear circuit diagram, labeling all resistors and voltage sources. This step includes a list of the knowns for the problem, since they are labeled in your circuit diagram.
2. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful.
3. Determine whether resistors are in series, parallel, or a combination of both series and parallel. Examine the circuit diagram to make this assessment. Resistors are in series if the same current must pass sequentially through them.
4. Use the appropriate list of major features for series or parallel connections to solve for the unknowns. There is one list for series and another for parallel. If your problem has a combination of series and parallel, reduce it in steps by considering individual groups of series or parallel connections, as done in this module and the examples. Special note: When finding R_p, the reciprocal must be taken with care.
5. Check to see whether the answers are reasonable and consistent. Units and numerical results must be reasonable. Total series resistance should be greater, whereas total parallel resistance should be smaller, for example. Power should be greater for the same devices in parallel compared with series, and so on.

Section Summary

- The total resistance of an electrical circuit with resistors wired in a series is the sum of the individual resistances: $R_{\text{s}} = R_1 + R_2 + R_3 + \ldots$.
- Each resistor in a series circuit has the same amount of current flowing through it.
- The voltage drop, or power dissipation, across each individual resistor in a series is different, and their combined total adds up to the power source input.
- The total resistance of an electrical circuit with resistors wired in parallel is less than the lowest resistance of any of the components and can be determined using the formula:

$$\frac{1}{R_{\text{p}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots$$

- Each resistor in a parallel circuit has the same full voltage of the source applied to it.
- The current flowing through each resistor in a parallel circuit is different, depending on the resistance.
- If a more complex connection of resistors is a combination of series and parallel, it can be reduced to a single equivalent resistance by identifying its various parts as series or parallel, reducing each to its equivalent, and continuing until a single resistance is eventually reached.

Glossary

series
: a sequence of resistors or other components wired into a circuit one after the other

resistor
: a component that provides resistance to the current flowing through an electrical circuit

resistance
: causing a loss of electrical power in a circuit

Ohm's law
: the relationship between current, voltage, and resistance within an electrical circuit: $V = IR$

voltage
: the electrical potential energy per unit charge; electric pressure created by a power source, such as a battery

voltage drop
: the loss of electrical power as a current travels through a resistor, wire or other component

current
: the flow of charge through an electric circuit past a given point of measurement

Joule's law
: the relationship between potential electrical power, voltage, and resistance in an electrical circuit, given by: $P_e = IV$

parallel
: the wiring of resistors or other components in an electrical circuit such that each component receives an equal voltage from the power source; often pictured in a ladder-shaped diagram, with each component on a rung of the ladder

Answers

Exercise 3.1.11.1

Exercise 3.1.11.1

No, there are many ways to connect resistors that are not combinations of series and parallel, including loops and junctions. In such cases Kirchhoff's rules, to be introduced in Kirchhoff's Rules, will allow you to analyze the circuit.

This page titled 3.1.11: Resistors in Series and Parallel is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.11: Resistors in Series and Parallel by Andrew Park is licensed CC BY 4.0.

3.1.12: Electric Hazards and the Human Body

Learning Objectives

- Define thermal hazard, shock hazard, and short circuit.
- Explain what effects various levels of current have on the human body.

There are two known hazards of electricity—thermal and shock. A **thermal hazard** is one where excessive electric power causes undesired thermal effects, such as starting a fire in the wall of a house. A **shock hazard** occurs when electric current passes through a person. Shocks range in severity from painful, but otherwise harmless, to heart-stopping lethality. This section considers these hazards and the various factors affecting them in a quantitative manner.

Thermal Hazards

Electric power causes undesired heating effects whenever electric energy is converted to thermal energy at a rate faster than it can be safely dissipated. A classic example of this is the **short circuit**, a low-resistance path between terminals of a voltage source. An example of a short circuit is shown in Figure 3.1.12.1. Insulation on wires leading to an appliance has worn through, allowing the two wires to come into contact. Such an undesired contact with a high voltage is called a *short*. Since the resistance of the short, r, is very small, the power dissipated in the short, $P = V^2/r$, is very large. For example, if V is 120 V and r is $0.100\ \Omega$, then the power is 144 kW, *much* greater than that used by a typical household appliance. Thermal energy delivered at this rate will very quickly raise the temperature of surrounding materials, melting or perhaps igniting them.

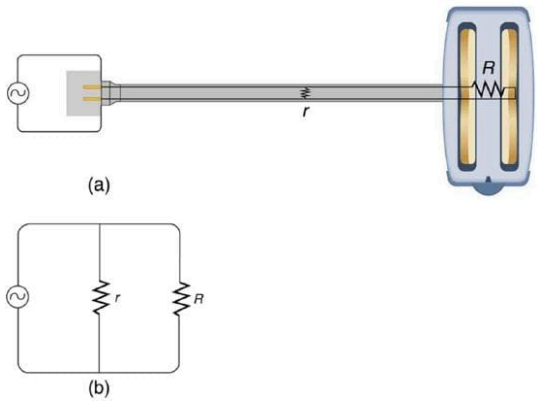


Figure 3.1.12.1: A short circuit is an undesired low-resistance path across a voltage source. (a) Worn insulation on the wires of a toaster allow them to come into contact with a low resistance r. Since $P = V^2/r$, thermal power is created so rapidly that the cord melts or burns. (b) A schematic of the short circuit.

One particularly insidious aspect of a short circuit is that its resistance may actually be decreased due to the increase in temperature. This can happen if the short creates ionization. These charged atoms and molecules are free to move and, thus, lower the resistance r. Since $P = V^2/r$, the power dissipated in the short rises, possibly causing more ionization, more power, and so on. High voltages, such as the 480-V AC used in some industrial applications, lend themselves to this hazard, because higher voltages create higher initial power production in a short.

Another serious, but less dramatic, thermal hazard occurs when wires supplying power to a user are overloaded with too great a current. As discussed in the previous section, the power dissipated in the supply wires is $P = I^2R_{\text{w}}$, where R_{w} is the resistance of the wires and I the current flowing through them. If either I or R_{w} is too large, the wires overheat. For example, a worn appliance cord (with some of its braided wires broken) may have $R_{\text{w}} = 2.00\ \Omega$ rather than the $0.100\ \Omega$ it should be. If 10.0 A of current passes through the cord, then $P = I^2R_{\text{w}} = 200\ \text{W}$ is dissipated in the cord—much more than is safe. Similarly, if a wire with a $0.100 - \Omega$ resistance is meant to carry a few amps, but is instead carrying 100 A, it will severely overheat. The power dissipated in the wire will in that case be $P = 1000\ \text{W}$. Fuses and circuit breakers are used to limit excessive currents. (See Figure 3.1.12.2 and Figure 3.1.12.3.) Each device opens the circuit automatically when a sustained current exceeds safe limits.

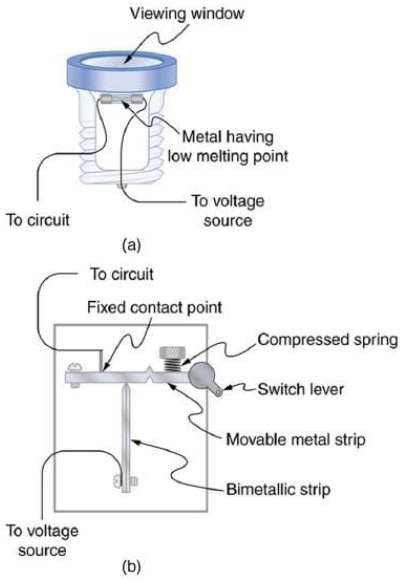


Figure 3.1.12.2: (a) A fuse has a metal strip with a low melting point that, when overheated by an excessive current, permanently breaks the connection of a circuit to a voltage source. (b) A circuit breaker is an automatic but restorable electric switch. The one shown here has a bimetallic strip that bends to the right and into the notch if overheated. The spring then forces the metal strip downward, breaking the electrical connection at the points.

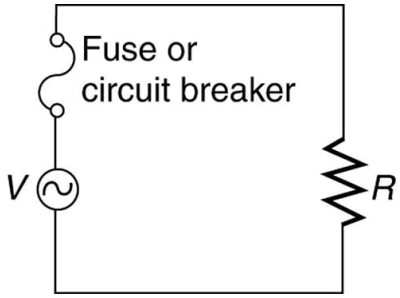


Figure 3.1.12.3: Schematic of a circuit with a fuse or circuit breaker in it. Fuses and circuit breakers act like automatic switches that open when sustained current exceeds desired limits.

Fuses and circuit breakers for typical household voltages and currents are relatively simple to produce, but those for large voltages and currents experience special problems. For example, when a circuit breaker tries to interrupt the flow of high-voltage electricity, a spark can jump across its points that ionizes the air in the gap and allows the current to continue flowing. Large circuit breakers found in power-distribution systems employ insulating gas and even use jets of gas to blow out such sparks. Here AC is safer than DC, since AC current goes through zero 120 times per second, giving a quick opportunity to extinguish these arcs.

Shock Hazards

Electrical currents through people produce tremendously varied effects. An electrical current can be used to block back pain. The possibility of using electrical current to stimulate muscle action in paralyzed limbs, perhaps allowing paraplegics to walk, is under study. TV dramatizations in which electrical shocks are used to bring a heart attack victim out of ventricular fibrillation (a massively irregular, often fatal, beating of the heart) are more than common. Yet most electrical shock fatalities occur because a current put the heart into fibrillation. A pacemaker uses electrical shocks to stimulate the heart to beat properly. Some fatal shocks do not produce burns, but warts can be safely burned off with electric current (though freezing using liquid nitrogen is now more common). Of course, there are consistent explanations for these disparate effects. The major factors upon which the effects of electrical shock depend are

1. The amount of current I
2. The path taken by the current
3. The duration of the shock
4. The frequency f of the current ($f = 0$ for DC)

Table 3.1.12.1 gives the effects of electrical shocks as a function of current for a typical accidental shock. The effects are for a shock that passes through the trunk of the body, has a duration of 1 s, and is caused by 60-Hz power.

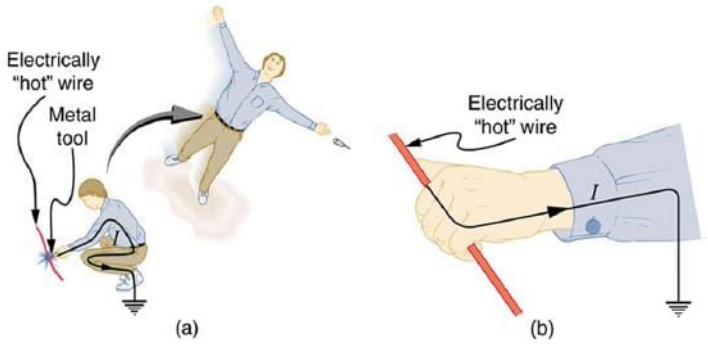


Figure 3.1.12.4: An electric current can cause muscular contractions with varying effects. (a) The victim is "thrown" backward by involuntary muscle contractions that extend the legs and torso. (b) The victim can't let go of the wire that is stimulating all the muscles in the hand. Those that close the fingers are stronger than those that open them.

Table 3.1.12.1*: Effects of Electrical Shock as a Function of Current*[1]

Current (mA)	Effect
1	Threshold of sensation
5	Maximum harmless current
10–20	Onset of sustained muscular contraction; cannot let go for duration of shock; contraction of chest muscles may stop breathing during shock
50	Onset of pain
100–300+	Ventricular fibrillation possible; often fatal
300	Onset of burns depending on concentration of current
6000 (6 A)	Onset of sustained ventricular contraction and respiratory paralysis; both cease when shock ends; heartbeat may return to normal; used to defibrillate the heart

Our bodies are relatively good conductors due to the water in our bodies. Given that larger currents will flow through sections with lower resistance (to be further discussed in the next chapter), electric currents preferentially flow through paths in the human body that have a minimum resistance in a direct path to earth. The earth is a natural electron sink. Wearing insulating shoes, a requirement in many professions, prohibits a pathway for electrons by providing a large resistance in that path. Whenever working with high-power tools (drills), or in risky situations, ensure that you do not provide a pathway for current flow (especially through the heart).

Very small currents pass harmlessly and unfelt through the body. This happens to you regularly without your knowledge. The threshold of sensation is only 1 mA and, although unpleasant, shocks are apparently harmless for currents less than 5 mA. A great number of safety rules take the 5-mA value for the maximum allowed shock. At 10 to 20 mA and above, the current can stimulate sustained muscular contractions much as regular nerve impulses do. People sometimes say they were knocked across the room by a shock, but what really happened was that certain muscles contracted, propelling them in a manner not of their own choosing. (See Figure 3.1.12.4(a).) More frightening, and potentially more dangerous, is the "can't let go" effect illustrated in Figure 3.1.12.4(b). The muscles that close the fingers are stronger than those that open them, so the hand closes involuntarily on the wire shocking it. This can prolong the shock indefinitely. It can also be a danger to a person trying to rescue the victim, because the rescuer's hand may close about the victim's wrist. Usually the best way to help the victim is to give the fist a hard knock/blow/jar with an insulator or to throw an insulator at the fist. Modern electric fences, used in animal enclosures, are now pulsed on and off to allow people who touch them to get free, rendering them less lethal than in the past.

Greater currents may affect the heart. Its electrical patterns can be disrupted, so that it beats irregularly and ineffectively in a condition called "ventricular fibrillation." This condition often lingers after the shock and is fatal due to a lack of blood circulation. The threshold for ventricular fibrillation is between 100 and 300 mA. At about 300 mA and above, the shock can cause burns, depending on the concentration of current—the more concentrated, the greater the likelihood of burns.

Very large currents cause the heart and diaphragm to contract for the duration of the shock. Both the heart and breathing stop. Interestingly, both often return to normal following the shock. The electrical patterns on the heart are completely erased in a manner that the heart can start afresh with normal beating, as opposed to the permanent disruption caused by smaller currents that can put the heart into ventricular fibrillation. The latter is something like scribbling on a blackboard, whereas the former completely erases it. TV dramatizations of electric shock used to bring a heart attack victim out of ventricular fibrillation also show large paddles. These are used to spread out current passed through the victim to reduce the likelihood of burns.

Current is the major factor determining shock severity (given that other conditions such as path, duration, and frequency are fixed, such as in the table and preceding discussion). A larger voltage is more hazardous, but since $I = V/R$, the severity of the shock depends on the combination of voltage and resistance. For example, a person with dry skin has a resistance of about $200\ \text{k}\Omega$. If he comes into contact with 120-V AC, a current $I = (120\ \text{V})/(200\ \text{k}\Omega) = 0.6\ \text{mA}$ passes harmlessly through him. The same person soaking wet may have a resistance of $10.0\ \text{k}\Omega$ and the same 120 V will produce a current of 12 mA—above the "can't let go" threshold and potentially dangerous.

Most of the body's resistance is in its dry skin. When wet, salts go into ion form, lowering the resistance significantly. The interior of the body has a much lower resistance than dry skin because of all the ionic solutions and fluids it contains. If skin resistance is bypassed, such as by an intravenous infusion, a catheter, or exposed pacemaker leads, a person is rendered **microshock sensitive**. In this condition, currents about 1/1000 those listed in Table 3.1.12.1 produce similar effects. During open-heart surgery, currents as small as $20\ \mu\text{A}$ can be used to still the heart. Stringent electrical safety requirements in hospitals, particularly in surgery and intensive care, are related to the doubly disadvantaged microshock-sensitive patient. The break in the skin has reduced his resistance, and so the same voltage causes a greater current, and a much smaller current has a greater effect.

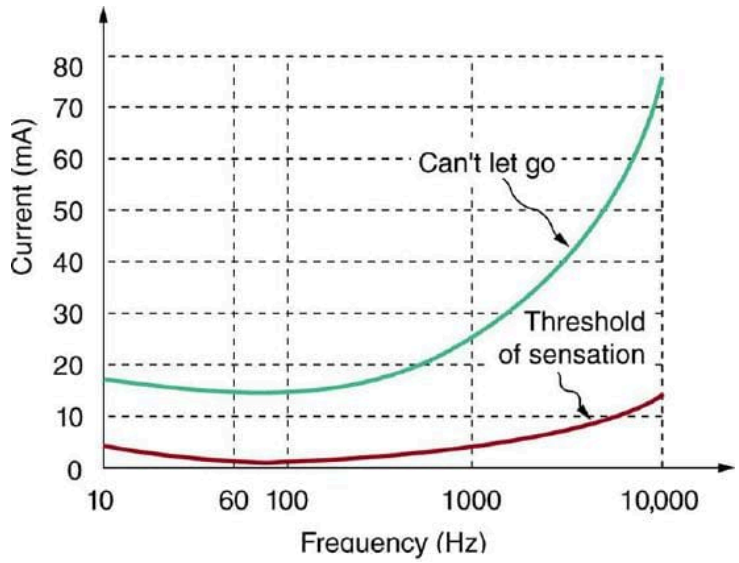


Figure 3.1.12.5: Graph of average values for the threshold of sensation and the "can't let go" current as a function of frequency. The lower the value, the more sensitive the body is at that frequency.

Factors other than current that affect the severity of a shock are its path, duration, and AC frequency. Path has obvious consequences. For example, the heart is unaffected by an electric shock through the brain, such as may be used to treat manic depression. And it is a general truth that the longer the duration of a shock, the greater its effects. Figure 3.1.12.5 presents a graph that illustrates the effects of frequency on a shock. The curves show the minimum current for two different effects, as a function of frequency. The lower the current needed, the more sensitive the body is at that frequency. Ironically, the body is most sensitive to frequencies near the 50- or 60-Hz frequencies in common use. The body is slightly less sensitive for DC ($f = 0$), mildly confirming Edison's claims that AC presents a greater hazard. At higher and higher frequencies, the body becomes progressively less sensitive to any effects that involve nerves. This is related to the maximum rates at which nerves can fire or be stimulated. At very high frequencies, electrical current travels only on the surface of a person. Thus a wart can be burned off with very high frequency current without causing the heart to stop. (Do not try this at home with 60-Hz AC!) Some of the spectacular demonstrations of electricity, in which high-voltage arcs are passed through the air and over people's bodies, employ high frequencies and low currents. (See Figure 3.1.12.6.)

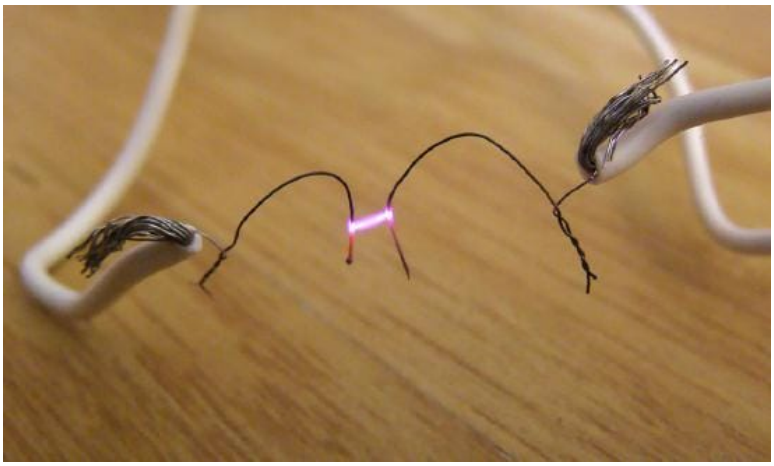


Figure 3.1.12.6: Is this electric arc dangerous? The answer depends on the AC frequency and the power involved. (credit: Khimich Alex, Wikimedia Commons)

Section Summary

- The two types of electric hazards are thermal (excessive power) and shock (current through a person).
- Shock severity is determined by current, path, duration, and AC frequency.
- Table 3.1.12.1 lists shock hazards as a function of current.
- Figure 3.1.12.5 graphs the threshold current for two hazards as a function of frequency.

Footnotes

- 1 For an average male shocked through trunk of body for 1 s by 60-Hz AC. Values for females are 60–80% of those listed.

Glossary

thermal hazard
: a hazard in which electric current causes undesired thermal effects

shock hazard
: when electric current passes through a person

short circuit
: also known as a “short,” a low-resistance path between terminals of a voltage source

microshock sensitive
: a condition in which a person’s skin resistance is bypassed, possibly by a medical procedure, rendering the person vulnerable to electrical shock at currents about 1/1000 the normally required level

This page titled 3.1.12: Electric Hazards and the Human Body is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.12: Electric Hazards and the Human Body by Andrew Park is licensed CC BY 4.0.

3.1.13: Electricity (Exercise)

Conceptual Questions

9.1: Static Electricity and Charge- Conservation of Charge

1. There are very large numbers of charged particles in most objects. Why, then, don't most objects exhibit static electricity?

2. Why do most objects tend to contain nearly equal numbers of positive and negative charges?

9.2: Coulomb's Law

3. Figure 3.1.13.1 shows the charge distribution in a water molecule, which is called a polar molecule because it has an inherent separation of charge. Given water's polar character, explain what effect humidity has on removing excess charge from objects.

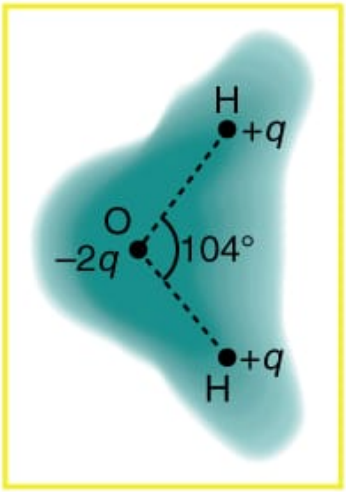


Figure 3.1.13.1: Schematic representation of the outer electron cloud of a neutral water molecule. The electrons spend more time near the oxygen than the hydrogens, giving a permanent charge separation as shown. Water is thus a *polar molecule*. It is more easily affected by electrostatic forces than molecules with uniform charge distributions.

4. Using Figure 3.1.13.1, explain, in terms of Coulomb's law, why a polar molecule (such as in Figure 3.1.13.1) is attracted by both positive and negative charges.

5. Given the polar character of water molecules, explain how ions in the air form nucleation centers for rain droplets.

9.3: Electric Field- Concept of a Field Revisited

6. Why must the test charge q in the definition of the electric field be vanishingly small?

7. Are the direction and magnitude of the Coulomb force unique at a given point in space? What about the electric field?

9.4: Electric Field Lines

8. Compare and contrast the Coulomb force field and the electric field. To do this, make a list of five properties for the Coulomb force field analogous to the five properties listed for electric field lines. Compare each item in your list of Coulomb force field properties with those of the electric field—are they the same or different? (For example, electric field lines cannot

cross. Is the same true for Coulomb field lines?)

9.5: Electric Potential and Potential Energy

9. Voltage is the common word for potential difference. Which term is more descriptive, voltage or potential difference?

10. If the voltage between two points is zero, can a test charge be moved between them with zero net work being done? Can this necessarily be done without exerting a force? Explain.

11. What is the relationship between voltage and energy? More precisely, what is the relationship between potential difference and electric potential energy?

12. Voltages are always measured between two points. Why?

13. How are units of volts and electron volts related? How do they differ?

9.6: Conductors and Applications of Electrostatics

14. Is the object in Figure 3.1.13.2 a conductor or an insulator? Justify your answer.

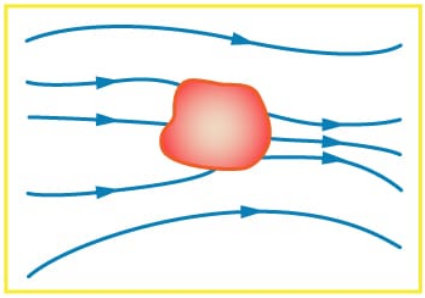


Figure
3.1.13.2

15. If the electric field lines in the figure above were perpendicular to the object, would it necessarily be a conductor? Explain.

16. The discussion of the electric field between two parallel conducting plates, in this module states that edge effects are less important if the plates are close together. What does close mean? That is, is the actual plate separation crucial, or is the ratio of plate separation to plate area crucial?

17. Would the self-created electric field at the end of a pointed conductor, such as a lightning rod, remove positive or negative charge from the conductor? Would the same sign charge be removed from a neutral pointed conductor by the application of a similar externally created electric field? (The answers to both questions have implications for charge transfer utilizing points.)

18. Why is a golfer with a metal club over her shoulder vulnerable to lightning in an open fairway? Would she be any safer under a tree?

19. Can the belt of a Van de Graaff accelerator be a conductor? Explain.

20. Are you relatively safe from lightning inside an automobile? Give two reasons.

21. Discuss pros and cons of a lightning rod being grounded versus simply being attached to a building.

22. Using the symmetry of the arrangement, show that the net Coulomb force on the charge q at the center of the square below (Figure 3.1.13.3) is zero if the charges on the four corners are exactly equal.

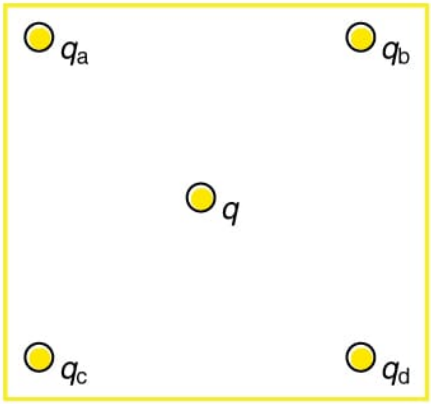


Figure 3.1.13.3: Four point charges q_a, q_b, q_c, and q_d lie on the corners of a square and q is located at its center.

23. (a) Using the symmetry of the arrangement, show that the electric field at the center of the square in Figure 3.1.13.3 is zero if the charges on the four corners are exactly equal.

(b) Show that this is also true for any combination of charges in which $q_a = q_d$ and $q_b = q_c$

24. (a) What is the direction of the total Coulomb force on q in Figure 3.1.13.3 if q is negative, $q_a = q_c$ and both are negative, and $q_b = q_c$ and both are positive?

(b) What is the direction of the electric field at the center of the square in this situation?

25. Considering Figure 3.1.13.3, suppose that $q_a = q_d$ and $q_b = q_c$. First show that q is in static equilibrium. (You may neglect the gravitational force.) Then discuss whether the equilibrium is stable or unstable, noting that this may depend on the signs of the charges and the direction of displacement of q from the center of the square.

26. If $q_a = 0$ in Figure 3.1.13.3, under what conditions will there be no net Coulomb force on q?

27. In regions of low humidity, one develops a special “grip” when opening car doors, or touching metal door knobs. This involves placing as much of the hand on the device as possible, not just the ends of one’s fingers. Discuss the induced charge and explain why this is done.

28. Tollbooth stations on roadways and bridges usually have a piece of wire stuck in the pavement before them that will touch a car as it approaches. Why is this done?

29. Suppose a woman carries an excess charge. To maintain her charged status can she be standing on ground wearing just any pair of shoes? How would you discharge her? What are the consequences if she simply walks away?

9.7: Current

30. Can a wire carry a current and still be neutral—that is, have a total charge of zero? Explain.

31. Car batteries are rated in ampere-hours ($A \cdot h$). To what physical quantity do ampere-hours correspond (voltage, charge, . . .), and what relationship do ampere-hours have to energy content?

32. Why are two conducting paths from a voltage source to an electrical device needed to operate the device?

33. In cars, one battery terminal is connected to the metal body. How does this allow a single wire to supply current to electrical devices rather than two wires?

34. Why isn’t a bird sitting on a high-voltage power line electrocuted? Contrast this with the situation in which a large bird hits two wires simultaneously with its wings.

9.8: Ohm's Law- Resistance and Simple Circuits

35. The IR drop across a resistor means that there is a change in potential or voltage across the resistor. Is there any change in current as it passes through a resistor? Explain.

36. How is the IR drop in a resistor similar to the pressure drop in a fluid flowing through a pipe?

9.9: Electric Power and Energy

37. Why do incandescent lightbulbs grow dim late in their lives, particularly just before their filaments break?

38. The power dissipated in a resistor is given by $P = V^2/R$, which means power decreases if resistance increases. Yet this power is also given by $P = I^2R$, which means power increases if resistance increases. Explain why there is no contradiction here.

9.10: Resistors in Series and Parallel

39. A switch has a variable resistance that is nearly zero when closed and extremely large when open, and it is placed in series with the device it controls. Explain the effect the switch in Figure 3.1.13.4 has on current when open and when closed.

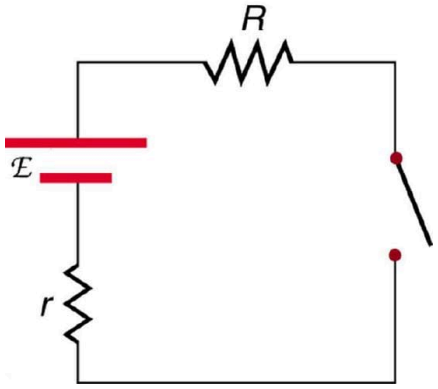


Figure 3.1.13.4: A switch is ordinarily in series with a resistance and voltage source. Ideally, the switch has nearly zero resistance when closed but has an extremely large resistance when open. (Note that in this diagram, the script E represents the voltage (or electromotive force) of the battery.)

40. What is the voltage across the open switch in Figure 3.1.13.4?

41. There is a voltage across an open switch, such as in Figure 3.1.13.4. Why, then, is the power dissipated by the open switch small?

42. Why is the power dissipated by a closed switch, such as in Figure 3.1.13.4, small?

43. A student in a physics lab mistakenly wired a light bulb, battery, and switch as shown in Figure 3.1.13.5. Explain why the bulb is on when the switch is open, and off when the switch is closed. (Do not try this—it is hard on the battery!)

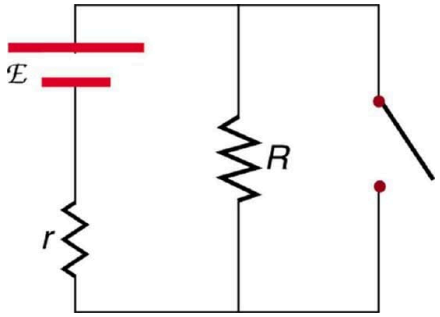


Figure 3.1.13.5: A wiring mistake put this switch in parallel with the device represented by R. (Note that in this diagram, the script E represents the voltage (or electromotive force) of the battery.)

44. Knowing that the severity of a shock depends on the magnitude of the current through your body, would you prefer to be in series or parallel with a resistance, such as the heating element of a toaster, if shocked by it? Explain.

45. Would your headlights dim when you start your car's engine if the wires in your automobile were superconductors? (Do not neglect the battery's internal resistance.) Explain.

46. Some strings of holiday lights are wired in series to save wiring costs. An old version utilized bulbs that break the electrical connection, like an open switch, when they burn out. If one such bulb burns out, what happens to the others? If such a string operates on 120 V and has 40 identical bulbs, what is the normal operating voltage of each? Newer versions use bulbs that short circuit, like a closed switch, when they burn out. If one such bulb burns out, what happens to the others? If such a string operates on 120 V and has 39 remaining identical bulbs, what is then the operating voltage of each?

47. If two household lightbulbs rated 60 W and 100 W are connected in series to household power, which will be brighter? Explain.

48. Suppose you are doing a physics lab that asks you to put a resistor into a circuit, but all the resistors supplied have a larger resistance than the requested value. How would you connect the available resistances to attempt to get the smaller value asked for?

49. Before World War II, some radios got power through a "resistance cord" that had a significant resistance. Such a resistance cord reduces the voltage to a desired level for the radio's tubes and the like, and it saves the expense of a transformer. Explain why resistance cords become warm and waste energy when the radio is on.

50. Some light bulbs have three power settings (not including zero), obtained from multiple filaments that are individually switched and wired in parallel. What is the minimum number of filaments needed for three power settings?

9.11: Electric Hazards and the Human Body

51. Using an ohmmeter, a student measures the resistance between various points on his body. He finds that the resistance between two points on the same finger is about the same as the resistance between two points on opposite hands—both are several hundred thousand ohms. Furthermore, the resistance decreases when more skin is brought into contact with the probes of the ohmmeter. Finally, there is a dramatic drop in resistance (to a few thousand ohms) when the skin is wet. Explain these observations and their implications regarding skin and internal resistance of the human body.

52. What are the two major hazards of electricity?

53. Why isn't a short circuit a shock hazard?

54. What determines the severity of a shock? Can you say that a certain voltage is hazardous without further information?

55. An electrified needle is used to burn off warts, with the circuit being completed by having the patient sit on a large butt plate. Why is this plate large?

56. Some surgery is performed with high-voltage electricity passing from a metal scalpel through the tissue being cut. Considering the nature of electric fields at the surface of conductors, why would you expect most of the current to flow from the sharp edge of the scalpel? Do you think high- or low-frequency AC is used?

57. Some devices often used in bathrooms, such as hairdryers, often have safety messages saying “Do not use when the bathtub or basin is full of water.” Why is this so?

58. We are often advised to not flick electric switches with wet hands, dry your hand first. We are also advised to never throw water on an electric fire. Why is this so?

59. Before working on a power transmission line, linemen will touch the line with the back of the hand as a final check that the voltage is zero. Why the back of the hand?

60. Why is the resistance of wet skin so much smaller than dry, and why do blood and other bodily fluids have low resistances?

61. Could a person on intravenous infusion (an IV) be microshock sensitive?

62. In view of the small currents that cause shock hazards and the larger currents that circuit breakers and fuses interrupt, how do they play a role in preventing shock hazards?

Problems & Exercises

9.1: Static Electricity and Charge- Conservation of Charge

63. Common static electricity involves charges ranging from nanocoulombs to microcoulombs.

(a) How many electrons are needed to form a charge of –2.00 nC

(b) How many electrons must be removed from a neutral object to leave a net charge of $0.500\ \mu\text{C}$?

Solution

(a) 1.25×1010

(b) 3.13×1012

64. If 1.80×10^{20} electrons move through a pocket calculator during a full day’s operation, how many coulombs of charge moved through it?

65. To start a car engine, the car battery moves 3.75×10^{21} electrons through the starter motor. How many coulombs of charge were moved?

Solution

-600 C

66. A certain lightning bolt moves 40.0 C of charge. How many fundamental units of charge $|q_e|$ is this?

9.2: Coulomb’s Law

67. What is the repulsive force between two pith balls that are 8.00 cm apart and have equal charges of – 30.0 nC?

68. (a) How strong is the attractive force between a glass rod with a $0.700\ \mu\text{C}$ charge and a silk cloth with a $-0.600\ \mu\text{C}$ charge, which are 12.0 cm apart, using the approximation that they act like point charges?

(b) Discuss how the answer to this problem might be affected if the charges are distributed over some area and do not act like point charges.

Solution

(a) 0.263 N

(b) If the charges are distributed over some area, there will be a concentration of charge along the side closest to the oppositely charged object. This effect will increase the net force.

69. Two point charges exert a 5.00 N force on each other. What will the force become if the distance between them is increased by a factor of three?

70. Two point charges are brought closer together, increasing the force between them by a factor of 25. By what factor was their separation decreased?

Solution

The separation decreased by a factor of 5.

71. How far apart must two point charges of 75.0 nC (typical of static electricity) be to have a force of 1.00 N between them?

72. If two equal charges each of 1 C each are separated in air by a distance of 1 km, what is the magnitude of the force acting between them? You will see that even at a distance as large as 1 km, the repulsive force is substantial because 1 C is a very significant amount of charge.

73. Bare free charges do not remain stationary when close together. To illustrate this, calculate the acceleration of two isolated protons separated by 2.00 nm (a typical distance between gas atoms).

Solution

$$\begin{aligned} F = k\frac{|q_1 q_2|}{r^2} = ma \Rightarrow a = \frac{kq^2}{mr^2} \\ = \frac{\left(9.00 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2\right)\left(1.60 \times 10^{-19} \text{ m}\right)^2}{\left(1.67 \times 10^{-27} \text{ kg}\right)\left(2.00 \times 10^{-9} \text{ m}\right)^2} \\ = 3.45 \times 10^{16} \text{ m/s}^2 \end{aligned}$$

74. (a) By what factor must you change the distance between two point charges to change the force between them by a factor of 10?

(b) Explain how the distance can either increase or decrease by this factor and still cause a factor of 10 change in the force.

Solution

(a) 3.2

(b) If the distance increases by 3.2, then the force will decrease by a factor of 10 ; if the distance decreases by 3.2, then the force will increase by a factor of 10. Either way, the force changes by a factor of 10.

75. Suppose you have a total charge q_{tot} that you can split in any manner. Once split, the separation distance is fixed. How do you split the charge to achieve the greatest force?

76. (a) Common transparent tape becomes charged when pulled from a dispenser. If one piece is placed above another, the repulsive force can be great enough to support the top piece's weight. Assuming equal point charges (only an approximation), calculate the magnitude of the charge if electrostatic force is great enough to support the weight of a 10.0 mg piece of tape held 1.00 cm above another.

(b) Discuss whether the magnitude of this charge is consistent with what is typical of static electricity.

Solution

(a) 1.04×10^{-9} C

(b) This charge is approximately 1 nC, which is consistent with the magnitude of charge typical for static electricity

77. (a) Find the ratio of the electrostatic to gravitational force between two electrons.

(b) What is this ratio for two protons?

(c) Why is the ratio different for electrons and protons?

78. At what distance is the electrostatic force between two protons equal to the weight of one proton?

79. A certain five cent coin contains 5.00 g of nickel. What fraction of the nickel atoms' electrons, removed and placed 1.00 m above it, would support the weight of this coin? The atomic mass of nickel is 58.7, and each nickel atom contains 28 electrons and 28 protons.

Solution

1.02×10^{-11}

80. (a) Two point charges totaling $8.00\ \mu\text{C}$ exert a repulsive force of 0.150 N on one another when separated by 0.500 m. What is the charge on each?

(b) What is the charge on each if the force is attractive?

9.3: Electric Field- Concept of a Field Revisited

81. What is the magnitude and direction of an electric field that exerts a 2.00×10^{-5} N upward force on a $-1.75\ \mu\text{C}$ charge?

82. What is the magnitude and direction of the force exerted on a $3.50\ \mu\text{C}$ charge by a 250 N/C electric field that points due east?

Solution

8.75×10^{-4} N

83. Calculate the magnitude of the electric field 2.00 m from a point charge of 5.00 mC (such as found on the terminal of a Van de Graaff).

84. (a) What magnitude point charge creates a 10,000 N/C electric field at a distance of 0.250 m?

(b) How large is the field at 10.0 m?

Solution

(a) 6.94×10^{-8} C

(b) 6.25 N/C

85. Calculate the initial (from rest) acceleration of a proton in a 5.00×10^{6} N/C electric field (such as created by a research Van de Graaff). Explicitly show how you follow the steps in the Problem-Solving Strategy for electrostatics.

86. (a) Find the direction and magnitude of an electric field that exerts a 4.80×10^{-17} N westward force on an electron.

(b) What magnitude and direction force does this field exert on a proton?

Solution

(a) 300 N/C (east)

(b) 4.80×10^{-17} N (east)

9.4: Electric Field Lines

87. (a) Sketch the electric field lines near a point charge $+q$. (b) Do the same for a point charge $-3.00q$.

88. Sketch the electric field lines a long distance from the charge distributions shown in Figure 10.4.4 (a) and (b)

89. Figure $3.1.13.6$ shows the electric field lines near two charges q_1 and q_2. What is the ratio of their magnitudes? (b) Sketch the electric field lines a long distance from the charges shown in the figure.

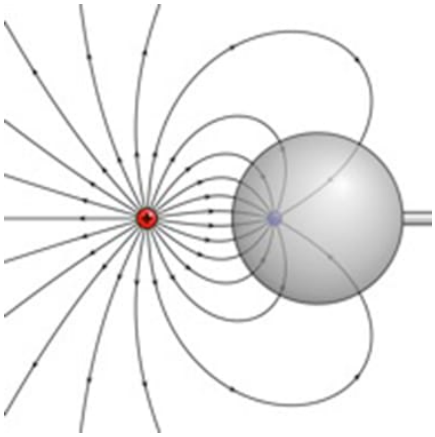


Figure $3.1.13.6$: The electric field near two charges.

90. Sketch the electric field lines in the vicinity of two opposite charges, where the negative charge is three times greater in magnitude than the positive. (See Figure $3.1.13.6$ for a similar situation).

9.5: Electric Potential and Potential Energy

91. Find the ratio of speeds of an electron and a negative hydrogen ion (one having an extra electron) accelerated through the same voltage, assuming non-relativistic final speeds. Take the mass of the hydrogen ion to be 1.67×10^{-27} kg.

Solution

42.8

92. An evacuated tube uses an accelerating voltage of 40 kV to accelerate electrons to hit a copper plate and produce X-rays. Non-relativistically, what would be the maximum speed of these electrons?

93. A bare helium nucleus has two positive charges and a mass of 6.64×10^{-27} kg

(a) Calculate its kinetic energy in joules at 2.00% of the speed of light.

(b) What is this in electron volts?

(c) What voltage would be needed to obtain this energy?

Integrated Concepts

94. Singly charged gas ions are accelerated from rest through a voltage of 13.0 V. At what temperature will the average kinetic energy of gas molecules be the same as that given these ions?

Solution

1.00×10^5 K

Integrated Concepts

95. The temperature near the center of the Sun is thought to be 15 million degrees Celsius $(1.5 \times 10^7 ^\circ\text{C})$. Through what voltage must a singly charged ion be accelerated to have the same energy as the average kinetic energy of ions at this temperature?

Integrated Concepts

96. (a) What is the average power output of a heart defibrillator that dissipates 400 J of energy in 10.0 ms?

(b) Considering the high-power output, why doesn't the defibrillator produce serious burns?

Solution

(a) $4 \times 10^4 \text{ W}$

(b) A defibrillator does not cause serious burns because the skin conducts electricity well at high voltages, like those used in defibrillators. The gel used aids in the transfer of energy to the body, and the skin doesn't absorb the energy, but rather lets it pass through to the heart.

Integrated Concepts

97. A lightning bolt strikes a tree, moving 20.0 C of charge through a potential difference of $1.00 \times 10^2 \text{ MV}$.

(a) What energy was dissipated?

(b) What mass of water could be raised from 15°C to the boiling point and then boiled by this energy?

(c) Discuss the damage that could be caused to the tree by the expansion of the boiling steam.

Integrated Concepts

98. A 12.0 V battery-operated bottle warmer heats 50.0 g of glass, $2.50 \times 10^2 \text{ g}$ of baby formula, and $2.00 \times 10^2 \text{ g}$ of aluminum from 20.0°C to 90.0°C.

(a) How much charge is moved by the battery?

(b) How many electrons per second flow if it takes 5.00 min to warm the formula? (Hint: Assume that the specific heat of baby formula is about the same as the specific heat of water.)

Solution

(a) $7.40 \times 10^3 \text{C}$

(b) $1.54 \times 10^{20} \text{ electrons per second}$

Integrated Concepts

99. A battery-operated car utilizes a 12.0 V system. Find the charge the batteries must be able to move in order to accelerate the 750 kg car from rest to 25.0 m/s, make it climb a $2.00 \times 10^2 \text{ m}$ high hill, and then cause it to travel at a constant 25.0 m/s by exerting a $5.00 \times 10^2 \text{ N}$ force for an hour.

Solution

$3.89 \times 10^6 \text{C}$

Integrated Concepts

100. Fusion probability is greatly enhanced when appropriate nuclei are brought close together, but mutual Coulomb repulsion must be overcome. This can be done using the kinetic energy of high-temperature gas ions or by accelerating the nuclei toward one another.

(a) Calculate the potential energy of two singly charged nuclei separated by $1.00 \times 10^{-12} \text{ m}$ by finding the voltage of

one at that distance and multiplying by the charge of the other.

(b) At what temperature will atoms of a gas have an average kinetic energy equal to this needed electrical potential energy?

9.7: Current

101. What is the current in milliamperes produced by the solar cells of a pocket calculator through which 4.00 C of charge passes in 4.00 h?

Solution

0.278 mA

102. A total of 600 C of charge passes through a flashlight in 0.500 h. What is the average current?

103. What is the current when a typical static charge of $0.250\ \mu\text{C}$ moves from your finger to a metal doorknob in $1.00\ \mu\text{s}$??

Solution

0.250 A

104. Find the current when 2.00 nC jumps between your comb and hair over a $0.500 - \mu\text{s}$ time interval.

105. A large lightning bolt had a 20,000-A current and moved 30.0 C of charge. What was its duration?

Solution

1.50ms

106. The 200-A current through a spark plug moves 0.300 mC of charge. How long does the spark last?

107. (a) A defibrillator passes 12.0 A of current through the torso of a person for 0.0100 s. How much charge moves?

(b) How many electrons pass through the wires connected to the patient? (See figure two problems earlier.)

Solution

(a) 0.120 C

(b) $7.50 \times 10^{17}\ \text{electrons}$

108. A clock battery wears out after moving 10,000 C of charge through the clock at a rate of 0.500 mA.

(a) How long did the clock run?

(b) How many electrons per second flowed?

109. The batteries of a submerged non-nuclear submarine supply 1000 A at full speed ahead. How long does it take to move Avogadro's number (6.02×10^{23}) of electrons at this rate?

Solution

96.3 s

110. Electron guns are used in X-ray tubes. The electrons are accelerated through a relatively large voltage and directed onto a metal target, producing X-rays.

(a) How many electrons per second strike the target if the current is 0.500 mA?

(b) What charge strikes the target in 0.750 s?

112. A large cyclotron directs a beam of He^{++} nuclei onto a target with a beam current of 0.250 mA.

(a) How many He^{++} nuclei per second is this?

(b) How long does it take for 1.00 C to strike the target?

(c) How long before 1.00 mol of He^{++} nuclei strike the target?

Solution

(a) $7.81 \times 10^{14}\ \mathrm{He}^{++}\mathrm{nuclei/s}$

(b) $4.00 \times 10^{3}\ \mathrm{s}$

(c) $7.71 \times 10^{8}\ \mathrm{s}$

9.8: Ohm's Law- Resistance and Simple Circuits

113. What current flows through the bulb of a 3.00-V flashlight when its hot resistance is $3.60\ \Omega$?

Solution

0.833 A

114. Calculate the effective resistance of a pocket calculator that has a 1.35-V battery and through which 0.200 mA flows.

115. What is the effective resistance of a car's starter motor when 150 A flows through it as the car battery applies 11.0 V to the motor?

Solution

$$7.33 \times 10^{-2}\ \Omega$$

116. How many volts are supplied to operate an indicator light on a DVD player that has a resistance of $140\ \Omega$, given that 25.0 mA passes through it?

117. (a) Find the voltage drop in an extension cord having a $0.0600 - \Omega$ resistance and through which 5.00 A is flowing.

(b) A cheaper cord utilizes thinner wire and has a resistance of $0.300\ \Omega$. What is the voltage drop in it when 5.00 A flows?

(c) Why is the voltage to whatever appliance is being used reduced by this amount? What is the effect on the appliance?

Solution

(a) 0.300 V

(b) 1.50 V

(c) The voltage supplied to whatever appliance is being used is reduced because the total voltage drop from the wall to the final output of the appliance is fixed. Thus, if the voltage drop across the extension cord is large, the voltage drop across the appliance is significantly decreased, so the power output by the appliance can be significantly decreased, reducing the ability of the appliance to work properly.

118. A power transmission line is hung from metal towers with glass insulators having a resistance of $1.00 \times 10^{9}\ \Omega$ What current flows through the insulator if the voltage is 200 kV? (Some high-voltage lines are DC.)

9.9: Electric Power and Energy

119. What is the power of a $1.00 \times 10^2 \text{MV}$ lightning bolt having a current of 2.00×10^4 A?

Solution

2.00×10^{12} W

120. What power is supplied to the starter motor of a large truck that draws 250 A of current from a 24.0-V battery hookup?

121. A charge of 4.00 C of charge passes through a pocket calculator's solar cells in 4.00 h. What is the power output, given the calculator's voltage output is 3.00 V? (See Figure 3.1.13.7.)



Figure 3.1.13.7: The strip of solar cells just above the keys of this calculator convert light to electricity to supply its energy needs. (credit: Evan-Amos, Wikimedia Commons)

122. How many watts does a flashlight that has 6.00×10^2 C pass through it in 0.500 h use if its voltage is 3.00 V?

123. Find the power dissipated in each of these extension cords:

(a) an extension cord having a $0.0600 - \Omega$ resistance and through which 5.00 A is flowing;

(b) a cheaper cord utilizing thinner wire and with a resistance of $0.300\ \Omega$.

Solution

(a) 1.50 W

(b) 7.50 W

124. Verify that the units of a volt-ampere are watts, as implied by the equation $P = IV$.

125. Show that the units $1\ \text{V}^2/\Omega = 1\ \text{W}$, as implied by the equation $P = V^2/R$.

Solution

$\frac{\text{V}^2}{\Omega} = \frac{\text{V}^2}{\text{V/A}} = \text{AV} = \left(\frac{\text{C}}{\text{s}}\right)\left(\frac{\text{J}}{\text{C}}\right) = \frac{\text{J}}{\text{s}} = 1\ \text{W}$

126. Show that the units $1\ \text{A}^2 \cdot \Omega = 1\ \text{W}$, as implied by the equation $P = I^2R$.

127. Verify the energy unit equivalence that $1\ \text{kW} \cdot \text{h} = 3.60 \times 10^6$ J.

Solution

$1\ \text{kW} \cdot \text{h} = \left(\frac{1 \times 10^3\ \text{J}}{1\ \text{s}}\right)(1\ \text{h})\left(\frac{3600\ \text{s}}{1\ \text{h}}\right) = 3.60 \times 10^6\ \text{J}$

128. Electrons in an X-ray tube are accelerated through 1.00×10^2 kV and directed toward a target to produce X-rays. Calculate the power of the electron beam in this tube if it has a current of 15.0 mA.

129. An electric water heater consumes 5.00 kW for 2.00 h per day. What is the cost of running it for one year if electricity costs 12.0 cents $/\mathrm{kW} \cdot \mathrm{h}$? See Figure 3.1.13.8.



Figure 3.1.13.8: On-demand electric hot water heater. Heat is supplied to water only when needed. (credit: aviddavid, Flickr)

Solution

$438/y

130. With a 1200-W toaster, how much electrical energy is needed to make a slice of toast (cooking time = 1 minute)? At 9.0 cents $/\mathrm{kW} \cdot \mathrm{h}$, how much does this cost?

131. Some makes of older cars have 6.00-V electrical systems.

(a) What is the hot resistance of a 30.0-W headlight in such a car?

(b) What current flows through it?

132. Alkaline batteries have the advantage of putting out constant voltage until very nearly the end of their life. How long will an alkaline battery rated at $1.00\ \mathrm{A} \cdot \mathrm{h}$ and 1.58 V keep a 1.00-W flashlight bulb burning?

Solution

1.58 h

133. A cauterizer, used to stop bleeding in surgery, puts out 2.00 mA at 15.0 kV.

(a) What is its power output?

(b) What is the resistance of the path?

134. The average television is said to be on 6 hours per day. Estimate the yearly cost of electricity to operate 100 million TVs, assuming their power consumption averages 150 W and the cost of electricity averages 12.0 cents$/\mathrm{kW} \cdot \mathrm{h}$.

Solution

$3.94 billion/year

9.10: Resistors in Series and Parallel

Note: Data taken from figures can be assumed to be accurate to three significant digits.

135. (a) What is the resistance of ten $275 - \Omega$ resistors connected in series?

(b) In parallel?

Solution

(a) $2.75\ \text{k}\Omega$

(b) $27.5\ \Omega$

136. (a) What is the resistance of a $1.00 \times 10^2 - \Omega$, a $2.50 - \text{k}\ \Omega$, and a $4.00 - \text{k}\ \Omega$ resistor connected in series?

(b) In parallel?

137. What are the largest and smallest resistances you can obtain by connecting a $36.0 - \Omega$, a $50.0 - \Omega$, and a $700 - \Omega$ resistor together?

Solution

(a) $786\ \Omega$

(b) $20.3\ \Omega$

138. An 1800-W toaster, a 1400-W electric frying pan, and a 75-W lamp are plugged into the same outlet in a 15-A, 120-V circuit. (The three devices are in parallel when plugged into the same socket.).

(a) What current is drawn by each device?

(b) Will this combination blow the 15-A fuse?

139. Your car's 30.0-W headlight and 2.40-kW starter are ordinarily connected in parallel in a 12.0-V system. What power would one headlight and the starter consume if connected in series to a 12.0-V battery? (Neglect any other resistance in the circuit and any change in resistance in the two devices.)

Solution

$29.6\ \text{W}$

140. (a) Given a 48.0-V battery and $24.0 - \Omega$ and $96.0 - \Omega$ resistors, find the current and power for each when connected in series.

(b) Repeat when the resistances are in parallel.

141. Referring to the example combining series and parallel circuits and Figure 10.10.5, calculate I_3 in the following two different ways: (a) from the known values of I and I_2; (b) using Ohm's law for R_3. In both parts explicitly show how you follow the steps in the Problem-Solving Strategies for Series and Parallel Resistors.

Solution

(a) 0.74 A

(b) 0.742 A

142. Referring to Figure: 10.10.5(a) Calculate P_3 and note how it compares with P_3 found in the first two example problems in this module. (b) Find the total power supplied by the source and compare it with the sum of the powers dissipated by the resistors.

143. Refer to Figure 10.10.6 and the discussion of lights dimming when a heavy appliance comes on. (a) Given the voltage source is 120 V, the wire resistance is $0.400\ \Omega$, and the bulb is nominally 75.0 W, what power will the bulb dissipate if a total of 15.0 A passes through the wires when the motor comes on? Assume negligible change in bulb resistance. (b) What power is consumed by the motor?

Solution

(a) 60.8 W

(b) 3.18 kW

144. A 240-kV power transmission line carrying 5.00×10^2 A is hung from grounded metal towers by ceramic insulators, each having a $1.00 \times 10^9 - \Omega$ resistance. Figure 3.1.13.9. (a) What is the resistance to ground of 100 of these insulators? (b) Calculate the power dissipated by 100 of them. (c) What fraction of the power carried by the line is this? Explicitly show how you follow the steps in the Problem-Solving Strategies for Series and Parallel Resistors.

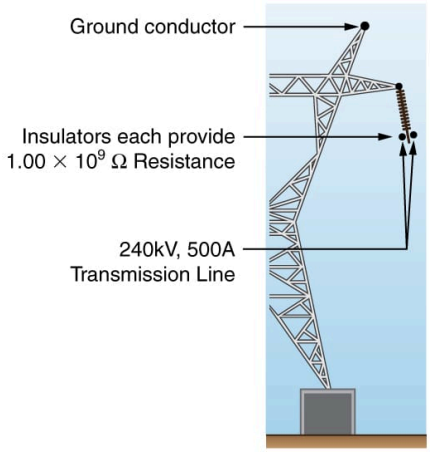


Figure 3.1.13.9: High-voltage (240-kV) transmission line carrying 5.00×10^2 A is hung from a grounded metal transmission tower. The row of ceramic insulators provide 1.00×10^9 Ω of resistance each.

145. Show that if two resistors R_1 and R_2 are combined and one is much greater than the other $(R_1 >> R_2)$:

(a) Their series resistance is very nearly equal to the greater resistance R_1.

(b) Their parallel resistance is very nearly equal to smaller resistance R_2.

Solution

(a) $$R_s = R_1 + R_2$$
$$\Rightarrow R_s \approx R_1 \, (R_1 >> R_2)$$

(b) $\frac{1}{R_p} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{R_1 + R_2}{R_1 R_2}$,

so that

$R_p = \frac{R_1 R_2}{R_1 + R_2} \approx \frac{R_1 R_2}{R_1} = R_2 \, (R_1 >> R_2)$.

Unreasonable Results

146. Two resistors, one having a resistance of $145\ \Omega$, are connected in parallel to produce a total resistance of $150\ \Omega$.

(a) What is the value of the second resistance?

(b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Unreasonable Results

147. Two resistors, one having a resistance of $900\ \text{k}\Omega$, are connected in series to produce a total resistance of $0.500\text{M}\ \Omega$.

(a) What is the value of the second resistance?

(b) What is unreasonable about this result?

(c) Which assumptions are unreasonable or inconsistent?

Solution

(a) $-400\ \text{k}\Omega$

(b) Resistance cannot be negative.

(c) Series resistance is said to be less than one of the resistors, but it must be greater than any of the resistors.

9.11: Electric Hazards and the Human Body

148. (a) How much power is dissipated in a short circuit of 240-V AC through a resistance of $0.250\ \Omega$?

(b) What current flows?

Solution

(a) 230 kW

(b) 960 A

149. What voltage is involved in a 1.44-kW short circuit through a $0.100 - \Omega$ resistance?

150. Find the current through a person and identify the likely effect on her if she touches a 120-V AC source:

(a) if she is standing on a rubber mat and offers a total resistance of $300\ \text{k}\Omega$;

(b) if she is standing barefoot on wet grass and has a resistance of only $4000\ \text{k}\Omega$.

Solution

(a) 0.400 mA, no effect

(b) 26.7 mA, muscular contraction for duration of the shock (can't let go)

151. While taking a bath, a person touches the metal case of a radio. The path through the person to the drainpipe and ground has a resistance of $4000\ \Omega$. What is the smallest voltage on the case of the radio that could cause ventricular fibrillation?

152. Foolishly trying to fish a burning piece of bread from a toaster with a metal butter knife, a man comes into contact with 120-V AC. He does not even feel it since, luckily, he is wearing rubber-soled shoes. What is the minimum resistance of the path the current follows through the person?

Solution

$1.20 \times 10^5\ \Omega$

153. (a) During surgery, a current as small as $20.0\mu\text{A}$ applied directly to the heart may cause ventricular fibrillation. If the resistance of the exposed heart is $300\ \Omega$, what is the smallest voltage that poses this danger?

(b) Does your answer imply that special electrical safety precautions are needed?

154. (a) What is the resistance of a 220-V AC short circuit that generates a peak power of 96.8 kW?

(b) What would the average power be if the voltage was 120 V AC?

Solution

(a) $1.00\ \Omega$

(b) 14.4 kW

155. A heart defibrillator passes 10.0 A through a patient's torso for 5.00 ms in an attempt to restore normal beating.

(a) How much charge passed?

(b) What voltage was applied if 500 J of energy was dissipated?

(c) What was the path's resistance?

(d) Find the temperature increase caused in the 8.00 kg of affected tissue.

Integrated Concepts

156. A short circuit in a 120-V appliance cord has a $0.500\ -\Omega$ resistance. Calculate the temperature rise of the 2.00 g of surrounding materials, assuming their specific heat capacity is $0.200 \mathrm{cal/g} \cdot {}^{\circ}\mathrm{C}$ and that it takes 0.0500 s for a circuit breaker to interrupt the current. Is this likely to be damaging?

Solution

Temperature increases $860^{\circ}\ \mathrm{C}$. It is very likely to be damaging.

Construct Your Own Problem

157. Consider a person working in an environment where electric currents might pass through her body. Construct a problem in which you calculate the resistance of insulation needed to protect the person from harm. Among the things to be considered are the voltage to which the person might be exposed, likely body resistance (dry, wet, ...), and acceptable currents (safe but sensed, safe and unfelt, ...).

This page titled 3.1.13: Electricity (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 10.13: Electricity (Exercise) by Andrew Park is licensed CC BY 4.0.

3.2: MAGNETISM

3.2.1: Introduction to Magnetism

3.2.2: Magnets

3.2.3: Ferromagnets and Electromagnets

3.2.4: Magnetic Fields and Magnetic Field Lines

3.2.5: Magnetic Field Strength- Force on a Moving Charge in a Magnetic Field

3.2.6: Magnetic Force on a Current-Carrying Conductor

3.2.7: Motors and Meters

3.2.8: Magnetic Fields Produced by Currents- Ampere's Law

3.2.9: Induced Voltage and Magnetic Flux

3.2.10: Faraday's Law of Induction- Lenz's Law

3.2.11: Transformers

3.2.12: Alternating Current versus Direct Current

3.2.13: Magnetism (Exercise)

This page titled 3.2: Magnetism is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

3.2.1: Introduction to Magnetism



Figure 3.2.1.1: The magnificent spectacle of the Aurora Borealis, or northern lights, glows in the northern sky above Bear Lake near Eielson Air Force Base, Alaska. Shaped by the Earth's magnetic field, this light is produced by radiation spewed from solar storms. (credit: Senior Airman Joshua Strang, via Flickr)

One evening, an Alaskan sticks a note to his refrigerator with a small magnet. Through the kitchen window, the Aurora Borealis glows in the night sky. This grand spectacle is shaped by the same force that holds the note to the refrigerator.

People have been aware of magnets and magnetism for thousands of years. The earliest records date to well before the time of Christ, particularly in a region of Asia Minor called Magnesia (the name of this region is the source of words like *magnetic*). Magnetic rocks found in Magnesia, which is now part of western Turkey, stimulated interest during ancient times. A practical application for magnets was found later, when they were employed as navigational compasses. The use of magnets in compasses resulted not only in improved long-distance sailing, but also in the names of "north" and "south" being given to the two types of magnetic poles.

Today magnetism plays many important roles in our lives. Physicists' understanding of magnetism has enabled the development of technologies that affect our everyday lives. The iPod in your purse or backpack, for example, wouldn't have been possible without the applications of magnetism and electricity on a small scale.

The discovery that weak changes in a magnetic field in a thin film of iron and chromium could bring about much larger changes in electrical resistance was one of the first large successes of nanotechnology. The 2007 Nobel Prize in Physics went to Albert Fert from France and Peter Grunberg from Germany for this discovery of *giant magnetoresistance* and its applications to computer memory.

All electric motors, with uses as diverse as powering refrigerators, starting cars, and moving elevators, contain magnets. Generators, whether producing hydroelectric power or running bicycle lights, use magnetic fields. Recycling facilities employ magnets to separate iron from other refuse. Hundreds of millions of dollars are spent annually on magnetic containment of fusion as a future energy source. Magnetic resonance imaging (MRI) has become an important diagnostic tool in the field of medicine, and the use of magnetism to explore brain activity is a subject of contemporary research and development. The list of applications also includes computer hard drives, tape recording, detection of inhaled asbestos, and levitation of high-speed trains. Magnetism is used to explain atomic energy levels, cosmic rays, and charged particles trapped in the Van Allen belts. Once again, we will find all these disparate phenomena are linked by a small number of underlying physical principles.



Figure 3.2.1.2: Engineering of technology like iPods would not be possible without a deep understanding magnetism. (credit: Jesse! S?, Flickr)

This page titled 3.2.1: Introduction to Magnetism is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.1: Introduction to Magnetism by Andrew Park is licensed CC BY 4.0.

3.2.2: Magnets

Learning Objectives

- Describe the difference between the north and south poles of a magnet.
- Describe how magnetic poles interact with each other.

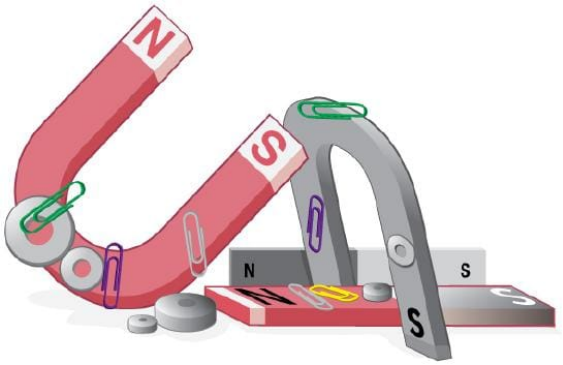


Figure 3.2.2.1: Magnets come in various shapes, sizes, and strengths. All have both a north pole and a south pole. There is never an isolated pole (a monopole).

All magnets attract iron, such as that in a refrigerator door. However, magnets may attract or repel other magnets. Experimentation shows that all magnets have two poles. If freely suspended, one pole will point toward the north. The two poles are thus named the **north magnetic pole** and the **south magnetic pole** (or more properly, north-seeking and south-seeking poles, for the attractions in those directions).

UNIVERSAL CHARACTERISTICS OF MAGNETS AND MAGNETIC POLES

It is a universal characteristic of all magnets that *like poles repel and unlike poles attract*. (Note the similarity with electrostatics: unlike charges attract and like charges repel.)

Further experimentation shows that it is *impossible to separate north and south poles* in the manner that + and – charges can be separated.

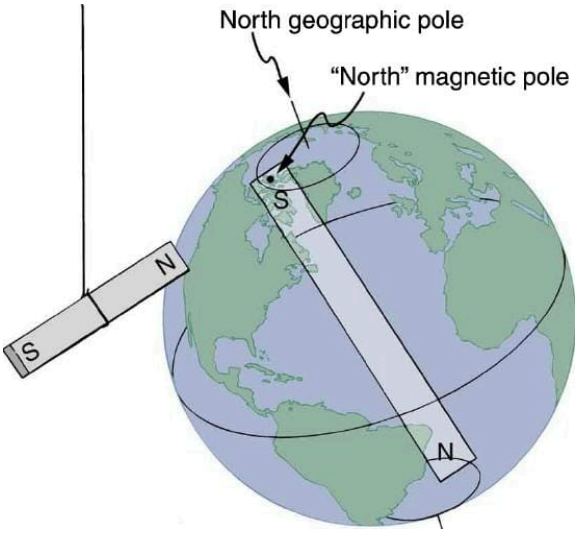


Figure 3.2.2.2: One end of a bar magnet is suspended from a thread that points toward north. The magnet's two poles are labeled N and S for north-seeking and south-seeking poles, respectively.

MISCONCEPTION ALERT: EARTH'S GEOGRAPHIC NORTH POLE HIDES AN S

The Earth acts like a very large bar magnet with its south-seeking pole near the geographic North Pole. That is why the north pole of your compass is attracted toward the geographic north pole of the Earth—because the magnetic pole that is near the geographic North Pole is actually a south magnetic pole! Confusion arises because the geographic term "North Pole" has come to be used (incorrectly) for the magnetic pole that is near the North Pole. Thus, "North magnetic pole" is actually a misnomer—it should be called the South magnetic pole.

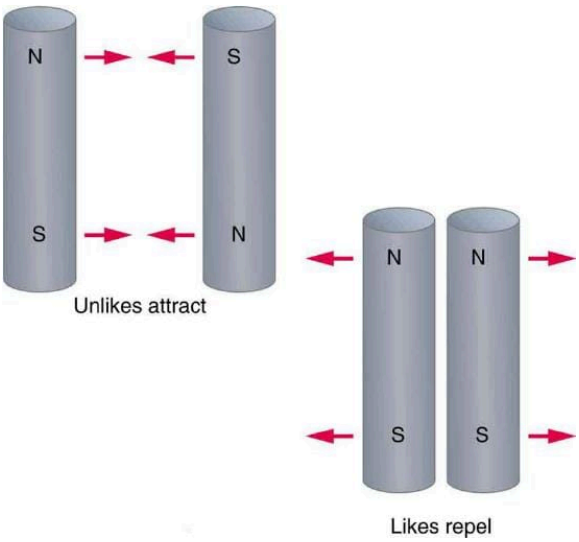


Figure 3.2.2.3: Unlike poles attract, whereas like poles repel.

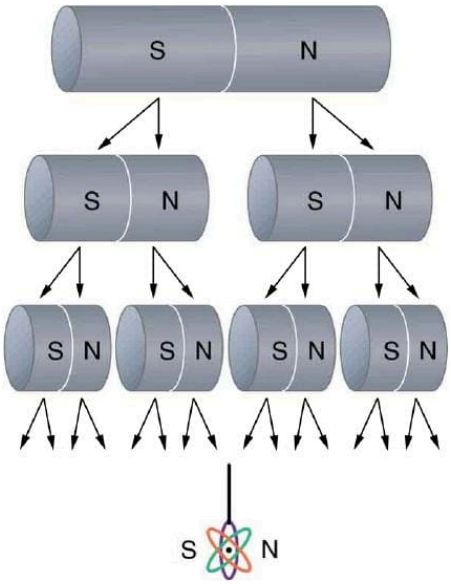


Figure 3.2.2.4: North and south poles always occur in pairs. Attempts to separate them result in more pairs of poles. If we continue to split the magnet, we will eventually get down to an iron atom with a north pole and a south pole—these, too, cannot be separated.

The fact that magnetic poles always occur in pairs of north and south is true from the very large scale—for example, sunspots always occur in pairs that are north and south magnetic poles—all the way down to the very small scale. Magnetic atoms have both a north pole and a south pole, as do many types of subatomic particles, such as electrons, protons, and neutrons.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—REFRIGERATOR MAGNETS

We know that like magnetic poles repel and unlike poles attract. See if you can show this for two refrigerator magnets. Will the magnets stick if you turn them over? Why do they stick to the door anyway? What can you say about the magnetic properties of the door next to the magnet? Do refrigerator magnets stick to metal or plastic spoons? Do they stick to all types of metal?

Section Summary

- Magnetism is a subject that includes the properties of magnets, the effect of the magnetic force on moving charges and currents, and the creation of magnetic fields by currents.
- There are two types of magnetic poles, called the north magnetic pole and south magnetic pole.
- North magnetic poles are those that are attracted toward the Earth's geographic north pole.
- Like poles repel and unlike poles attract.
- Magnetic poles always occur in pairs of north and south—it is not possible to isolate north and south poles.

Glossary

north magnetic pole
: the end or the side of a magnet that is attracted toward Earth's geographic north pole

south magnetic pole
: the end or the side of a magnet that is attracted toward Earth's geographic south pole

This page titled 3.2.2: Magnets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.2: Magnets by Andrew Park is licensed CC BY 4.0.

3.2.3: Ferromagnets and Electromagnets

Learning Objectives

- Define ferromagnet.
- Describe the role of magnetic domains in magnetization.
- Explain the significance of the Curie temperature.
- Describe the relationship between electricity and magnetism.

Ferromagnets

Only certain materials, such as iron, cobalt, nickel, and gadolinium, exhibit strong magnetic effects. Such materials are called **ferromagnetic**, after the Latin word for iron, *ferrum*. A group of materials made from the alloys of the rare earth elements are also used as strong and permanent magnets; a popular one is neodymium. Other materials exhibit weak magnetic effects, which are detectable only with sensitive instruments. Not only do ferromagnetic materials respond strongly to magnets (the way iron is attracted to magnets), they can also be **magnetized** themselves—that is, they can be induced to be magnetic or made into permanent magnets.

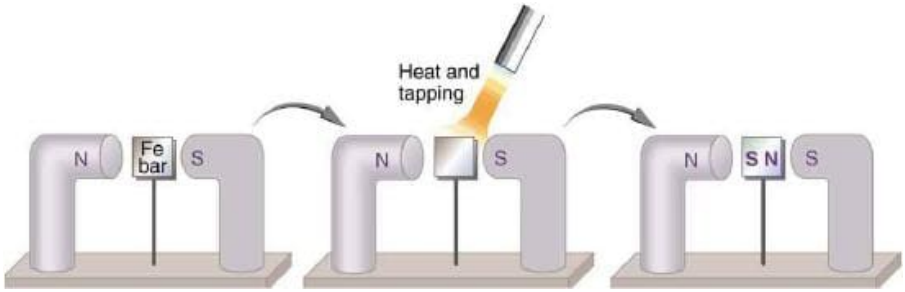


Figure 3.2.3.1: An unmagnetized piece of iron is placed between two magnets, heated, and then cooled, or simply tapped when cold. The iron becomes a permanent magnet with the poles aligned as shown: its south pole is adjacent to the north pole of the original magnet, and its north pole is adjacent to the south pole of the original magnet. Note that there are attractive forces between the magnets.

When a magnet is brought near a previously unmagnetized ferromagnetic material, it causes local magnetization of the material with unlike poles closest, as in Figure 3.2.3.1. (This results in the attraction of the previously unmagnetized material to the magnet.) What happens on a microscopic scale is illustrated in Figure 3.2.3.2. The regions within the material called **domains** act like small bar magnets. Within domains, the poles of individual atoms are aligned. Each atom acts like a tiny bar magnet. Domains are small and randomly oriented in an unmagnetized ferromagnetic object. In response to an external magnetic field, the domains may grow to millimeter size, aligning themselves as shown in Figure 3.2.3.2(b). This induced magnetization can be made permanent if the material is heated and then cooled, or simply tapped in the presence of other magnets.

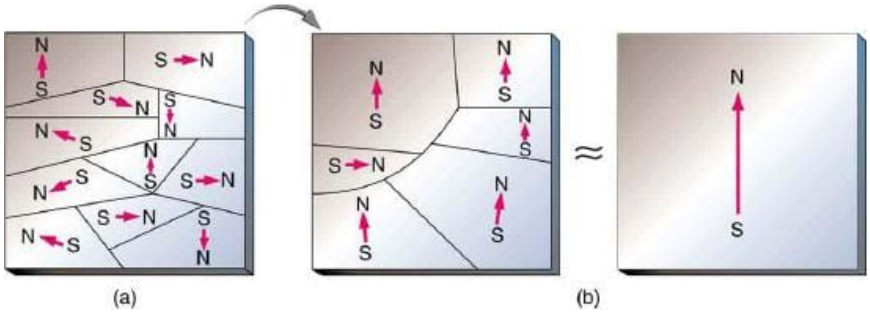


Figure 3.2.3.2: (a) An unmagnetized piece of iron (or other ferromagnetic material) has randomly oriented domains. (b) When magnetized by an external field, the domains show greater alignment, and some grow at the expense of others. Individual atoms are aligned within domains; each atom acts like a tiny bar magnet.

Conversely, a permanent magnet can be demagnetized by hard blows or by heating it in the absence of another magnet. Increased thermal motion at higher temperature can disrupt and randomize the orientation and the size of the domains. There is a well-defined temperature for ferromagnetic materials, which is called the **Curie temperature**, above which they cannot be magnetized. The Curie temperature for iron is 1043 K $(770^{\circ}\mathrm{C})$, which is well above room temperature. There are several elements and alloys that have Curie temperatures much lower than room temperature and are ferromagnetic only below those temperatures.

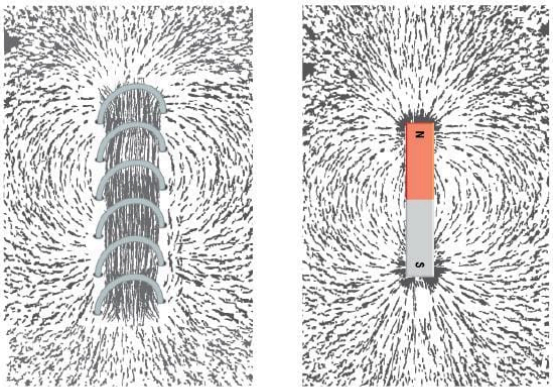
Electromagnets

Early in the 19th century, it was discovered that electrical currents cause magnetic effects. The first significant observation was by the Danish scientist Hans Christian Oersted (1777–1851), who found that a compass needle was deflected by a current-carrying wire. This was the first significant evidence that the movement of charges had any connection with magnets. **Electromagnetism** is the use of electric current to make magnets. These temporarily induced magnets are called **electromagnets**. Electromagnets are employed for everything from a wrecking yard crane that lifts scrapped cars to controlling the beam of a 90-km-circumference particle accelerator to the magnets in medical imaging machines (See Figure 3.2.3.3).



Figure 3.2.3.3: Instrument for magnetic resonance imaging (MRI). The device uses a superconducting cylindrical coil for the main magnetic field. The patient goes into this "tunnel" on the gurney. (credit: Bill McChesney, Flickr)

Figure 3.2.3.4 shows that the response of iron filings to a current-carrying coil and to a permanent bar magnet. The patterns are similar. In fact, electromagnets and ferromagnets have the same basic characteristics—for example, they have north and south poles that cannot be separated and for which like poles repel and unlike poles attract.



(a) (b)

Figure 3.2.3.4: Iron filings near (a) a current-carrying coil and (b) a magnet act like tiny compass needles, showing the shape of their fields. Their response to a current-carrying coil and a permanent magnet is seen to be very similar, especially near the ends of the coil and the magnet.

Combining a ferromagnet with an electromagnet can produce particularly strong magnetic effects. (See Figure 3.2.3.5.) Whenever strong magnetic effects are needed, such as lifting scrap metal, or in particle accelerators, electromagnets are enhanced by

ferromagnetic materials. Limits to how strong the magnets can be made are imposed by coil resistance (it will overheat and melt at sufficiently high current), and so superconducting magnets may be employed. These are still limited, because superconducting properties are destroyed by too great a magnetic field.

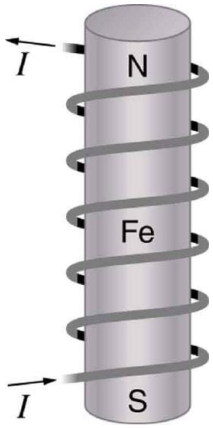


Figure 3.2.3.5: An electromagnet with a ferromagnetic core can produce very strong magnetic effects. Alignment of domains in the core produces a magnet, the poles of which are aligned with the electromagnet.

Figure 3.2.3.6 shows a few uses of combinations of electromagnets and ferromagnets. Ferromagnetic materials can act as memory devices, because the orientation of the magnetic fields of small domains can be reversed or erased. Magnetic information storage on videotapes and computer hard drives are among the most common applications. This property is vital in our digital world.

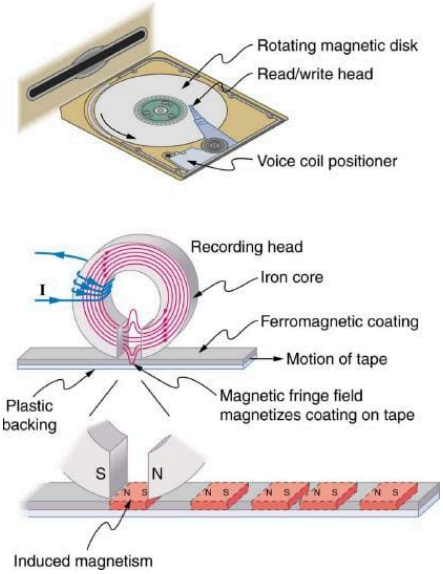


Figure 3.2.3.6: An electromagnet induces regions of permanent magnetism on a floppy disk coated with a ferromagnetic material. The information stored here is digital (a region is either magnetic or not); in other applications, it can be analog (with a varying strength), such as on audiotapes.

Current: The Source of All Magnetism

An electromagnet creates magnetism with an electric current. In later sections we explore this more quantitatively, finding the strength and direction of magnetic fields created by various currents. But what about ferromagnets? Figure 3.2.3.7 shows models of how electric currents create magnetism at the submicroscopic level. (Note that we cannot directly observe the paths of individual electrons about atoms, and so a model or visual image, consistent with all direct observations, is made. We can directly observe the electron's orbital angular momentum, its spin momentum, and subsequent magnetic moments, all of which are explained with electric-current-creating subatomic magnetism.) Currents, including those associated with other submicroscopic particles like protons, allow us to explain ferromagnetism and all other magnetic effects. Ferromagnetism, for example, results

from an internal cooperative alignment of electron spins, possible in some materials but not in others.

Crucial to the statement that electric current is the source of all magnetism is the fact that it is impossible to separate north and south magnetic poles. (This is far different from the case of positive and negative charges, which are easily separated.) A current loop always produces a magnetic dipole—that is, a magnetic field that acts like a north pole and south pole pair. Since isolated north and south magnetic poles, called **magnetic monopoles**, are not observed, currents are used to explain all magnetic effects. If magnetic monopoles did exist, then we would have to modify this underlying connection that all magnetism is due to electrical current. There is no known reason that magnetic monopoles should not exist—they are simply never observed—and so searches at the subnuclear level continue. If they do *not* exist, we would like to find out why not. If they *do* exist, we would like to see evidence of them.

ELECTRIC CURRENTS AND MAGNETISM

Electric current is the source of all magnetism.

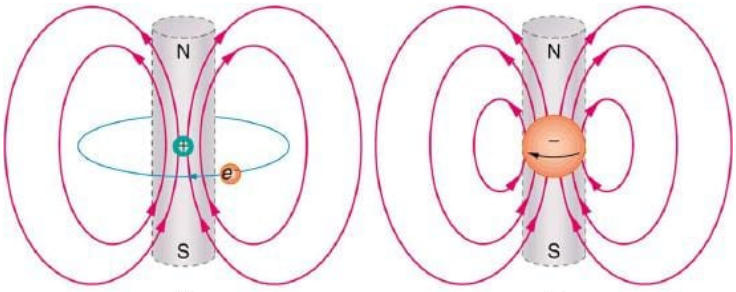


Figure 3.2.3.7: (a) In the planetary model of the atom, an electron orbits a nucleus, forming a closed-current loop and producing a magnetic field with a north pole and a south pole. (b) Electrons have spin and can be crudely pictured as rotating charge, forming a current that produces a magnetic field with a north pole and a south pole. Neither the planetary model nor the image of a spinning electron is completely consistent with modern physics. However, they do provide a useful way of understanding phenomena.

Section Summary

- Magnetic poles always occur in pairs of north and south—it is not possible to isolate north and south poles.
- All magnetism is created by electric current.
- Ferromagnetic materials, such as iron, are those that exhibit strong magnetic effects.
- The atoms in ferromagnetic materials act like small magnets (due to currents within the atoms) and can be aligned, usually in millimeter-sized regions called domains.
- Domains can grow and align on a larger scale, producing permanent magnets. Such a material is magnetized, or induced to be magnetic.
- Above a material's Curie temperature, thermal agitation destroys the alignment of atoms, and ferromagnetism disappears.
- Electromagnets employ electric currents to make magnetic fields, often aided by induced fields in ferromagnetic materials.

Glossary

ferromagnetic
: materials, such as iron, cobalt, nickel, and gadolinium, that exhibit strong magnetic effects

magnetized
: to be turned into a magnet; to be induced to be magnetic

domains
: regions within a material that behave like small bar magnets

Curie temperature
: the temperature above which a ferromagnetic material cannot be magnetized

electromagnetism
: the use of electrical currents to induce magnetism

electromagnet
: an object that is temporarily magnetic when an electrical current is passed through it

magnetic monopoles
: an isolated magnetic pole; a south pole without a north pole, or vice versa (no magnetic monopole has ever been observed)

This page titled 3.2.3: Ferromagnets and Electromagnets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.3: Ferromagnets and Electromagnets by Andrew Park is licensed CC BY 4.0.

3.2.4: Magnetic Fields and Magnetic Field Lines

Learning Objectives

- Define magnetic field and describe the magnetic field lines of various magnetic fields.

Einstein is said to have been fascinated by a compass as a child, perhaps musing on how the needle felt a force without direct physical contact. His ability to think deeply and clearly about action at a distance, particularly for gravitational, electric, and magnetic forces, later enabled him to create his revolutionary theory of relativity. Since magnetic forces act at a distance, we define a **magnetic field** to represent magnetic forces. The pictorial representation of **magnetic field lines** is very useful in visualizing the strength and direction of the magnetic field. As shown in Figure 3.2.4.1, the **direction of magnetic field lines** is defined to be the direction in which the north end of a compass needle points. The magnetic field is traditionally called the ***B*-field**.

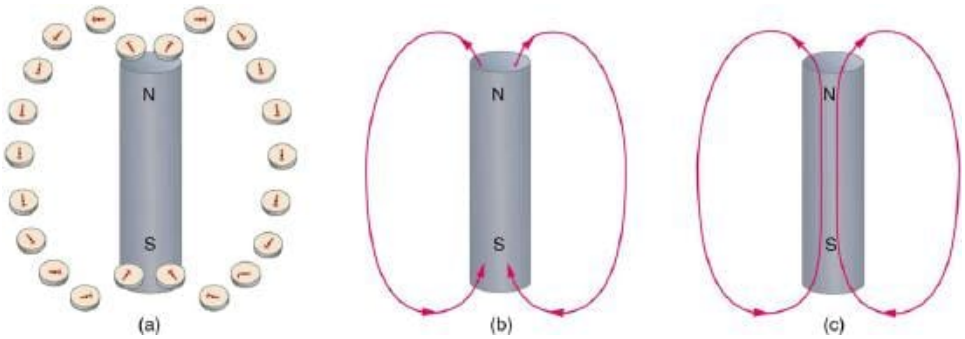


Figure 3.2.4.1: Magnetic field lines are defined to have the direction that a small compass points when placed at a location. (a) If small compasses are used to map the magnetic field around a bar magnet, they will point in the directions shown: away from the north pole of the magnet, toward the south pole of the magnet. (Recall that the Earth's north magnetic pole is really a south pole in terms of definitions of poles on a bar magnet.) (b) Connecting the arrows gives continuous magnetic field lines. The strength of the field is proportional to the closeness (or density) of the lines. (c) If the interior of the magnet could be probed, the field lines would be found to form continuous closed loops.

Small compasses used to test a magnetic field will not disturb it. (This is analogous to the way we tested electric fields with a small test charge. In both cases, the fields represent only the object creating them and not the probe testing them.) Figure 3.2.4.2 shows how the magnetic field appears for a current loop and a long straight wire, as could be explored with small compasses. A small compass placed in these fields will align itself parallel to the field line at its location, with its north pole pointing in the direction of *B*. Note the symbols used for field into and out of the paper.

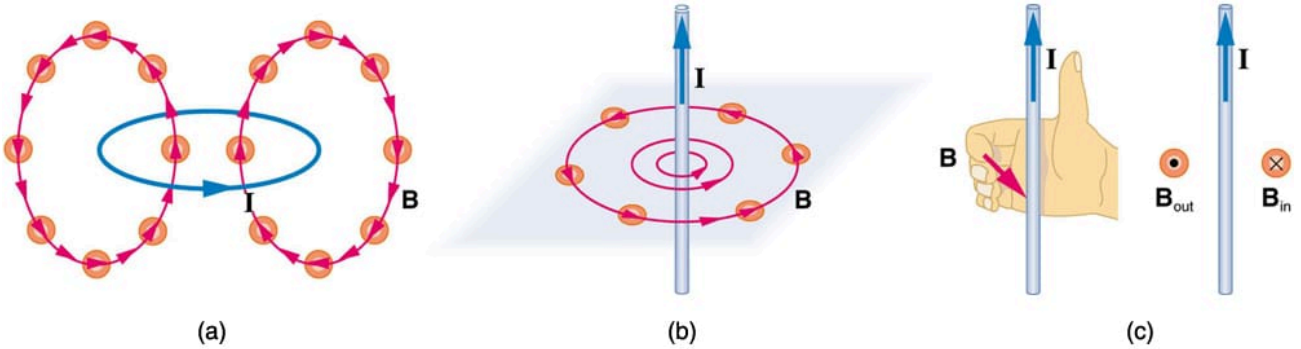


Figure 3.2.4.2: Small compasses could be used to map the fields shown here. (a) The magnetic field of a circular current loop is similar to that of a bar magnet. (b) A long and straight wire creates a field with magnetic field lines forming circular loops. (c) When the wire is in the plane of the paper, the field is perpendicular to the paper. Note that the symbols used for the field pointing inward (like the tail of an arrow) and the field pointing outward (like the tip of an arrow).

MAKING CONNECTIONS: CONCEPT OF A FIELD

A field is a way of mapping forces surrounding any object that can act on another object at a distance without apparent physical connection. The field represents the object generating it. Gravitational fields map gravitational forces, electric fields map electrical forces, and magnetic fields map magnetic forces.

Extensive exploration of magnetic fields has revealed a number of hard-and-fast rules. We use magnetic field lines to represent the field (the lines are a pictorial tool, not a physical entity in and of themselves). The properties of magnetic field lines can be summarized by these rules:

1. The direction of the magnetic field is tangent to the field line at any point in space. A small compass will point in the direction of the field line.
2. The strength of the field is proportional to the closeness of the lines. It is exactly proportional to the number of lines per unit area perpendicular to the lines (called the areal density).
3. Magnetic field lines can never cross, meaning that the field is unique at any point in space.
4. Magnetic field lines are continuous, forming closed loops without beginning or end. They go from the north pole to the south pole.

The last property is related to the fact that the north and south poles cannot be separated. It is a distinct difference from electric field lines, which begin and end on the positive and negative charges. If magnetic monopoles existed, then magnetic field lines would begin and end on them.

Section Summary

- Magnetic fields can be pictorially represented by magnetic field lines, the properties of which are as follows:

1. The field is tangent to the magnetic field line.
2. Field strength is proportional to the line density.
3. Field lines cannot cross.
4. Field lines are continuous loops.

Glossary

magnetic field
: the representation of magnetic forces

***B*-field**
: another term for magnetic field

magnetic field lines
: the pictorial representation of the strength and the direction of a magnetic field

direction of magnetic field lines
: the direction that the north end of a compass needle points

This page titled 3.2.4: Magnetic Fields and Magnetic Field Lines is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.4: Magnetic Fields and Magnetic Field Lines by Andrew Park is licensed CC BY 4.0.

3.2.5: Magnetic Field Strength- Force on a Moving Charge in a Magnetic Field

Learning Objectives

- Describe the effects of magnetic fields on moving charges.
- Use the right hand rule 1 to determine the velocity of a charge, the direction of the magnetic field, and the direction of the magnetic force on a moving charge.
- Calculate the magnetic force on a moving charge.

What is the mechanism by which one magnet exerts a force on another? The answer is related to the fact that all magnetism is caused by current, the flow of charge. *Magnetic fields exert forces on moving charges*, and so they exert forces on other magnets, all of which have moving charges.

Magnetic Force on a Moving Charge

The magnetic force on a moving charge is one of the most fundamental known. Magnetic force is as important as the electrostatic or Coulomb force. Yet the magnetic force is more complex, in both the number of factors that affects it and in its direction, than the relatively simple Coulomb force.

The magnitude of the magnetic force F on a charge depends on: the quantity of charge q, its speed v, the strength of magnetic field B, and **the direction of motion relative to the magnetic field's direction**. Motion, and its direction, are critical.

The **maximum** force occurs when the direction of motion and the magnetic field's direction are *perpendicular* to one another (i.e. ninety degree angle between directions).

$$\mathbf{v} \perp \mathbf{B}$$

In that situation, the magnitude of the magnetic force is

$$F = qvB$$

The **minimum** force occurs when the direction of motion and the magnetic field's direction are *parallel* to one another (i.e. zero or 180 degree angle between directions).

$$\mathbf{v} \| \mathbf{B}$$

In that situation, the magnitude of the magnetic force is

$$F = 0$$

We define the magnetic field strength B in terms of the force on a charged particle moving in a magnetic field. The SI unit for magnetic field strength B is called the **tesla** (T) after the eccentric but brilliant inventor Nikola Tesla (1856–1943). To determine how the tesla relates to other SI units, we solve for the magnetic field strength.

$$B = \frac{F}{qv}$$

So, the tesla is

$$1\ \mathrm{T} = \frac{1\ \mathrm{N}}{\mathrm{C} \cdot \mathrm{m/s}} = \frac{1\ \mathrm{N}}{\mathrm{A} \cdot \mathrm{m}}$$

(note that C/s = A).

Another smaller unit, called the **gauss** (G), where $1\ \mathrm{G} = 10^{-4}\ \mathrm{T}$, is sometimes used. The strongest permanent magnets have fields near 2 T; superconducting electromagnets may attain 10 T or more. The Earth's magnetic field on its surface is only about $5 \times 10^{-5}\ \mathrm{T}$, or 0.5 G.

MAKING CONNECTIONS: CHARGES AND MAGNETS

There is no magnetic force on static charges. However, there is a magnetic force on moving charges. When charges are stationary, their electric fields do not affect magnets. But, when charges move, they produce magnetic fields that exert forces on other magnets. When there is relative motion, a connection between electric and magnetic fields emerges—each affects the other.

Direction of Force: Right Hand Rule 1

The *direction* of the magnetic force $\mathbf{F}$ is perpendicular to the plane formed by $\mathbf{v}$ and $\mathbf{B}$, as determined by the **right hand rule 1** (or RHR-1), which is illustrated in Figure 3.2.5.1. RHR-1 states that, to determine the direction of the magnetic force on a positive moving charge, you point the thumb of the right hand in the direction of $\mathbf{v}$, the fingers in the direction of $\mathbf{B}$, and a perpendicular to the palm points in the direction of $\mathbf{F}$. One way to remember this is that there is one velocity, and so the thumb represents it. There are many field lines, and so the fingers represent them. The force is in the direction you would push with your palm. The force on a negative charge is in exactly the opposite direction to that on a positive charge.

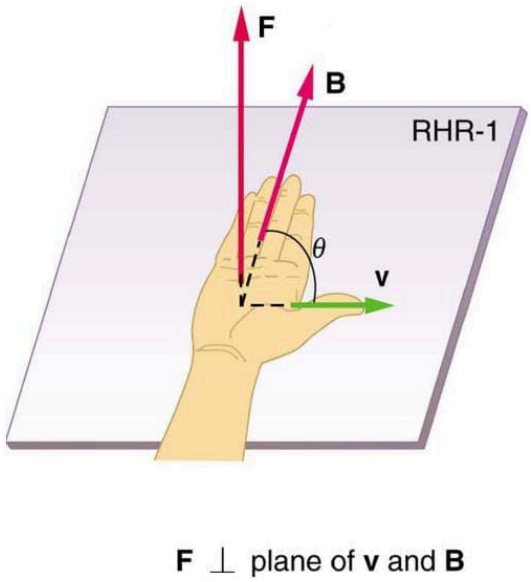


Figure 3.2.5.1: Magnetic fields exert forces on moving charges. This force is one of the most basic known. The direction of the magnetic force on a moving charge is perpendicular to the plane formed by $\mathbf{v}$ and B and follows right hand rule–1 (RHR-1) as shown. The magnitude of the force is proportional to q, v, B, and depends on the angle between $\mathbf{v}$ and B.

Section Summary

- The maximum force a magnetic field can exert on a moving charge is

$$F = qvB$$

- The SI unit for magnetic field strength B is the tesla (T), which is related to other units by

$$1\ \mathrm{T} = \frac{1\ \mathrm{N}}{\mathrm{C} \cdot \mathrm{m/s}} = \frac{1\ \mathrm{N}}{\mathrm{A} \cdot \mathrm{m}}.$$

- The *direction* of the force on a moving charge is given by right hand rule 1 (RHR-1): Point the thumb of the right hand in the direction of v, the fingers in the direction of B, and a perpendicular to the palm points in the direction of F.

- The force is perpendicular to the plane formed by $\mathbf{V}$ and B. Since the force is zero if $\mathbf{V}$ is parallel to B, charged particles often follow magnetic field lines rather than cross them.

Glossary

right hand rule 1 (RHR-1)
: the rule to determine the direction of the magnetic force on a positive moving charge: when the thumb of the right hand points in the direction of the charge's velocity v and the fingers point in the direction of the magnetic field $\mathbf{B}$, then the force on the charge is perpendicular and away from the palm; the force on a negative charge is perpendicular and into the palm

tesla
: T, the SI unit of the magnetic field strength; $1\ \mathrm{T} = \frac{1\ \mathrm{N}}{\mathrm{A \cdot m}}$

magnetic force
: the force on a charge produced by its motion through a magnetic field

gauss
: G, the unit of the magnetic field strength; $1\mathrm{G} = 10^{-4}\ \mathrm{T}$

This page titled 3.2.5: Magnetic Field Strength- Force on a Moving Charge in a Magnetic Field is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.5: Magnetic Field Strength- Force on a Moving Charge in a Magnetic Field by Andrew Park is licensed CC BY 4.0.

3.2.6: Magnetic Force on a Current-Carrying Conductor

Learning Objectives

- Describe the effects of a magnetic force on a current-carrying conductor.
- Calculate the magnetic force on a current-carrying conductor.

Because charges ordinarily cannot escape a conductor, the magnetic force on charges moving in a conductor is transmitted to the conductor itself.

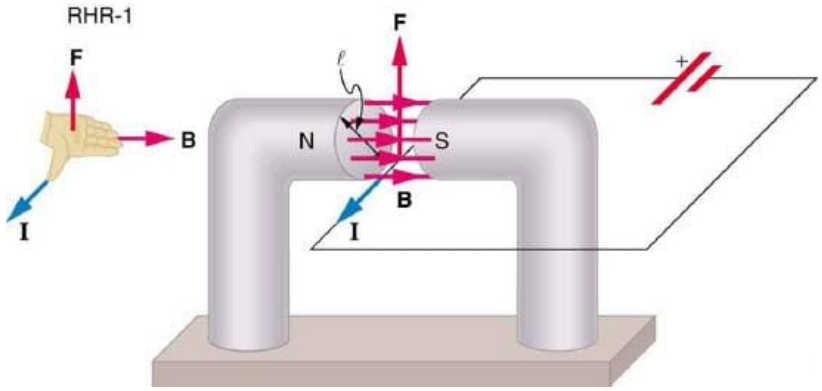


Figure 3.2.6.1: The magnetic field exerts a force on a current-carrying wire in a direction given by the right hand rule 1 (the same direction as that on the individual moving charges). This force can easily be large enough to move the wire, since typical currents consist of very large numbers of moving charges.

The maximum force on a current-carrying conductor occurs when the current direction and the magnetic field's direction are perpendicular to one another (i.e. ninety degree angle between directions). We can derive an expression for the maximum magnetic force on a current by taking a sum of the magnetic forces on individual charges. (The forces add because they are in the same direction.) The force on an individual charge moving at the drift velocity v_d is given by $F = qv_dB$. Taking B to be uniform over a length of wire l and zero elsewhere, the total magnetic force on the wire is then $F = (qv_dB)\,(N)$, where N is the number of charge carriers in the section of wire of length l. Now, $N = nV$, where n is the number of charge carriers per unit volume and V is the volume of wire in the field. Noting that $V = Al$, where A is the cross-sectional area of the wire, then the force on the wire is $F = (qv_dB)\,(nAl)$. Gathering terms,

$$F = (nqAv_d)\,(lB).$$

Because $nqAv_{\mathrm{d}} = I$,

$$F = IlB$$

is the equation for *maximum magnetic force on a length l of wire carrying a current I in a uniform magnetic field B*, as shown in Figure 3.2.6.2. If we divide both sides of this expression by l, we find that the magnetic force per unit length of wire in a uniform field is $\frac{F}{l} = IB$. The direction of this force is given by RHR-1, with the thumb in the direction of the current I. Then, with the fingers in the direction of B, a perpendicular to the palm points in the direction of F, as in Figure 3.2.6.2.

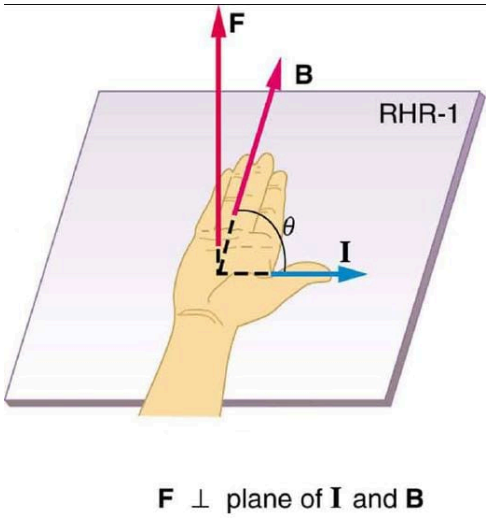


Figure 3.2.6.2: The force on a current-carrying wire in a magnetic field is $F = IlB$. Its direction is given by RHR-1.

Example 3.2.6.1: Calculating Magnetic Force on a Current-Carrying Wire: A Strong Magnetic Field

Calculate the force on the wire shown in Figure 3.2.6.1, given $B = 1.50\ \text{T}$, $l = 5.00\ \text{cm}$, and $I = 20.0\ \text{A}$.

Strategy

The force can be found with the given information by using $F = IlB$ because the angle between I and B is 90°.

Solution

Entering the given values into $F = IlB$ yields

$$F = IlB = (20.0\ \text{A})(0.0500\ \text{m})(1.50\ \text{T}).$$

The units for tesla are $1\ \text{T} = \frac{\text{N}}{\text{A}\cdot\text{m}}$; thus,

$$F = 1.50\ \text{N}.$$

Discussion

This large magnetic field creates a significant force on a small length of wire.

Magnetic force on current-carrying conductors is used to convert electric energy to work. (Motors are a prime example—they employ loops of wire and are considered in the next section.) Magnetohydrodynamics (MHD) is the technical name given to a clever application where magnetic force pumps fluids without moving mechanical parts. (See Figure 3.2.6.3.)

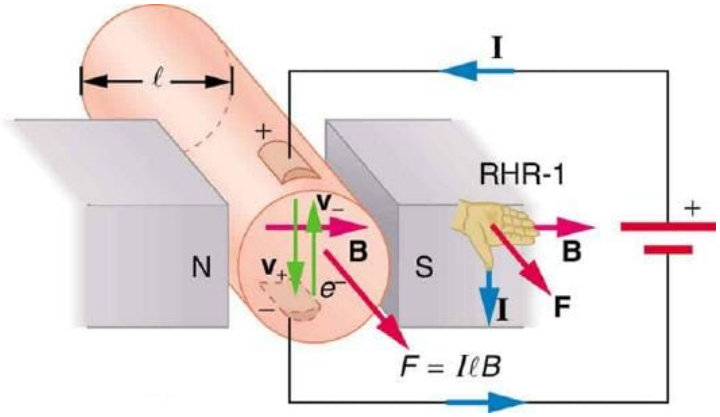


Figure 3.2.6.3: Magnetohydrodynamics. The magnetic force on the current passed through this fluid can be used as a nonmechanical pump.

A strong magnetic field is applied across a tube and a current is passed through the fluid at right angles to the field, resulting in a force on the fluid parallel to the tube axis as shown. The absence of moving parts makes this attractive for moving a hot, chemically active substance, such as the liquid sodium employed in some nuclear reactors. Experimental artificial hearts are testing with this technique for pumping blood, perhaps circumventing the adverse effects of mechanical pumps. (Cell membranes, however, are affected by the large fields needed in MHD, delaying its practical application in humans.) MHD propulsion for nuclear submarines has been proposed, because it could be considerably quieter than conventional propeller drives. The deterrent value of nuclear submarines is based on their ability to hide and survive a first or second nuclear strike. As we slowly disassemble our nuclear weapons arsenals, the submarine branch will be the last to be decommissioned because of this ability (See Figure 3.2.6.4 .) Existing MHD drives are heavy and inefficient—much development work is needed.

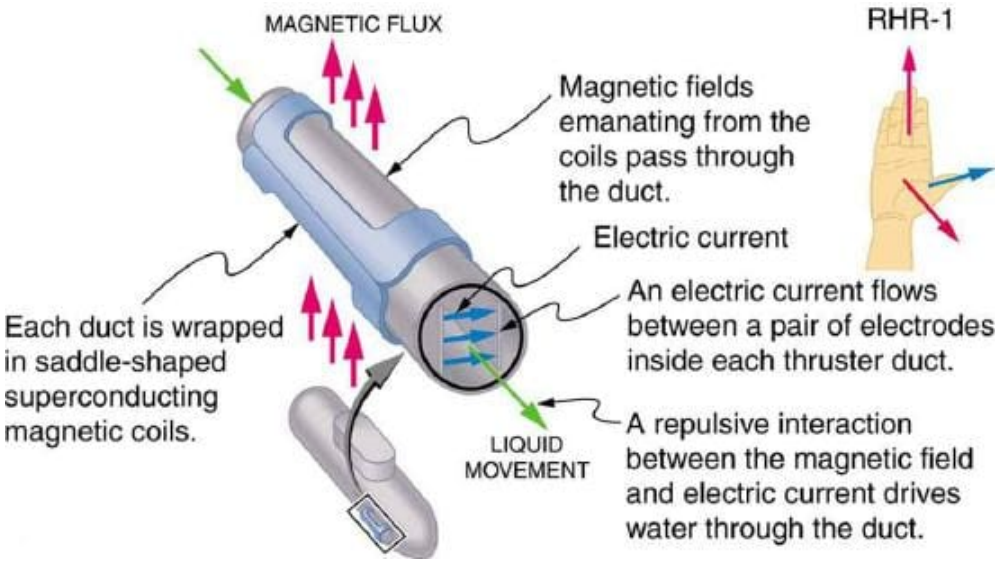


Figure 3.2.6.4: An MHD propulsion system in a nuclear submarine could produce significantly less turbulence than propellers and allow it to run more silently. The development of a silent drive submarine was dramatized in the book and the film *The Hunt for Red October*.

Section Summary

- The magnetic force on current-carrying conductors (when current direction and magnetic field direction are perpendicular) is given by

$$F = IlB,$$

where I is the current, l is the length of a straight conductor in a uniform magnetic field B, and $I \perp B$. The force follows RHR-1 with the thumb in the direction of I.

This page titled 3.2.6: Magnetic Force on a Current-Carrying Conductor is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.6: Magnetic Force on a Current-Carrying Conductor by Andrew Park is licensed CC BY 4.0.

3.2.7: Motors and Meters

Learning Objectives

- Describe how motors and meters work in terms of force on a current loop.

Motors are the most common application of magnetic force on current-carrying wires. Motors have loops of wire in a magnetic field. When current is passed through the loops, the magnetic field exerts force on the loops, which rotates a shaft. Electrical energy is converted to mechanical work in the process. (See Figure 3.2.7.1.)

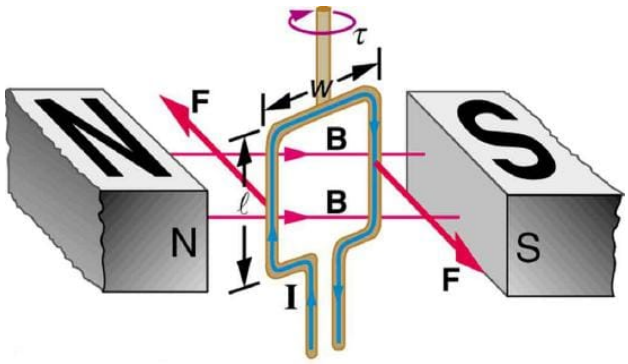


Figure 3.2.7.1: Force on a current loop. A current-carrying loop of wire attached to a vertically rotating shaft feels magnetic forces that produce a clockwise rotation as viewed from above.

As the coil rotates, the force decreases to zero at $\theta = 0$. The force then *reverses* its direction once the coil rotates past $\theta = 0$. This means that, unless we do something, the coil will oscillate back and forth about equilibrium at $\theta = 0$. To get the coil to continue rotating in the same direction, we can reverse the current as it passes through $\theta = 0$ with automatic switches called *brushes*. (See Figure 3.2.7.2.)

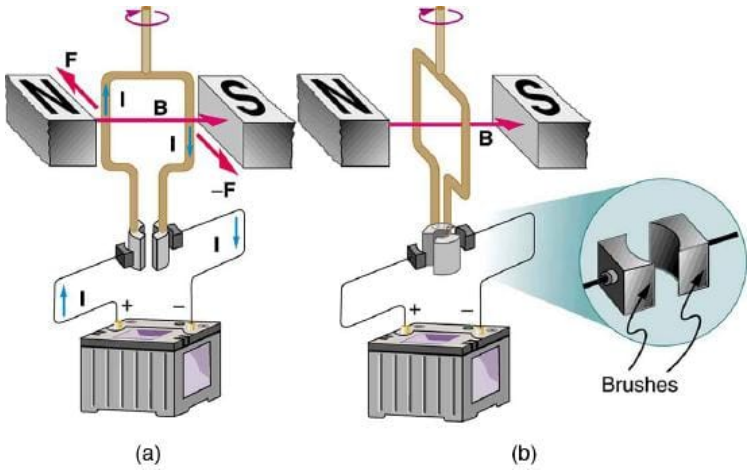


Figure 3.2.7.2: (a) As the momentum of the coil carries it through $\theta = 0$, the brushes reverse the current to keep the motion clockwise. (b) The coil will rotate continuously in the clockwise direction, with the current reversing each half revolution to maintain the motion.

Meters, such as those in analog fuel gauges on a car, are another common application of magnetic force on a current-carrying loop. Figure 3.2.7.3 shows that a meter is very similar in construction to a motor. The meter in the figure has its magnets shaped to limit the effect of θ by making B perpendicular to the loop over a large angular range. A linear spring exerts a counter-force that balances the current-produced force. This makes the needle deflection proportional to I. If an exact proportionality cannot be achieved, the gauge reading can be calibrated. To produce a galvanometer for use in analog voltmeters and ammeters that

have a low resistance and respond to small currents, we use a large loop area A, high magnetic field B, and low-resistance coils.

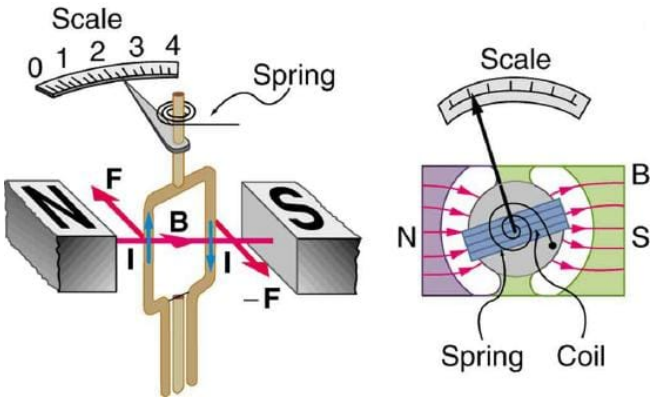


Figure 3.2.7.3: Meters are very similar to motors but only rotate through a part of a revolution. The magnetic poles of this meter are shaped to keep the component of B perpendicular to the loop constant, so that the force does not depend on θ and the deflection against the return spring is proportional only to the current I.

Glossary

motor
: loop of wire in a magnetic field; when current is passed through the loops, the magnetic field exerts force on the loops, which rotates a shaft; electrical energy is converted to mechanical work in the process

meter
: common application of magnetic force on a current-carrying loop that is very similar in construction to a motor; by design, the force is proportional to I and not θ, so the needle deflection is proportional to the current

This page titled 3.2.7: Motors and Meters is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.7: Motors and Meters by Andrew Park is licensed CC BY 4.0.

3.2.8: Magnetic Fields Produced by Currents- Ampere's Law

Learning Objectives

- Calculate current that produces a magnetic field.
- Use the right hand rule 2 to determine the direction of current or the direction of magnetic field loops.

How much current is needed to produce a significant magnetic field, perhaps as strong as the Earth's field? Surveyors will tell you that overhead electric power lines create magnetic fields that interfere with their compass readings. Indeed, when Oersted discovered in 1820 that a current in a wire affected a compass needle, he was not dealing with extremely large currents. How does the shape of wires carrying current affect the shape of the magnetic field created? We noted earlier that a current loop created a magnetic field similar to that of a bar magnet, but what about a straight wire or a toroid (doughnut)? How is the direction of a current-created field related to the direction of the current? Answers to these questions are explored in this section, together with a brief discussion of the law governing the fields created by currents.

Magnetic Field Created by a Long Straight Current-Carrying Wire: Right Hand Rule 2

Magnetic fields have both direction and magnitude. As noted before, one way to explore the direction of a magnetic field is with compasses, as shown for a long straight current-carrying wire in Figure 3.2.8.1. Hall probes can determine the magnitude of the field. The field around a long straight wire is found to be in circular loops. The **right hand rule 2** (RHR-2) emerges from this exploration and is valid for any current segment—*point the thumb in the direction of the current, and the fingers curl in the direction of the magnetic field loops* created by it.

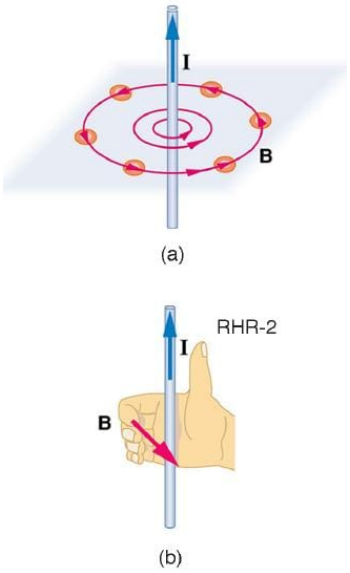


Figure 3.2.8.1: (a) Compasses placed near a long straight current-carrying wire indicate that field lines form circular loops centered on the wire. (b) Right hand rule 2 states that, if the right hand thumb points in the direction of the current, the fingers curl in the direction of the field. This rule is consistent with the field mapped for the long straight wire and is valid for any current segment.

The **magnetic field strength (magnitude) produced by a long straight current-carrying wire** is found by experiment to be

$$B = \frac{\mu_0 I}{2\pi r} \text{ (long straight wire)},$$

where I is the current, r is the shortest distance to the wire, and the constant $\mu_0 = 4\pi \times 10^{-7}\ \text{T} \cdot \text{m/A}$ is the **permeability of free space**. (μ_0 is one of the basic constants in nature. We will see later that μ_0 is related to the speed of light.) Since the wire is very

long, the magnitude of the field depends only on distance from the wire r, not on position along the wire.

Example 3.2.8.1: Calculating Current that Produces a Magnetic Field

Find the current in a long straight wire that would produce a magnetic field twice the strength of the Earth's at a distance of 5.0 cm from the wire.

Strategy

The Earth's field is about 5.0×10^{-5} T, and so here B due to the wire is taken to be 1.0×10^{-4} T. The equation $B = \frac{\mu_0 I}{2\pi r}$ can be used to find I, since all other quantities are known.

Solution

Solving for I and entering known values gives

$$I = \frac{2\pi r B}{\mu_0} = \frac{2\pi \left(5.0 \times 10^{-2} \text{ m}\right)\left(1.0 \times 10^{-4} \text{ T}\right)}{4\pi \times 10^{-7} \text{ T} \cdot \text{m/A}}$$
$$= 25 \text{ A}.$$

Discussion

So a moderately large current produces a significant magnetic field at a distance of 5.0 cm from a long straight wire. Note that the answer is stated to only two digits, since the Earth's field is specified to only two digits in this example.

Ampere's Law and Others

The magnetic field of a long straight wire has more implications than you might at first suspect. *Each segment of current produces a magnetic field like that of a long straight wire, and the total field of any shape current is the vector sum of the fields due to each segment*. The formal statement of the direction and magnitude of the field due to each segment is called the **Biot-Savart law**. Integral calculus is needed to sum the field for an arbitrary shape current. This results in a more complete law, called **Ampere's law**, which relates magnetic field and current in a general way. Ampere's law in turn is a part of **Maxwell's equations**, which give a complete theory of all electromagnetic phenomena. Considerations of how Maxwell's equations appear to different observers led to the modern theory of relativity, and the realization that electric and magnetic fields are different manifestations of the same thing. Most of this is beyond the scope of this text in both mathematical level, requiring calculus, and in the amount of space that can be devoted to it. But for the interested student, and particularly for those who continue in physics, engineering, or similar pursuits, delving into these matters further will reveal descriptions of nature that are elegant as well as profound. In this text, we shall keep the general features in mind, such as RHR-2 and the rules for magnetic field lines listed in Magnetic Fields and Magnetic Field Lines, while concentrating on the fields created in certain important situations.

Magnetic Field Produced by a Current-Carrying Circular Loop

The magnetic field near a current-carrying loop of wire is shown in Figure 3.2.8.2. Both the direction and the magnitude of the magnetic field produced by a current-carrying loop are complex. RHR-2 can be used to give the direction of the field near the loop, but mapping with compasses and the rules about field lines given in "Magnetic Fields and Magnetic Field Lines" are needed for more detail. There is a simple formula for the **magnetic field strength at the center of a circular loop**. It is

$$B = \frac{\mu_0 I}{2R} \text{ (at center of loop)},$$

where R is the radius of the loop. This equation is very similar to that for a straight wire, but it is valid *only* at the center of a circular loop of wire. The similarity of the equations does indicate that similar field strength can be obtained at the center of a loop. One way to get a larger field is to have N loops; then, the field is $B = N\mu_0 I/(2R)$. Note that the larger the loop, the smaller the field at its center, because the current is farther away.

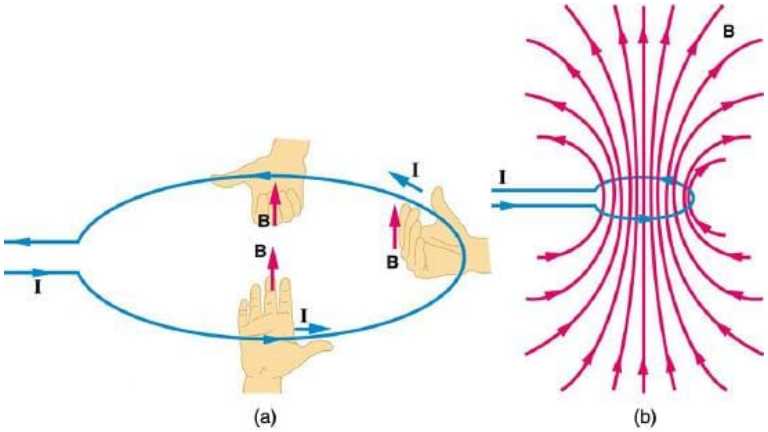


Figure 3.2.8.2: (a) RHR-2 gives the direction of the magnetic field inside and outside a current-carrying loop. (b) More detailed mapping with compasses or with a Hall probe completes the picture. The field is similar to that of a bar magnet.

Magnetic Field Produced by a Current-Carrying Solenoid

A **solenoid** is a long coil of wire (with many turns or loops, as opposed to a flat loop). Because of its shape, the field inside a solenoid can be very uniform, and also very strong. The field just outside the coils is nearly zero. Figure 3.2.8.3 shows how the field looks and how its direction is given by RHR-2.

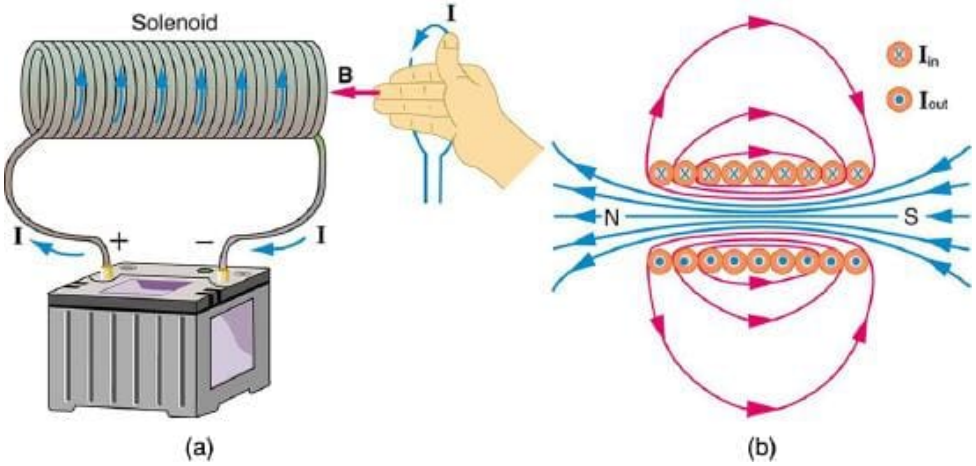


Figure 3.2.8.3: (a) Because of its shape, the field inside a solenoid of length l is remarkably uniform in magnitude and direction, as indicated by the straight and uniformly spaced field lines. The field outside the coils is nearly zero. (b) This cutaway shows the magnetic field generated by the current in the solenoid.

The magnetic field inside of a current-carrying solenoid is very uniform in direction and magnitude. Only near the ends does it begin to weaken and change direction. The field outside has similar complexities to flat loops and bar magnets, but the **magnetic field strength inside a solenoid** is simply

$$B = \mu_0 nI \text{ (inside a solenoid)},$$

where n is the number of loops per unit length of the solenoid ($n = N/l$, with N being the number of loops and l the length). Note that B is the field strength anywhere in the uniform region of the interior and not just at the center. Large uniform fields spread over a large volume are possible with solenoids, as Example 3.2.8.2 implies.

Example 3.2.8.2: Calculating Field Strength inside a Solenoid

What is the field inside a 2.00-m-long solenoid that has 2000 loops and carries a 1600-A current?

Strategy

To find the field strength inside a solenoid, we use $B = \mu_0 nI$. First, we note the number of loops per unit length is

$$n^{-1} = \frac{N}{l} = \frac{2000}{2.00 \text{ m}} = 1000 \text{ m}^{-1} = 10 \text{ cm}^{-1}.$$

Solution

Substituting known values gives

$$\begin{aligned} B = \mu_0 nI &= \left(4\pi \times 10^{-7} \text{ T} \cdot \text{m/A}\right)\left(1000 \text{ m}^{-1}\right)(1600 \text{ A}) \\ &= 2.01 \text{ T}. \end{aligned}$$

Discussion

This is a large field strength that could be established over a large-diameter solenoid, such as in medical uses of magnetic resonance imaging (MRI). The very large current is an indication that the fields of this strength are not easily achieved, however. Such a large current through 1000 loops squeezed into a meter's length would produce significant heating. Higher currents can be achieved by using superconducting wires, although this is expensive. There is an upper limit to the current, since the superconducting state is disrupted by very large magnetic fields.

There are interesting variations of the flat coil and solenoid. For example, the toroidal coil used to confine the reactive particles in tokamaks is much like a solenoid bent into a circle. The field inside a toroid is very strong but circular. Charged particles travel in circles, following the field lines, and collide with one another, perhaps inducing fusion. But the charged particles do not cross field lines and escape the toroid. A whole range of coil shapes are used to produce all sorts of magnetic field shapes. Adding ferromagnetic materials produces greater field strengths and can have a significant effect on the shape of the field. Ferromagnetic materials tend to trap magnetic fields (the field lines bend into the ferromagnetic material, leaving weaker fields outside it) and are used as shields for devices that are adversely affected by magnetic fields, including the Earth's magnetic field.

Section Summary

- The strength of the magnetic field created by current in a long straight wire is given by

$$B = \frac{\mu_0 I}{2\pi r}(\text{ long straight wire }),$$

  where I is the current, r is the shortest distance to the wire, and the constant $\mu_0 = 4\pi \times 10^{-7} \text{ T} \cdot \text{m/A}$ is the permeability of free space.
- The direction of the magnetic field created by a long straight wire is given by right hand rule 2 (RHR-2): *Point the thumb of the right hand in the direction of current, and the fingers curl in the direction of the magnetic field loops* created by it.
- The magnetic field created by current following any path is the sum (or integral) of the fields due to segments along the path (magnitude and direction as for a straight wire), resulting in a general relationship between current and field known as Ampere's law.
- The magnetic field strength at the center of a circular loop is given by

$$B = \frac{\mu_0 I}{2R}(\text{ at center of loop }),$$

  where R is the radius of the loop. This equation becomes $B = \mu_0 nI/(2R)$ for a flat coil of N loops. RHR-2 gives the direction of the field about the loop. A long coil is called a solenoid.
- The magnetic field strength inside a solenoid is

$$B = \mu_0 nI \text{ (inside a solenoid)},$$

  where N is the number of loops per unit length of the solenoid. The field inside is very uniform in magnitude and direction.

Glossary

right hand rule 2 (RHR-2)
: a rule to determine the direction of the magnetic field induced by a current-carrying wire: Point the thumb of the right hand in the direction of current, and the fingers curl in the direction of the magnetic field loops

magnetic field strength (magnitude) produced by a long straight current-carrying wire
: defined as $B = \frac{\mu_0 I}{2\pi r}$, where I is the current, r is the shortest distance to the wire, and μ_0 is the permeability of free space

permeability of free space
: the measure of the ability of a material, in this case free space, to support a magnetic field; the constant $\mu_0 = 4\pi \times 10^{-7}\ \mathrm{T \cdot m/A}$

magnetic field strength at the center of a circular loop
: defined as $B = \frac{\mu_0 I}{2R}$ where R is the radius of the loop

solenoid
: a thin wire wound into a coil that produces a magnetic field when an electric current is passed through it

magnetic field strength inside a solenoid
: defined as $B = \mu_0 n I$ where n is the number of loops per unit length of the solenoid ($n = N/l$, with N being the number of loops and l the length)

Biot-Savart law
: a physical law that describes the magnetic field generated by an electric current in terms of a specific equation

Ampere's law
: the physical law that states that the magnetic field around an electric current is proportional to the current; each segment of current produces a magnetic field like that of a long straight wire, and the total field of any shape current is the vector sum of the fields due to each segment

Maxwell's equations
: a set of four equations that describe electromagnetic phenomena

This page titled 3.2.8: Magnetic Fields Produced by Currents- Ampere's Law is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.8: Magnetic Fields Produced by Currents- Ampere's Law by Andrew Park is licensed CC BY 4.0.

3.2.9: Induced Voltage and Magnetic Flux

Learning Objectives

- Describe methods to produce an induced voltage with a magnetic field or magnet and a loop of wire.

The apparatus used by Faraday to demonstrate that magnetic fields can create currents is illustrated in Figure 3.2.9.1. When the switch is closed, a magnetic field is produced in the coil on the top part of the iron ring and transmitted to the coil on the bottom part of the ring. The galvanometer is used to detect any current induced in the coil on the bottom. It was found that each time the switch is closed, the galvanometer detects a current in one direction in the coil on the bottom. Each time the switch is opened, the galvanometer detects a current in the opposite direction. Interestingly, if the switch remains closed or open for any length of time, there is no current through the galvanometer. *Closing and opening the switch* induces the current. It is the *change* in magnetic field that creates the current in the following way: *a changing magnetic field induces an electric field*, which results in the induced voltage. When this induced voltage occurs over a conducting path, as in this example, the induced voltage causes a current to flow. As a shorthand, we call the resulting current **induced current**; the changing magnetic field does not induce the current directly but through the induced voltage and an application of Ohm's law.

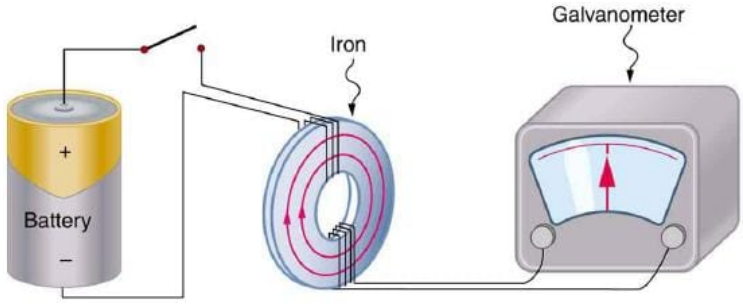


Figure 3.2.9.1: Faraday's apparatus for demonstrating that a magnetic field can produce a current. A change in the field produced by the top coil induces an voltage and, hence, a current in the bottom coil. When the switch is opened and closed, the galvanometer registers currents in opposite directions. No current flows through the galvanometer when the switch remains closed or open.

An experiment easily performed and often done in physics labs is illustrated in Figure 3.2.9.2. A voltage is induced in the coil when a bar magnet is pushed in and out of it. Voltages of opposite signs are produced by motion in opposite directions, and the voltages are also reversed by reversing poles. The same results are produced if the coil is moved rather than the magnet—it is the relative motion that is important. The faster the motion, the greater the voltage, and there is no voltage when the magnet is stationary relative to the coil.

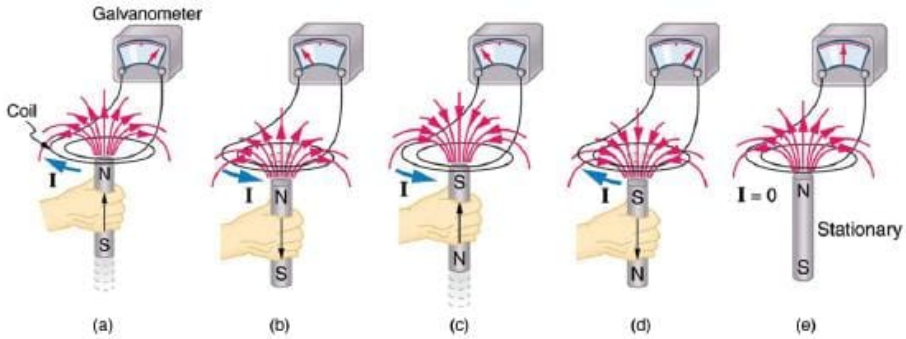


Figure 3.2.9.2: Movement of a magnet relative to a coil produces voltage as shown. The same voltages are produced if the coil is moved relative to the magnet. The greater the speed, the greater the magnitude of the voltage, and the voltage is zero when there is no motion.

The method of inducing a voltage used in most electric generators is shown in Figure 3.2.9.3. A coil is rotated in a magnetic field, producing an alternating voltage (and current), which depends on rotation rate and other factors that will be explored in later sections. Note that the generator is remarkably similar in construction to a motor.

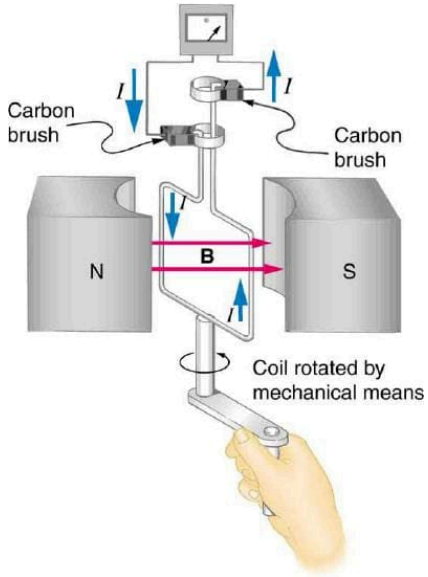


Figure 3.2.9.3: Rotation of a coil in a magnetic field produces a voltage. This is the basic construction of a generator, where work done to turn the coil is converted to electric energy. Note the generator is very similar in construction to a motor.

So we see that changing the magnitude or direction of a magnetic field produces a voltage. Experiments revealed that there is a crucial quantity called the **magnetic flux**, Φ , given by

$$\Phi = B_{\perp}A,$$

where B is the magnetic field strength over an area A, at an angle θ with the perpendicular to the area as shown in Figure 3.2.9.4. **Any change in magnetic flux Φ induces a voltage.** This process is defined to be **electromagnetic induction**. Units of magnetic flux Φ are $\mathrm{T \cdot m^2}$.

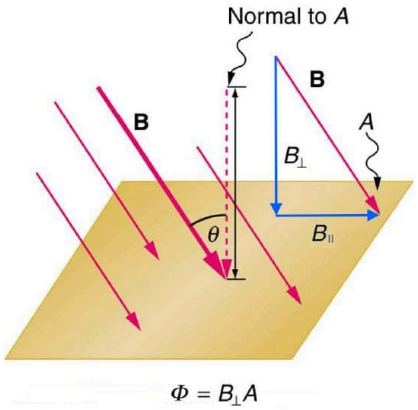


Figure 3.2.9.4: Magnetic flux Φ is related to the magnetic field and the area A over which it exists. Only the portion of the magnetic field that is perpendicular to the area ($B_{\perp}$) contributes to the flux. The flux $\Phi = B_{\perp}A$ is related to induction; any change in Φ induces a voltage.

All induction, including the examples given so far, arises from some change in magnetic flux Φ. For example, Faraday changed B and hence Φ when opening and closing the switch in his apparatus (shown in Figure 3.2.9.1). This is also true for the bar magnet and coil shown in Figure 3.2.9.2. When rotating the coil of a generator, the angle θ and, hence, Φ is changed. Just how great a voltage and what direction it takes depend on the change in Φ and how rapidly the change is made, as examined in the next section.

Section Summary

- The crucial quantity in induction is magnetic flux Φ, defined to be $\Phi = B_{\perp}A$, where $B_{\perp}$ is the magnetic field strength perpendicular to the area A.
- Units of magnetic flux Φ are $\mathrm{T \cdot m^2}$.

- Any change in magnetic flux Φ induces a voltage—the process is defined to be electromagnetic induction.

Glossary

induced current
: the current created by a changing magnetic field through voltage induced over a conducting path

magnetic flux
: the amount of magnetic field going through a particular area, calculated with $\Phi = B_{\perp}A$, where $B_{\perp}$ is the magnetic field strength perpendicular to the area A

electromagnetic induction
: the process of inducing a voltage with a change in magnetic flux

This page titled 3.2.9: Induced Voltage and Magnetic Flux is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.9: Induced Voltage and Magnetic Flux by Andrew Park is licensed CC BY 4.0.

3.2.10: Faraday's Law of Induction- Lenz's Law

Learning Objectives

- Calculate voltage, current, and magnetic fields using Faraday's Law.
- Explain the physical results of Lenz's Law

Faraday's and Lenz's Law

Faraday's experiments showed that the voltage induced by a change in magnetic flux depends on only a few factors. First, voltage is directly proportional to the change in flux $\Delta\Phi$. Second, voltage is greatest when the change in time Δt is smallest—that is, voltage is inversely proportional to Δt. Finally, if a coil has N turns, a voltage will be produced that is N times greater than for a single coil, so that voltage is directly proportional to N. The equation for the voltage induced by a change in magnetic flux is

$$V = -N\frac{\Delta\Phi}{\Delta t}.$$

This relationship is known as **Faraday's law of induction**.

The minus sign in Faraday's law of induction is very important. The minus means that *the induced voltage creates a current I and magnetic field B that oppose the change in flux $\Delta\Phi$—this is known as Lenz's law*. Faraday was aware of the direction, but Lenz stated it so clearly that he is credited for its discovery. (See Figure 3.2.10.1.)

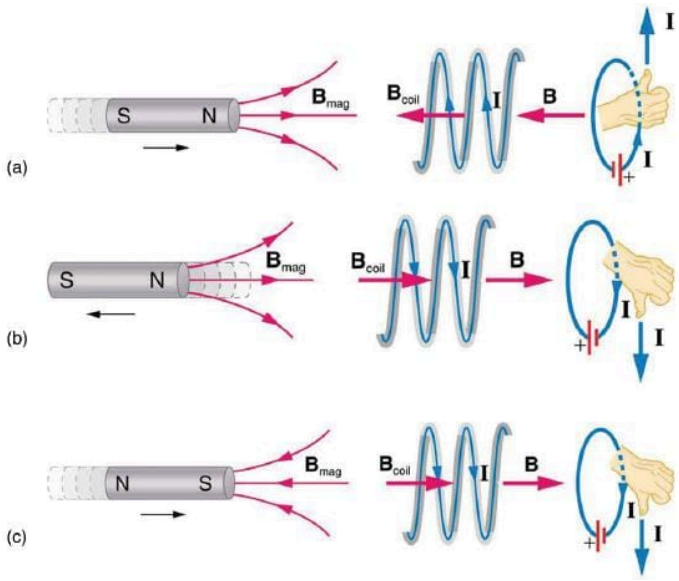


Figure 3.2.10.1: (a) When this bar magnet is thrust into the coil, the strength of the magnetic field increases in the coil. The current induced in the coil creates another field, in the opposite direction of the bar magnet's to oppose the increase. This is one aspect of *Lenz's law—induction opposes any change in flux*. (b) and (c) are two other situations. Verify for yourself that the direction of the induced B_{coil} shown indeed opposes the change in flux and that the current direction shown is consistent with RHR-2.

Applications of Electromagnetic Induction

There are many applications of Faraday's Law of induction, as we will explore in this chapter and others. At this juncture, let us mention several that have to do with data storage and magnetic fields. A very important application has to do with audio and video *recording tapes*. A plastic tape, coated with iron oxide, moves past a recording head. This recording head is basically a round iron ring about which is wrapped a coil of wire—an electromagnet (Figure 3.2.10.2). A signal in the form of a varying input current from a microphone or camera goes to the recording head. These signals (which are a function of the signal amplitude and

frequency) produce varying magnetic fields at the recording head. As the tape moves past the recording head, the magnetic field orientations of the iron oxide molecules on the tape are changed thus recording the signal. In the playback mode, the magnetized tape is run past another head, similar in structure to the recording head. The different magnetic field orientations of the iron oxide molecules on the tape induces a voltage in the coil of wire in the playback head. This signal then is sent to a loudspeaker or video player.

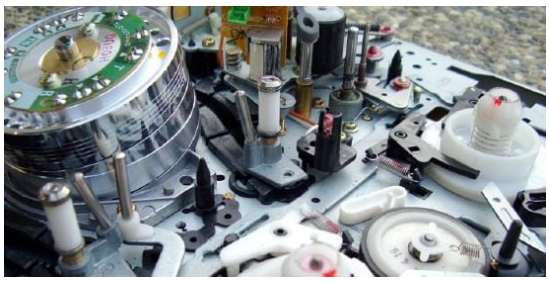


Figure 3.2.10.2: Recording and playback heads used with audio and video magnetic tapes. (credit: Steve Jurvetson)

Similar principles apply to computer hard drives, except at a much faster rate. Here recordings are on a coated, spinning disk. Read heads historically were made to work on the principle of induction. However, the input information is carried in digital rather than analog form – a series of 0's or 1's are written upon the spinning hard drive. Today, most hard drive readout devices do not work on the principle of induction, but use a technique known as *giant magnetoresistance*. (The discovery that weak changes in a magnetic field in a thin film of iron and chromium could bring about much larger changes in electrical resistance was one of the first large successes of nanotechnology.) Another application of induction is found on the magnetic stripe on the back of your personal credit card as used at the grocery store or the ATM machine. This works on the same principle as the audio or video tape mentioned in the last paragraph in which a head reads personal information from your card.

Another application of electromagnetic induction is when electrical signals need to be transmitted across a barrier. Consider the *cochlear implant* shown below. Sound is picked up by a microphone on the outside of the skull and is used to set up a varying magnetic field. A current is induced in a receiver secured in the bone beneath the skin and transmitted to electrodes in the inner ear. Electromagnetic induction can be used in other instances where electric signals need to be conveyed across various media.



Figure 3.2.10.3: Electromagnetic induction used in transmitting electric currents across mediums. The device on the baby's head induces an electrical current in a receiver secured in the bone beneath the skin. (credit: Bjorn Knetsch)

Another contemporary area of research in which electromagnetic induction is being successfully implemented (and with substantial potential) is transcranial magnetic simulation. A host of disorders, including depression and hallucinations can be traced to irregular localized electrical activity in the brain. In *transcranial magnetic stimulation*, a rapidly varying and very localized magnetic field is placed close to certain sites identified in the brain. Weak electric currents are induced in the identified sites and can result in recovery of electrical functioning in the brain tissue.

Sleep apnea ("the cessation of breath") affects both adults and infants (especially premature babies and it may be a cause of sudden infant deaths [SID]). In such individuals, breath can stop repeatedly during their sleep. A cessation of more than 20 seconds can be very dangerous. Stroke, heart failure, and tiredness are just some of the possible consequences for a person having sleep apnea. The concern in infants is the stopping of breath for these longer times. One type of monitor to alert parents when a child is not breathing uses electromagnetic induction. A wire wrapped around the infant's chest has an alternating current running through it. The expansion and contraction of the infant's chest as the infant breathes changes the area through the coil. A pickup coil located nearby has an alternating current induced in it due to the changing magnetic field of the initial wire. If the child stops breathing, there will be a change in the induced current, and so a parent can be alerted.

MAKING CONNECTIONS: CONSERVATION OF ENERGY

Lenz's law is a manifestation of the conservation of energy. The induced voltage produces a current that opposes the change in flux, because a change in flux means a change in energy. Energy can enter or leave, but not instantaneously. Lenz's law is a consequence. As the change begins, the law says induction opposes and, thus, slows the change. In fact, if the induced voltage were in the same direction as the change in flux, there would be a positive feedback that would give us free energy from no apparent source—conservation of energy would be violated.

Section Summary

- Faraday's law of induction states that the voltage induced by a change in magnetic flux is

$$V = -N\frac{\Delta\Phi}{\Delta t}$$

  when flux changes by $\Delta\Phi$ in a time Δt.

- If voltage is induced in a coil, N is its number of turns.
- The minus sign means that the induced voltage creates a current I and magnetic field B that *oppose the change in flux* $\Delta\Phi$ —this opposition is known as Lenz's law.

Glossary

Faraday's law of induction
: the means of calculating the voltage in a coil due to changing magnetic flux, given by $V = -N\frac{\Delta\Phi}{\Delta t}$

Lenz's law
: the minus sign in Faraday's law, signifying that the voltage induced in a coil opposes the change in magnetic flux

This page titled 3.2.10: Faraday's Law of Induction- Lenz's Law is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.10: Faraday's Law of Induction- Lenz's Law by Andrew Park is licensed CC BY 4.0.

3.2.11: Transformers

Learning Objectives

- Explain how a transformer works.
- Calculate voltage, current, and/or number of turns given the other quantities.

Transformers do what their name implies—they transform voltages from one value to another. For example, many cell phones, laptops, video games, and power tools and small appliances have a transformer built into their plug-in unit (like that in Figure 3.2.11.1) that changes 120 V or 240 V AC into whatever voltage the device uses. Transformers are also used at several points in the power distribution systems, such as illustrated in Figure 3.2.11.2. Power is sent long distances at high voltages, because less current is required for a given amount of power, and this means less line loss, as was discussed previously. But high voltages pose greater hazards, so that transformers are employed to produce lower voltage at the user's location.



Figure 3.2.11.1: The plug-in transformer has become increasingly familiar with the proliferation of electronic devices that operate on voltages other than common 120 V AC. Most are in the 3 to 12 V range. (credit: Shop Xtreme)

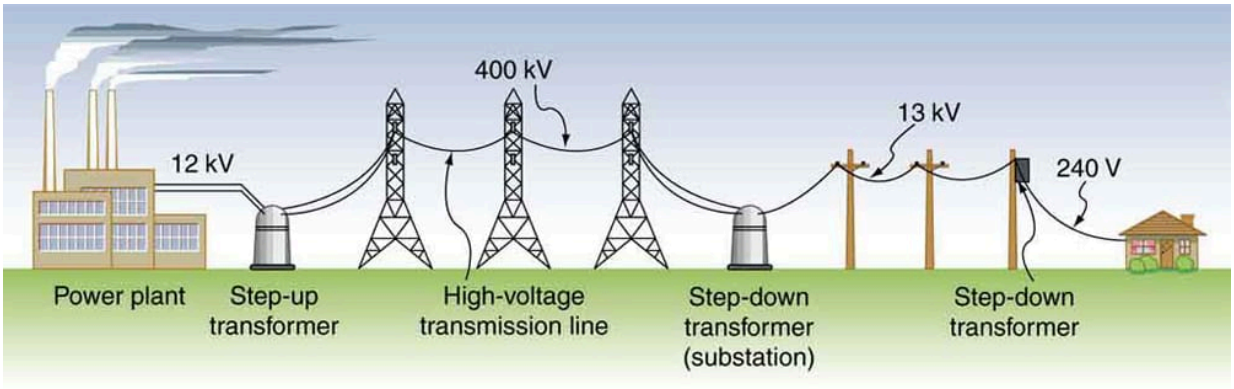


Figure 3.2.11.2: Transformers change voltages at several points in a power distribution system. Electric power is usually generated at greater than 10 kV, and transmitted long distances at voltages over 200 kV—sometimes as great as 700 kV—to limit energy losses. Local power distribution to neighborhoods or industries goes through a substation and is sent short distances at voltages ranging from 5 to 13 kV. This is reduced to 120, 240, or 480 V for safety at the individual user site.

The type of transformer considered in this text—see Figure 3.2.11.3 —is based on Faraday's law of induction and is very similar in construction to the apparatus Faraday used to demonstrate magnetic fields could cause currents. The two coils are called the *primary* and *secondary coils*. In normal use, the input voltage is placed on the primary, and the secondary produces the transformed output voltage. Not only does the iron core trap the magnetic field created by the primary coil, its magnetization increases the field strength. Since the input voltage is AC, a time-varying magnetic flux is sent to the secondary, inducing its AC output voltage.

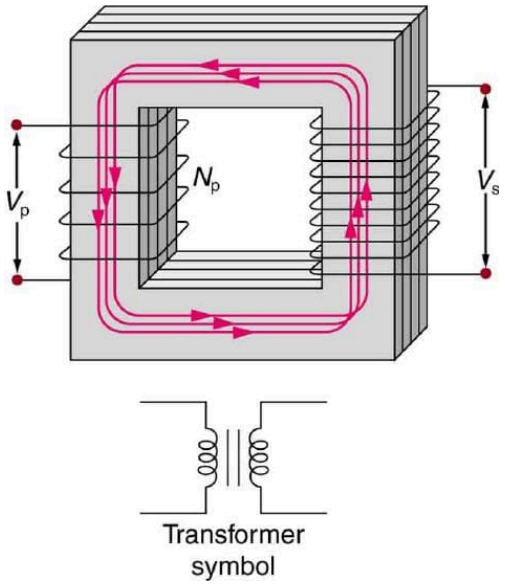


Figure 3.2.11.3: A typical construction of a simple transformer has two coils wound on a ferromagnetic core that is laminated to minimize eddy currents. The magnetic field created by the primary is mostly confined to and increased by the core, which transmits it to the secondary coil. Any change in current in the primary induces a current in the secondary.

For the simple transformer shown in Figure 3.2.11.3, the output voltage V_s depends almost entirely on the input voltage V_p and the ratio of the number of loops in the primary and secondary coils. Faraday's law of induction for the secondary coil gives its induced output voltage V_s to be

$$V_\mathrm{s} = -N_\mathrm{s}\frac{\Delta\Phi}{\Delta t},$$

where N_s is the number of loops in the secondary coil and $\Delta\Phi/\Delta t$ is the rate of change of magnetic flux. The cross-sectional area of the coils is the same on either side, as is the magnetic field strength, and so $\Delta\Phi/\Delta t$ is the same on either side. The input primary voltage V_p is also related to changing flux by

$$V_p = -N_\mathrm{p}\frac{\Delta\Phi}{\Delta t}.$$

The reason for this is a little more subtle. Lenz's law tells us that the primary coil opposes the change in flux caused by the input voltage V_p, hence the minus sign. Assuming negligible coil resistance, Kirchhoff's loop rule tells us that the induced voltage exactly equals the input voltage. Taking the ratio of these last two equations yields a useful relationship:

$$\frac{V_\mathrm{s}}{V_\mathrm{p}} = \frac{N_\mathrm{s}}{N_\mathrm{p}}.$$

This is known as the **transformer equation**, and it simply states that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of loops in their coils.

The output voltage of a transformer can be less than, greater than, or equal to the input voltage, depending on the ratio of the number of loops in their coils. Some transformers even provide a variable output by allowing connection to be made at different points on the secondary coil. A **step-up transformer** is one that increases voltage, whereas a **step-down transformer** decreases voltage. Assuming, as we have, that resistance is negligible, the electrical power output of a transformer equals its input. This is nearly true in practice—transformer efficiency often exceeds 99%. Equating the power input and output,

$$P_\mathrm{p} = I_\mathrm{p}V_\mathrm{p} = I_\mathrm{s}V_\mathrm{s} = P_\mathrm{s}.$$

Rearranging terms gives

$$\frac{V_\mathrm{s}}{V_\mathrm{p}} = \frac{I_\mathrm{p}}{I_\mathrm{s}}.$$

Combining this with $\frac{V_\mathrm{s}}{V_\mathrm{p}} = \frac{N_\mathrm{s}}{N_\mathrm{p}}$, we find that

$$\frac{I_s}{I_p} = \frac{N_p}{N_s}$$

is the relationship between the output and input currents of a transformer. So if voltage increases, current decreases. Conversely, if voltage decreases, current increases.

Example 3.2.11.1: Calculating Characteristics of a Step-Up Transformer

A portable x-ray unit has a step-up transformer, the 120 V input of which is transformed to the 100 kV output needed by the x-ray tube. The primary has 50 loops and draws a current of 10.00 A when in use. (a) What is the number of loops in the secondary? (b) Find the current output of the secondary.

Strategy and Solution for (a)

We solve $\frac{V_s}{V_p} = \frac{N_s}{N_p}$ for N_s, the number of loops in the secondary, and enter the known values. This gives

$$\begin{aligned} N_s &= N_p \frac{V_s}{V_p} \\ &= (50)\frac{100,000 \text{ V}}{120 \text{ V}} = 4.17 \times 10^4. \end{aligned}$$

Discussion for (a)

A large number of loops in the secondary (compared with the primary) is required to produce such a large voltage. This would be true for neon sign transformers and those supplying high voltage inside TVs and CRTs.

Strategy and Solution for (b)

We can similarly find the output current of the secondary by solving $\frac{I_s}{I_p} = \frac{N_p}{N_s}$ for I_s and entering known values. This gives

$$\begin{aligned} I_s &= I_p \frac{N_p}{N_s} \\ &= (10.00 \text{ A})\frac{50}{4.17 \times 10^4} = 12.0 \text{ mA}. \end{aligned}$$

Discussion for (b)

As expected, the current output is significantly less than the input. In certain spectacular demonstrations, very large voltages are used to produce long arcs, but they are relatively safe because the transformer output does not supply a large current. Note that the power input here is $P_p = I_p V_p = (10.00 \text{ A})(120 \text{ V}) = 1.20 \text{ kW}$. This equals the power output $P_p = I_s V_s = (12.0 \text{ mA})(100\text{kV}) = 1.20 \text{ kW}$, as we assumed in the derivation of the equations used.

Example 3.2.11.2: Calculating Characteristics of a Step-Down Transformer

A battery charger meant for a series connection of ten nickel-cadmium batteries needs to have a 15.0 V output to charge the batteries. It uses a step-down transformer with a 200-loop primary and a 120 V input. (a) How many loops should there be in the secondary coil? (b) If the charging current is 16.0 A, what is the input current?

Strategy and Solution for (a)

You would expect the secondary to have a small number of loops. Solving $\frac{V_s}{V_p} = \frac{N_s}{N_p}$ for N_s and entering known values gives

$$\begin{aligned} N_s &= N_p \frac{V_s}{V_p} \\ &= (200)\frac{15.0 \text{ V}}{120 \text{ V}} = 25. \end{aligned}$$

Strategy and Solution for (b)

The current input can be obtained by solving $\frac{I_s}{I_p} = \frac{N_p}{N_s}$ for I_p and entering known values. This gives

$$\begin{aligned} I_p &= I_s \frac{N_s}{N_p} \\ &= (16.0\ \text{A})\frac{25}{200} = 2.00\ \text{A}. \end{aligned}$$

Discussion

The number of loops in the secondary is small, as expected for a step-down transformer. We also see that a small input current produces a larger output current in a step-down transformer. When transformers are used to operate large magnets, they sometimes have a small number of very heavy loops in the secondary. This allows the secondary to have low internal resistance and produce large currents. Note again that this solution is based on the assumption of 100% efficiency—or power out equals power in ($P_p = P_s$)—reasonable for good transformers. In this case the primary and secondary power is 240 W. (Verify this for yourself as a consistency check.) Note that the Ni-Cd batteries need to be charged from a DC power source (as would a 12 V battery). So the AC output of the secondary coil needs to be converted into DC. This is done using something called a rectifier, which uses devices called diodes that allow only a one-way flow of current.

Section Summary

- Transformers use induction to transform voltages from one value to another.
- For a transformer, the voltages across the primary and secondary coils are related by

$$\frac{V_s}{V_p} = \frac{N_s}{N_p},$$

  where V_p and V_s are the voltages across primary and secondary coils having N_p and N_s turns.
- The currents I_p and I_s in the primary and secondary coils are related by $\frac{I_s}{I_p} = \frac{N_p}{N_s}$.
- A step-up transformer increases voltage and decreases current, whereas a step-down transformer decreases voltage and increases current.

Glossary

transformer
: a device that transforms voltages from one value to another using induction

transformer equation
: the equation showing that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of loops in their coils; $\frac{V_s}{V_p} = \frac{N_s}{N_p}$

step-up transformer
: a transformer that increases voltage

step-down transformer
: a transformer that decreases voltage

This page titled 3.2.11: Transformers is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.11: Transformers by Andrew Park is licensed CC BY 4.0.

3.2.12: Alternating Current versus Direct Current

Learning Objectives

- Explain the differences and similarities between AC and DC current.
- Describe rms voltage, current, and average power.
- Explain why AC current is used for power transmission.

Alternating Current

Most of the examples in electric circuits, and particularly those utilizing batteries, have constant voltage sources. Once the current is established, it is thus also a constant. **Direct current** (DC) is the flow of electric charge in only one direction. It is the steady state of a constant-voltage circuit. Many well-known applications, however, use a time-varying voltage source. **Alternating current** (AC) is the flow of electric charge that periodically reverses direction. If the source varies periodically, particularly sinusoidally, the circuit is known as an alternating current circuit. Examples include the commercial and residential power that serves so many of our needs. Figure 3.2.12.1 shows graphs of voltage and current versus time for typical DC and AC power. The AC voltages and frequencies commonly used in homes and businesses vary around the world. The AC voltages range from 100 V to 240 V; the frequencies range from 50 Hz to 60 Hz.

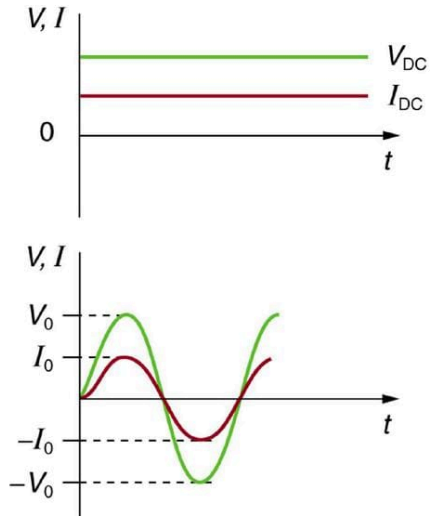


Figure 3.2.12.1: (a) DC voltage and current are constant in time, once the current is established. (b) A graph of voltage and current versus time for 60-Hz AC power. The voltage and current are sinusoidal and are in phase for a simple resistance circuit. The frequencies and peak voltages of AC sources differ greatly.

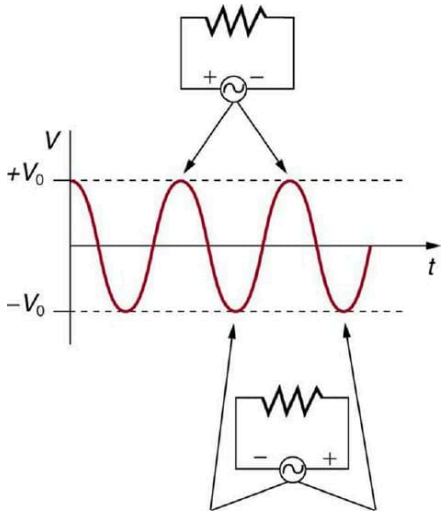


Figure 3.2.12.2: The potential difference V between the terminals of an AC voltage source fluctuates as shown.

Figure 3.2.12.2 shows a schematic of a simple circuit with an AC voltage source. The voltage between the terminals fluctuates as shown. For this example, the voltage and current are said to be in phase, as seen in Figure 3.2.12.1(b).

Current in the resistor alternates back and forth just like the driving voltage, since $I = V/R$. If the resistor is a fluorescent light bulb, for example, it brightens and dims 120 times per second as the current repeatedly goes through zero. A 120-Hz flicker is too rapid for your eyes to detect, but if you wave your hand back and forth between your face and a fluorescent light, you will see a stroboscopic effect evidencing AC. The fact that the light output fluctuates means that the power is fluctuating. The power supplied is $P = IV$.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—AC/DC LIGHTS

Wave your hand back and forth between your face and a fluorescent light bulb. Do you observe the same thing with the headlights on your car? Explain what you observe. *Warning: Do not look directly at very bright light*.

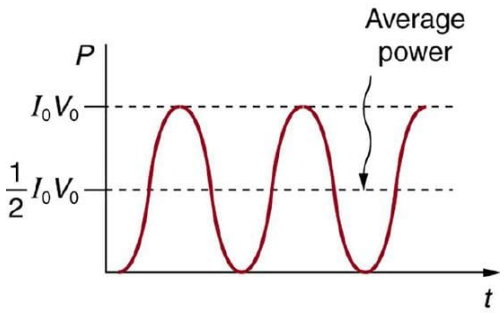


Figure 3.2.12.3: AC power as a function of time.

We are most often concerned with average power rather than its fluctuations—that 60-W light bulb in your desk lamp has an average power consumption of 60 W, for example. As illustrated in Figure 3.2.12.3. One common way to express an average is "root-mean-square," or "**rms**." For example, rms voltage of an AC voltage source is found by first squaring the voltage ("square"), taking an average of this value over one period of oscillation ("mean"), and taking the square root ("root").

It is standard practice to quote I_{rms} , V_{rms} , and P_{ave} rather than the peak values. For example, most household electricity is 120 V AC, which means that V_{rms} is 120 V. The common 10-A circuit breaker will interrupt a sustained I_{rms} greater than 10 A. Your 1.0-kW microwave oven consumes $P_{\text{ave}} = 1.0\ \text{kW}$, and so on. You can think of these rms and average values as the equivalent DC values for a simple resistive circuit.

To summarize, when dealing with AC, Ohm's law and the equations for power are completely analogous to those for DC, but rms and average values are used for AC. Thus, for AC, Ohm's law is written

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{R}.$$

The various expressions for AC power P_{ave} are

$$P_{\text{ave}} = I_{\text{rms}}\, V_{\text{rms}}\,,$$

$$P_{\text{ave}} = \frac{V_{\text{rms}}^2}{R},$$

and

$$P_{\text{ave}} = I_{\text{rms}}^2 R.$$

Why Use AC for Power Distribution?

Most large power-distribution systems are AC. Moreover, the power is transmitted at much higher voltages than the 120-V AC (240 V in most parts of the world) we use in homes and on the job. Economies of scale make it cheaper to build a few very large electric power-generation plants than to build numerous small ones. This necessitates sending power long distances, and it is obviously important that energy losses en route be minimized. High voltages can be transmitted with much smaller power losses than low voltages, as we shall show below. (See Figure 3.2.12.4.) For safety reasons, the voltage at the user is reduced to familiar values. The crucial factor is that AC voltages can be increased and decreased efficiently with transformers (which uses electromagnetic induction to produce time-varying voltages), while it is more difficult to change DC voltages without power losses. So AC is used in most large power distribution systems.



Figure 3.2.12.4: Power is distributed over large distances at high voltage to reduce power loss in the transmission lines. The voltages generated at the power plant are stepped up by passive devices called transformers to 330,000 volts (or more in some places worldwide). At the point of use, the transformers reduce the voltage transmitted for safe residential and commercial use. (Credit: GeorgHH, Wikimedia Commons)

Example 3.2.12.1: Power Losses Are Less for High-Voltage Transmission

(a) What current is needed to transmit 100 MW of power at 200 kV? (b) What is the power dissipated by the transmission lines if they have a resistance of $1.00\ \Omega$? (c) What percentage of the power is lost in the transmission lines?

Strategy

We are given $P_{\text{ave}} = 100\ \text{MW}$, $V_{\text{rms}} = 200\ \text{kV}$, and the resistance of the lines is $R = 1.00\ \Omega$. Using these givens, we can find the current flowing (from $P = IV$) and then the power dissipated in the lines ($P = I^2R$), and we take the ratio to the total power transmitted.

Solution

To find the current, we rearrange the relationship $P_{\text{ave}} = I_{\text{rms}}\, V_{\text{rms}}$ and substitute known values. This gives

$$I_{\text{rms}} = \frac{P_{\text{ave}}}{V_{\text{rms}}} = \frac{100 \times 10^6\ \text{W}}{200 \times 10^3\ \text{V}} = 500\ \text{A}.$$

Solution

Knowing the current and given the resistance of the lines, the power dissipated in them is found from $P_{\text{ave}} = I_{\text{rms}}^2 R$. Substituting the known values gives

$$P_{\text{ave}} = I_{\text{rms}}^2 R = (500 \text{ A})^2(1.00\Omega) = 250 \text{ kW}.$$

Solution

The percent loss is the ratio of this lost power to the total or input power, multiplied by 100:

$$\% \text{ loss } = \frac{250 \text{ kW}}{100 \text{ MW}} \times 100 = 0.250 \text{ \%}.$$

Discussion

One-fourth of a percent is an acceptable loss. Note that if 100 MW of power had been transmitted at 25 kV, then a current of 4000 A would have been needed. This would result in a power loss in the lines of 16.0 MW, or 16.0% rather than 0.250%. The lower the voltage, the more current is needed, and the greater the power loss in the fixed-resistance transmission lines. Of course, lower-resistance lines can be built, but this requires larger and more expensive wires. If superconducting lines could be economically produced, there would be no loss in the transmission lines at all. But, as we shall see in a later chapter, there is a limit to current in superconductors, too. In short, high voltages are more economical for transmitting power, and AC voltage is much easier to raise and lower, so that AC is used in most large-scale power distribution systems.

Section Summary

- Ohm's law for AC is $I_{\text{rms}} = \frac{V_{\text{rms}}}{R}$.
- Expressions for the average power of an AC circuit are $P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}}$, $P_{\text{ave}} = \frac{{V_{\text{rms}}}^2}{R}$, and $P_{\text{ave}} = I_{\text{rms}}^2 R$, analogous to the expressions for DC circuits.

Glossary

direct current
: (DC) the flow of electric charge in only one direction

alternating current
: (AC) the flow of electric charge that periodically reverses direction

AC voltage
: voltage that fluctuates sinusoidally with time.

AC current
: current that fluctuates sinusoidally with time.

rms
: a type of average taken for a time-varying quantity by squaring it, taking the mean of the square, and then taking the square-root of the mean.

This page titled 3.2.12: Alternating Current versus Direct Current is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.12: Alternating Current versus Direct Current by Andrew Park is licensed CC BY 4.0.

3.2.13: Magnetism (Exercise)

Conceptual Questions

10.1: Magnets

1. Volcanic and other such activity at the mid-Atlantic ridge extrudes material to fill the gap between separating tectonic plates associated with continental drift. The magnetization of rocks is found to reverse in a coordinated manner with distance from the ridge. What does this imply about the Earth's magnetic field and how could the knowledge of the spreading rate be used to give its historical record?

10.3: Magnetic Fields and Magnetic Field Lines

2. Explain why the magnetic field would not be unique (that is, not have a single value) at a point in space where magnetic field lines might cross. (Consider the direction of the field at such a point.)

3. List the ways in which magnetic field lines and electric field lines are similar. For example, the field direction is tangent to the line at any point in space. Also list the ways in which they differ. For example, electric force is parallel to electric field lines, whereas magnetic force on moving charges is perpendicular to magnetic field lines.

4. Noting that the magnetic field lines of a bar magnet resemble the electric field lines of a pair of equal and opposite charges, do you expect the magnetic field to rapidly decrease in strength with distance from the magnet? Is this consistent with your experience with magnets?

5. Is the Earth's magnetic field parallel to the ground at all locations? If not, where is it parallel to the surface? Is its strength the same at all locations? If not, where is it greatest?

10.4: Magnetic Field Strength- Force on a Moving Charge in a Magnetic Field

6. If a charged particle moves in a straight line through some region of space, can you say that the magnetic field in that region is necessarily zero?

10.5: Magnetic Force on a Current-Carrying Conductor

7. Draw a sketch of the situation in the figure showing the direction of electrons carrying the current, and use RHR-1 to verify the direction of the force on the wire.

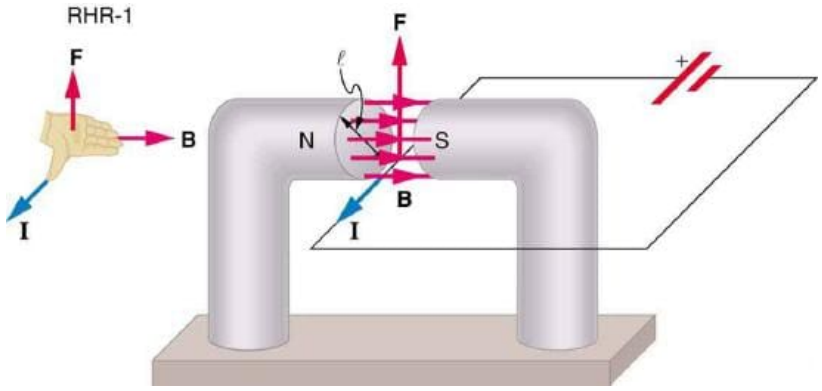


Figure 3.2.13.1: The magnetic field exerts a force on a current-carrying wire in a direction given by the right hand rule 1 (the same direction as that on the individual moving charges). This force can easily be large enough to move the wire, since typical currents consist of very large numbers of moving charges.

8. Verify that the direction of the force in an MHD drive, such as that in the figure, does not depend on the sign of the charges carrying the current across the fluid.

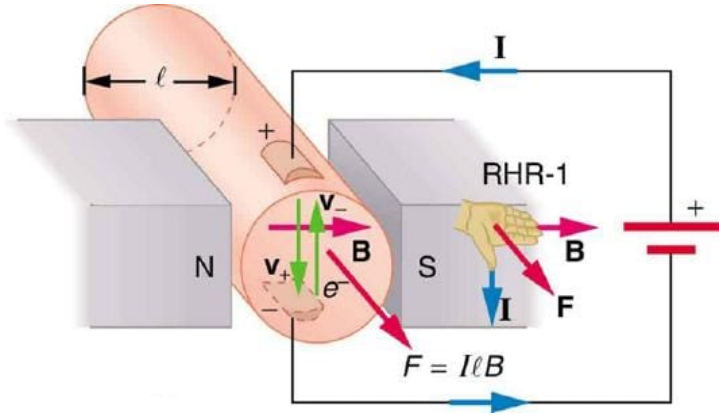


Figure 3.2.13.2: Magnetohydrodynamics. The magnetic force on the current passed through this fluid can be used as a nonmechanical pump.

9. Why would a magnetohydrodynamic drive work better in ocean water than in fresh water? Also, why would superconducting magnets be desirable?

10. Which is more likely to interfere with compass readings, AC current in your refrigerator or DC current when you start your car? Explain.

10.8: Induced Voltage and Magnetic Flux

11. How do the multiple-loop coils and iron ring in the version of Faraday's apparatus shown in the figure enhance the observation of induced voltage?

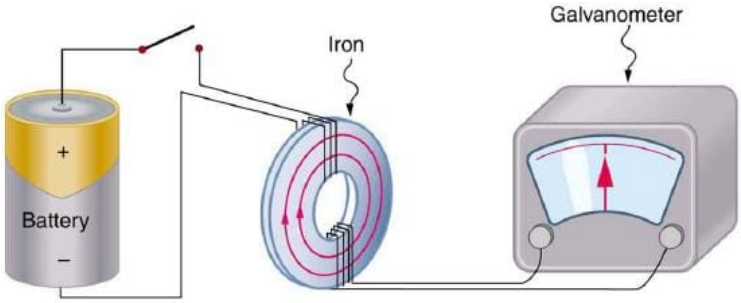


Figure 3.2.13.3: Faraday's apparatus for demonstrating that a magnetic field can produce a current. A change in the field produced by the top coil induces an voltage and, hence, a current in the bottom coil. When the switch is opened and closed, the galvanometer registers currents in opposite directions. No current flows through the galvanometer when the switch remains closed or open.

12. When a magnet is thrust into a coil as in figure (a), what is the direction of the force exerted by the coil on the magnet? Draw a diagram showing the direction of the current induced in the coil and the magnetic field it produces, to justify your response. How does the magnitude of the force depend on the resistance of the galvanometer?

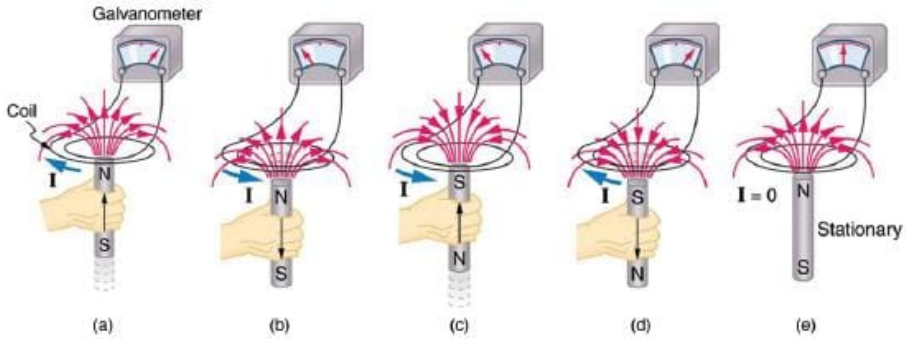


Figure 3.2.13.4: Movement of a magnet relative to a coil produces voltage as shown. The same voltages are produced if the coil is moved relative to the magnet. The greater the speed, the greater the magnitude of the voltage, and the voltage is zero when there is no motion.

13. Explain how magnetic flux can be zero when the magnetic field is not zero.

14. Is a voltage induced in the coil in the figure when it is stretched? If so, state why and give the direction of the induced current.

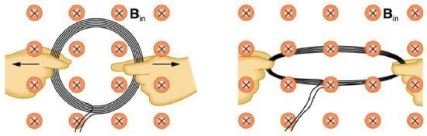


Figure 3.2.13.5: A circular coil of wire is stretched in a magnetic field.

10.9: Faraday's Law of Induction- Lenz's Law

15. A person who works with large magnets sometimes places her head inside a strong field. She reports feeling dizzy as she quickly turns her head. How might this be associated with induction?

16. A particle accelerator sends high-velocity charged particles down an evacuated pipe. Explain how a coil of wire wrapped around the pipe could detect the passage of individual particles. Sketch a graph of the voltage output of the coil as a single particle passes through it.

10.11: Alternating Current versus Direct Current

17. Give an example of a use of AC power other than in the household. Similarly, give an example of a use of DC power other than that supplied by batteries.

18. Why do voltage, current, and power go through zero 120 times per second for 60-Hz AC electricity?

19. You are riding in a train, gazing into the distance through its window. As close objects streak by, you notice that the nearby fluorescent lights make *dashed* streaks. Explain.

Problems & Exercises

10.4: Magnetic Field Strength- Force on a Moving Charge in a Magnetic Field

20. What is the direction of the magnetic force on a positive charge that moves as shown in each of the six cases shown in the figure? Note that ⊙⊙ indicates "coming out of the page" and ⊗⊗ means "going into the page."

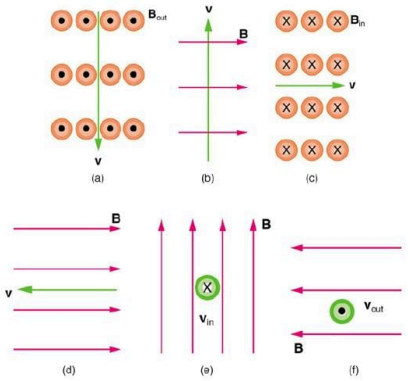


Figure 3.2.13.6

Solution

(a) Left (West)

(b) Into the page

(c) Up (North)

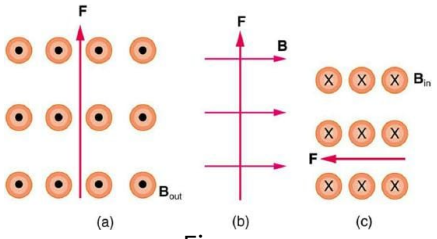
(d) No force

(e) Right (East)

(f) Down (South)

21. Repeat Exercise 11.4.20 for a negative charge.

22. What is the direction of the velocity of a negative charge that experiences the magnetic force shown in each of the three cases in the figure, assuming it moves perpendicular to B ? Note that $\odot\odot$ indicates "coming out of the page" and $\otimes\otimes$ means "going into the page."



Figure

Solution

(a) East (right)

(b) Into page

(c) South (down)

23. Repeat Exercise 11.4.21 for a positive charge.

24. What is the direction of the magnetic field that produces the magnetic force on a positive charge as shown in each of the three cases in the figure below, assuming B is perpendicular to v? Note that $\otimes\otimes$ means "going into the page."

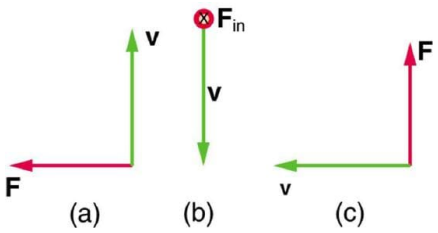


Figure 3.2.13.8

Solution

(a) Into page

(b) West (left)

(c) Out of page

25. Repeat Exercise 11.4.24 for a negative charge.

26. What is the maximum force on an aluminum rod with a $0.100 - \mu\text{C}$ charge that you pass between the poles of a 1.50-T permanent magnet at a speed of 5.00 m/s? In what direction is the force?

Solution

7.50×10^{-7} N perpendicular to both the magnetic field lines and the velocity

27. (a) Aircraft sometimes acquire small static charges. Suppose a supersonic jet has a $0.500 - \mu\text{C}$ charge and flies due west at a speed of 660 m/s over the Earth's south magnetic pole, where the $8.00 \times 10^{-5} - \text{T}$ magnetic field points straight up. What are the direction and the magnitude of the magnetic force on the plane? (b) Discuss whether the value obtained in part (a) implies this is a significant or negligible effect.

10.5: Magnetic Force on a Current-Carrying Conductor

28. What is the direction of the magnetic force on the current in each of the six cases in the figure? Note that ⊙⊙ indicates "coming out of the page" and ⊗⊗ means "going into the page."

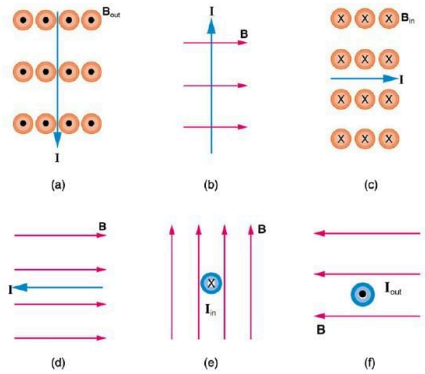


Figure 3.2.13.9

Solution

(a) west (left)

(b) into page

(c) north (up)

(d) no force

(e) east (right)

(f) south (down)

29. What is the direction of a current that experiences the magnetic force shown in each of the three cases in the figure, assuming the current runs perpendicular to B? Note that $\odot\odot$ indicates "coming out of the page" and $\otimes\otimes$ means "going into the page."

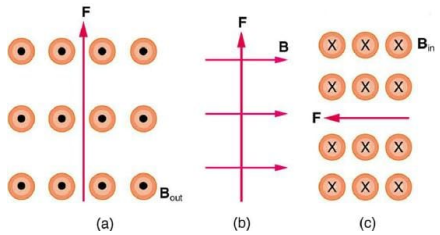


Figure
3.2.13.10

30. What is the direction of the magnetic field that produces the magnetic force shown on the currents in each of the three cases in the figure, assuming $\mathbf{B}$ is perpendicular to $\mathbf{I}$? Note that $\otimes\otimes$ means "going into the page."

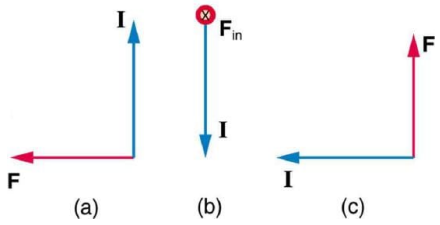


Figure
3.2.13.11

Solution

(a) into page

(b) west (left)

(c) out of page

31. (a) What is the force per meter on a lightning bolt at the equator that carries 20,000 A perpendicular to the Earth's $3.00 \times 10^{-5} - \mathrm{T}$ field?

(b) What is the direction of the force if the current is straight up and the Earth's field direction is due north, parallel to the ground?

32. (a) A DC power line for a light-rail system carries 1000 A. If Earth's magnetic field at this location is 5.00×10^{-5} T, what is the maximum possible magnetic force on a 100-m section of this line?

(b) Discuss practical concerns this presents, if any.

Solution

(a) 5.00 N

(b) This is about a pound of force per 100 m of wire, which is much less than the weight of the wire itself. Therefore, it does not cause any special concerns.

33. What force is exerted on the water in an MHD drive utilizing a 25.0-cm-diameter tube, if 100-A current is passed across the tube that is perpendicular to a 2.00-T magnetic field? (The relatively small size of this force indicates the need for very large currents and magnetic fields to make practical MHD drives.)

34. A wire carrying a 30.0-A current passes between the poles of a strong magnet that is perpendicular to its field and experiences a 2.16-N force on the 4.00 cm of wire in the field. What is the average field strength?

Solution

1.80 T

10.10: Transformers

35. A plug-in transformer, like that in Section 11.10, supplies 9.00 V to a video game system.

(a) How many turns are in its secondary coil, if its input voltage is 120 V and the primary coil has 400 turns?

(b) What is its input current when its output is 1.30 A?

Solution

(a) 30.0

(b) 9.75×10^{-2} A

36. An American traveler in New Zealand carries a transformer to convert New Zealand's standard 240 V to 120 V so that she can use some small appliances on her trip.

(a) What is the ratio of turns in the primary and secondary coils of her transformer?

(b) What is the ratio of input to output current?

(c) How could a New Zealander traveling in the United States use this same transformer to power her 240 V appliances from 120 V?

37. A digital recorder uses a plug-in transformer to convert 120 V to 12.0 V, with a maximum current output of 200 mA.

(a) What is the current input?

(b) What is the power input?

(c) Is this amount of power reasonable for a small appliance?

Solution

(a) 20.0 mA

(b) 2.40 W

(c) Yes, this amount of power is quite reasonable for a small appliance.

38. (a) What is the voltage output of a transformer used for rechargeable flashlight batteries, if its primary has 500 turns, its secondary 4 turns, and the input voltage is 120 V?

(b) What input current is required to produce a 4.00 A output?

(c) What is the power input?

39. (a) The plug-in transformer for a laptop computer puts out 7.50 V and can supply a maximum current of 2.00 A. What is the maximum input current if the input voltage is 240 V? Assume 100% efficiency.

(b) If the actual efficiency is less than 100%, would the input current need to be greater or smaller? Explain.

Solution

(a) 0.063 A

(b) Greater input current needed.

40. A multipurpose transformer has a secondary coil with several points at which a voltage can be extracted, giving outputs of 5.60, 12.0, and 480 V.

(a) The input voltage is 240 V to a primary coil of 280 turns. What are the numbers of turns in the parts of the secondary used to produce the output voltages?

(b) If the maximum input current is 5.00 A, what are the maximum output currents (each used alone)?

41. A large power plant generates electricity at 12.0 kV. Its old transformer once converted the voltage to 335 kV. The secondary of this transformer is being replaced so that its output can be 750 kV for more efficient cross-country transmission on upgraded transmission lines.

(a) What is the ratio of turns in the new secondary compared with the old secondary?

(b) What is the ratio of new current output to old output (at 335 kV) for the same power?

(c) If the upgraded transmission lines have the same resistance, what is the ratio of new line power loss to old?

Solution

(a) 2.2

(b) 0.45

(c) 0.20, or 20.0%

42. If the power output in the previous problem is 1000 MW and line resistance is $2.00\ \Omega$, what were the old and new line losses?

10.11: Alternating Current versus Direct Current

43. Military aircraft use 400-Hz AC power, because it is possible to design lighter-weight equipment at this higher frequency. What is the time for one complete cycle of this power?

Solution

2.50 ms

44. A North American tourist takes his 25.0-W, 120-V AC razor to Europe, finds a special adapter, and plugs it into 240 V AC. Assuming constant resistance, what power does the razor consume as it is ruined?

45. In this problem, you will verify statements made at the end of the power losses for Example 11.11.1.

(a) What current is needed to transmit 100 MW of power at a voltage of 25.0 kV?

(b) Find the power loss in a $1.00-\Omega$ transmission line.

(c) What percent loss does this represent?

Solution

(a) 4.00 kA

(b) 16.0 MW

(c) 16.0%

46. A small office-building air conditioner operates on 408-V AC and consumes 50.0 kW.

(a) What is its effective resistance?

(b) What is the cost of running the air conditioner during a hot summer month when it is on 8.00 h per day for 30 days and electricity costs $9.00\ \text{cents}\ /\text{kW}\cdot\text{h}$?

47. What is the average power consumption of a 120-V AC microwave oven that draws 10.0 A?

Solution

1.20 kW

48. What is the average current through a 500-W room heater that operates on 120-V AC power?

49. Two different electrical devices have the same power consumption, but one is meant to be operated on 120-V AC and the other on 240-V AC.

(a) What is the ratio of their resistances?

(b) What is the ratio of their currents?

(c) Assuming its resistance is unaffected, by what factor will the power increase if a 120-V AC device is connected to 240-V AC?

Solution

(a) 4.0

(b) 0.50

(c) 4.0

50. Nichrome wire is used in some radiative heaters. Find the resistance needed if the average power output is to be 1.00 kW utilizing 120-V AC.

This page titled 3.2.13: Magnetism (Exercise) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 11.13: Magnetism (Exercise) by Andrew Park is licensed CC BY 4.0.

3.3: LIGHT

3.3.1: Introduction to Light
3.3.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed
3.3.3: Production and Properties of Electromagnetic Waves
3.3.4: The Electromagnetic Spectrum- an Overview
3.3.5: The Electromagnetic Spectrum- Application Notes
3.3.6: Reflection
3.3.7: Refraction
3.3.8: Dispersion- The Rainbow and Prisms
3.3.9: Polarization
3.3.10: Image Formation by Lenses
3.3.11: Image Formation by Mirrors
3.3.12: Light (Exercises)

This page titled 3.3: Light is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

3.3.1: Introduction to Light

Light from this page or screen is formed into an image by the lens of your eye, much as the lens of the camera that made this photograph. Mirrors, like lenses, can also form images that in turn are captured by your eye.



Figure 3.3.1.1: Image seen as a result of reflection of light on a plane smooth surface. (credit: NASA Goddard Photo and Video, via Flickr)

Our lives are filled with light. Through vision, the most valued of our senses, light can evoke spiritual emotions, such as when we view a magnificent sunset or glimpse a rainbow breaking through the clouds. Light can also simply amuse us in a theater, or warn us to stop at an intersection. It has innumerable uses beyond vision. Light can carry telephone signals through glass fibers or cook a meal in a solar oven. Life itself could not exist without light's energy. From photosynthesis in plants to the sun warming a cold-blooded animal, its supply of energy is vital.



Figure 3.3.1.2: Double Rainbow over the bay of Pocitos in Montevideo, Uruguay. (credit: Madrax, Wikimedia Commons)

We will start our discussion of visible light as a type of electromagnetic wave. This knowledge will help us answer questions regarding the nature of light and vision. What is color, and how do our eyes detect it? Why do diamonds sparkle? How does light travel? How do lenses and mirrors form images? These are but a few of the questions that are answered by the study of optics. Optics is the branch of physics that deals with the behavior of visible light and other electromagnetic waves.

It is convenient to divide optics into two major parts based on the features we are interested in. The wave characteristics of light, such as frequency and wavelength, relates to the colors we perceive and to how we characterize different types of electromagnetic wave along the electromagnetic spectrum. The wave nature of light is also responsible for phenomena such as diffraction and interference. We call this part of optics "wave optics" or "physical optics." But when light interacts with an object that is several times as large as the light's wavelength, its observable behavior is like that of a ray; it does not prominently display its wave characteristics, and we can look at primarily how it refracts and reflects. We call this part of optics "geometric optics" or "ray optics."

This page titled 3.3.1: Introduction to Light is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.1: Introduction to Light by Andrew Park is licensed CC BY 4.0.

3.3.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed

Learning Objectives

- Restate Maxwell's equations.

The Scotsman James Clerk Maxwell (1831–1879) is regarded as the greatest theoretical physicist of the 19th century. (See Figure 3.3.2.1). Although he died young, Maxwell not only formulated a complete electromagnetic theory, represented by **Maxwell's equations**, he also developed the kinetic theory of gases and made significant contributions to the understanding of color vision and the nature of Saturn's rings.



Figure 3.3.2.1: James Clerk Maxwell, a 19th-century physicist, developed a theory that explained the relationship between electricity and magnetism and correctly predicted that visible light is caused by electromagnetic waves. (credit: G. J. Stodart)

Maxwell brought together all the work that had been done by brilliant physicists such as Oersted, Coulomb, Gauss, and Faraday, and added his own insights to develop the overarching theory of electromagnetism. The results of Maxwell's equations are paraphrased here in words because their mathematical statement is beyond the level of this text. However, the equations illustrate how apparently simple mathematical statements can elegantly unite and express a multitude of concepts—why mathematics is the language of science.

THE RESULTS OF MAXWELL'S EQUATIONS

1. Electric field lines originate on positive charges and terminate on negative charges. The electric field is defined as the force per unit charge on a test charge.
2. Magnetic field lines are continuous, having no beginning or end. No magnetic monopoles are known to exist.
3. A changing magnetic field induces an electric field.
4. Magnetic fields are generated by moving charges or by changing electric fields.

Maxwell's equations encompass the major laws of electricity and magnetism. What is not so apparent is the symmetry that Maxwell introduced in his mathematical framework. Especially important is his addition of the hypothesis that changing electric fields create magnetic fields. This is exactly analogous (and symmetric) to Faraday's law of induction and had been suspected for some time, but fits beautifully into Maxwell's equations.

MAKING CONNECTIONS: UNIFICATION OF FORCES

Maxwell's complete and symmetric theory showed that electric and magnetic forces are not separate, but different manifestations of the same thing—the electromagnetic force. This classical unification of forces is one motivation for current attempts to unify the four basic forces in nature—the gravitational, electrical, strong, and weak nuclear forces.

Since changing electric fields create relatively weak magnetic fields, they could not be easily detected at the time of Maxwell's hypothesis. Maxwell realized, however, that oscillating charges, like those in AC circuits, produce changing electric fields. He predicted that these changing fields would propagate from the source like waves generated on a lake by a jumping fish.

The waves predicted by Maxwell would consist of oscillating electric and magnetic fields—defined to be an electromagnetic wave (EM wave). Electromagnetic waves would be capable of exerting forces on charges great distances from their source, and they might thus be detectable. Maxwell calculated that electromagnetic waves would propagate at the speed of light,

$$c = 3.00 \times 10^8 \text{ m/s}.$$

In fact, Maxwell concluded that light is an electromagnetic wave having such wavelengths that it can be detected by the eye.

Other wavelengths should exist—it remained to be seen if they did. If so, Maxwell's theory and remarkable predictions would be verified, the greatest triumph of physics since Newton. Experimental verification came within a few years, but not before Maxwell's death.

Hertz's Observations

The German physicist Heinrich Hertz (1857–1894) was the first to generate and detect certain types of electromagnetic waves in the laboratory. Starting in 1887, he performed a series of experiments that not only confirmed the existence of electromagnetic waves, but also verified that they travel at the speed of light.

Hertz used an AC circuit that resonates at a known frequency and connected it to a loop of wire as shown in Figure 3.3.2.2. High voltages induced across the gap in the loop produced sparks that were visible evidence of the current in the circuit and that helped generate electromagnetic waves.

Across the laboratory, Hertz had another loop attached to another circuit, which could be tuned (as the dial on a radio) to the same resonant frequency as the first and could, thus, be made to receive electromagnetic waves. This loop also had a gap across which sparks were generated, giving solid evidence that electromagnetic waves had been received.

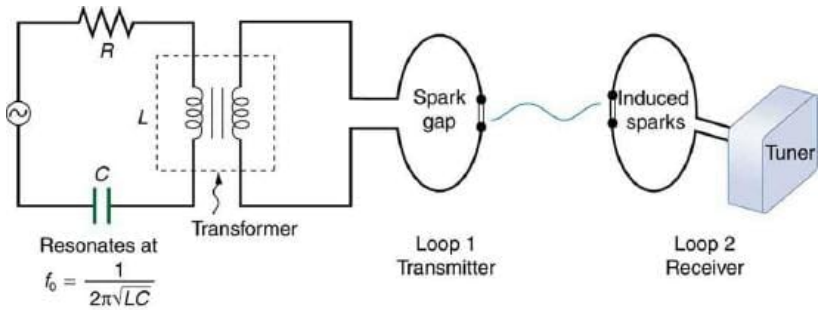


Figure 3.3.2.2: The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves. An AC circuit connected to the first loop caused sparks across a gap in the wire loop and generated electromagnetic waves. Sparks across a gap in the second loop located across the laboratory gave evidence that the waves had been received.

Hertz also studied the reflection, refraction, and interference patterns of the electromagnetic waves he generated, verifying their wave character. He was able to determine wavelength from the interference patterns, and knowing their frequency, he could calculate the propagation speed using the equation $v = f\lambda$ (velocity—or speed—equals frequency times wavelength). Hertz was thus able to prove that electromagnetic waves travel at the speed of light. The SI unit for frequency, the hertz $(1 \text{ Hz} = 1 \text{ cycle /sec})$, is named in his honor.

Section Summary

- Electromagnetic waves consist of oscillating electric and magnetic fields and propagate at the speed of light c. They were predicted by Maxwell, who also showed that
- Maxwell's prediction of electromagnetic waves resulted from his formulation of a complete and symmetric theory of electricity and magnetism, known as Maxwell's equations.
- These four equations are paraphrased in this text, rather than presented numerically, and encompass the major laws of electricity and magnetism. First is Gauss's law for electricity, second is Gauss's law for magnetism, third is Faraday's law of induction, including Lenz's law, and fourth is Ampere's law in a symmetric formulation that adds another source of magnetism—changing electric fields.

Glossary

electromagnetic waves
: radiation in the form of waves of electric and magnetic energy

Maxwell's equations
: a set of four equations that comprise a complete, overarching theory of electromagnetism

hertz
: an SI unit denoting the frequency of an electromagnetic wave, in cycles per second

speed of light
: in a vacuum, such as space, the speed of light is a constant 3×10^8 m/s

electric field lines
: a pattern of imaginary lines that extend between an electric source and charged objects in the surrounding area, with arrows pointed away from positively charged objects and toward negatively charged objects. The more lines in the pattern, the stronger the electric field in that region

magnetic field lines
: a pattern of continuous, imaginary lines that emerge from and enter into opposite magnetic poles. The density of the lines indicates the magnitude of the magnetic field

This page titled 3.3.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.2: Maxwell's Equations- Electromagnetic Waves Predicted and Observed by Andrew Park is licensed CC BY 4.0.

3.3.3: Production and Properties of Electromagnetic Waves

Learning Objectives

- Describe the electric and magnetic waves as they move out from a source, such as an AC generator.
- Describe the momentum carried by electromagnetic waves and its relationship to radiation pressure.

We can get a good understanding of **electromagnetic waves** (EM) by considering how they are produced. Whenever a current varies, associated electric and magnetic fields vary, moving out from the source like waves. Perhaps the easiest situation to visualize is a varying current in a long straight wire, produced by an AC generator at its center, as illustrated in Figure 3.3.3.1.

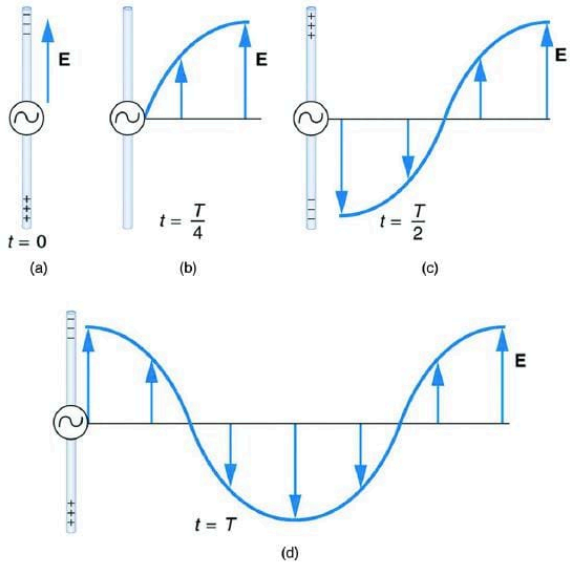


Figure 3.3.3.1: This long straight gray wire with an AC generator at its center becomes a broadcast antenna for electromagnetic waves. Shown here are the charge distributions at four different times. The electric field ($\mathbf{E}$) propagates away from the antenna at the speed of light, forming part of an electromagnetic wave.

The **electric field** ($\mathbf{E}$) shown surrounding the wire is produced by the charge distribution on the wire. Both the $\mathbf{E}$ and the charge distribution vary as the current changes. The changing field propagates outward at the speed of light.

There is an associated **magnetic field** ($\mathbf{B}$) which propagates outward as well (see Figure 3.3.3.2). The electric and magnetic fields are closely related and propagate as an electromagnetic wave. This is what happens in broadcast antennae such as those in radio and TV stations.

Closer examination of the one complete cycle shown in Figure 3.3.3.1 reveals the periodic nature of the generator-driven charges oscillating up and down in the antenna and the electric field produced. At time $t = 0$, there is the maximum separation of charge, with negative charges at the top and positive charges at the bottom, producing the maximum magnitude of the electric field (or E-field) in the upward direction. One-fourth of a cycle later, there is no charge separation and the field next to the antenna is zero, while the maximum E-field has moved away at speed c.

As the process continues, the charge separation reverses and the field reaches its maximum downward value, returns to zero, and rises to its maximum upward value at the end of one complete cycle. The outgoing wave has an **amplitude** proportional to the maximum separation of charge. Its **wavelength** (λ) is proportional to the period of the oscillation and, hence, is smaller for short periods or high frequencies. (As usual, wavelength and **frequency** (f) are inversely proportional.)

Electric and Magnetic Waves: Moving Together

Following Ampere's law, current in the antenna produces a magnetic field, as shown in Figure 3.3.3.2. The relationship between $\mathbf{E}$ and $\mathbf{B}$ is shown at one instant in Figure 3.3.3.2(a). As the current varies, the magnetic field varies in magnitude and direction.

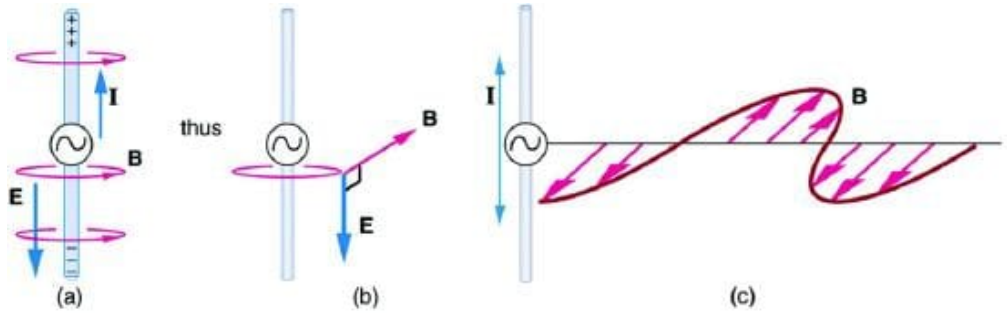


Figure 3.3.3.2: (a) The current in the antenna produces the circular magnetic field lines. The current (I) produces the separation of charge along the wire, which in turn creates the electric field as shown. (b) The electric and magnetic fields ($\mathbf{E}$ and $\mathbf{B}$) near the wire are perpendicular; they are shown here for one point in space. (c) The magnetic field varies with current and propagates away from the antenna at the speed of light.

The magnetic field lines also propagate away from the antenna at the speed of light, forming the other part of the electromagnetic wave, as seen in Figure 3.3.3.2(b). The magnetic part of the wave has the same period and wavelength as the electric part, since they are both produced by the same movement and separation of charges in the antenna.

The electric and magnetic waves are shown together at one instant in time in Figure 3.3.3.3. The electric and magnetic fields produced by a long straight wire antenna are exactly in phase. Note that they are perpendicular to one another and to the direction of propagation, making this a **transverse wave**.

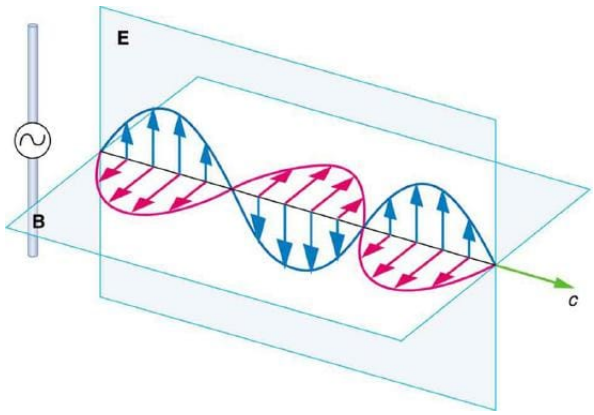


Figure 3.3.3.3: A part of the electromagnetic wave sent out from the antenna at one instant in time. The electric and magnetic fields ($\mathbf{E}$ and $\mathbf{B}$) are in phase, and they are perpendicular to one another and the direction of propagation. For clarity, the waves are shown only along one direction, but they propagate out in other directions too.

Electromagnetic waves generally propagate out from a source in all directions, sometimes forming a complex radiation pattern. A linear antenna like this one will not radiate parallel to its length, for example. The wave is shown in one direction from the antenna in Figure 3.3.3.3 to illustrate its basic characteristics.

Instead of the AC generator, the antenna can also be driven by an AC circuit. In fact, charges radiate whenever they are accelerated. But while a current in a circuit needs a complete path, an antenna has a varying charge distribution forming a **standing wave**, driven by the AC. The dimensions of the antenna are critical for determining the frequency of the radiated electromagnetic waves. This is a **resonant** phenomenon and when we tune radios or TV, we vary electrical properties to achieve appropriate resonant conditions in the antenna.

Receiving Electromagnetic Waves

Electromagnetic waves carry energy away from their source, similar to a sound wave carrying energy away from a standing wave on a guitar string. An antenna for receiving EM signals works in reverse. And like antennas that produce EM waves, receiver antennas are specially designed to resonate at particular frequencies.

An incoming electromagnetic wave accelerates electrons in the antenna, setting up a standing wave. If the radio or TV is switched on, electrical components pick up and amplify the signal formed by the accelerating electrons. The signal is then converted to audio and/or video format. Sometimes big receiver dishes are used to focus the signal onto an antenna.

In fact, charges radiate whenever they are accelerated. When designing circuits, we often assume that energy does not quickly escape AC circuits, and mostly this is true. A broadcast antenna is specially designed to enhance the rate of electromagnetic radiation, and shielding is necessary to keep the radiation close to zero. Some familiar phenomena are based on the production of electromagnetic waves by varying currents. Your microwave oven, for example, sends electromagnetic waves, called microwaves, from a concealed antenna that has an oscillating current imposed on it.

Momentum Carried by Electromagnetic Waves

As electromagnetic waves are received, as described above, the forces exerted on the charged particles (see [link][1] for diagrams) do work on the particles and increase its energy. The energy that sunlight carries is a familiar part of every warm sunny day. A much less familiar feature of electromagnetic radiation is the extremely weak pressure that electromagnetic radiation produces by exerting a force in the direction of the wave. This force occurs because electromagnetic waves contain and transport momentum.

To understand the direction of the force for a very specific case, consider a plane electromagnetic wave incident on a metal. The electrons in the metal move with the velocity v that is proportional to the force on it due to applied electric force (in opposite direction to E because electron is negatively charged). There are other, frictional forces in the metal that make this motion happen (this is the origin of Ohm's law in electric circuits). Note that there is also a magnetic field in the electromagnetic wave (Figure 3.3.3.4). Using the right-hand rule (magnetic force is $\boldsymbol{F} = q(\boldsymbol{v} \times \boldsymbol{B})$), we can get the direction of force on the electron (remember that electron is negatively charged), which is shown in Figure 3.3.3.4. This force is in the same direction as the direction of wave propagation, and it represents a transfer of momentum carried by the electromagnetic wave from the wave to the electron. This momentum is directly proportional to the energy carried by the electromagnetic wave and is given by,

$$p = \frac{E}{c},$$

where p is the magnitude of momentum of electromagnetic wave, E is the energy of electromagnetic wave, and c is the speed of light.

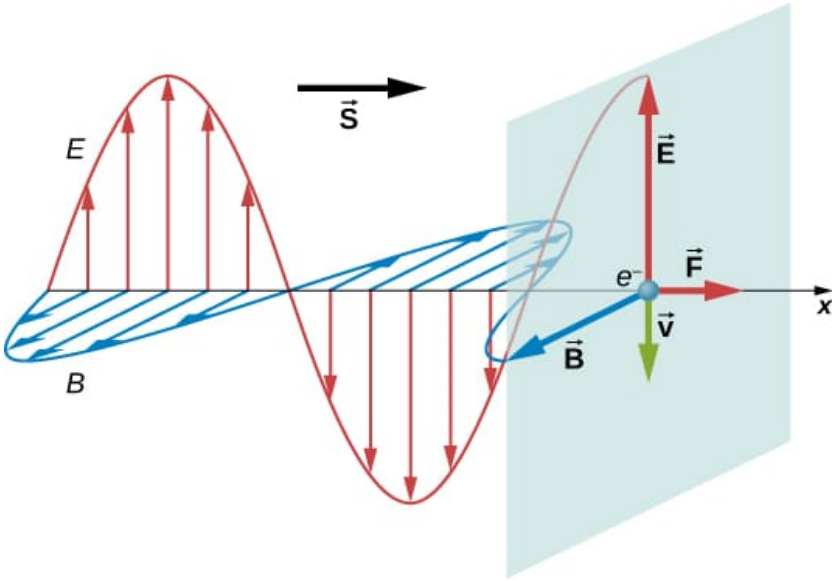


Figure 3.3.3.4: Electric and magnetic fields of an electromagnetic wave can combine to produce a force in the direction of propagation, as illustrated for the special case of electrons whose motion is highly damped by the resistance of a metal.

Maxwell predicted that an electromagnetic wave carries momentum. An object absorbing an electromagnetic wave would experience a force in the direction of propagation of the wave. The force corresponds to **radiation pressure** exerted on the object by the wave. The force would be twice as great if the radiation were reflected rather than absorbed, because the change in the momentum of electromagnetic wave itself is twice as great (and we enforce momentum conservation of the system as a whole).

Maxwell's prediction was confirmed in 1903 by Nichols and Hull by precisely measuring radiation pressures with a torsion balance. The schematic arrangement is shown in Figure 3.3.3.5. The mirrors suspended from a fiber were housed inside a glass container. Nichols and Hull were able to obtain a small measurable deflection of the mirrors from shining light on one of them. From the measured deflection, they could calculate the unbalanced force on the mirror, and obtained agreement with the predicted value of the force.

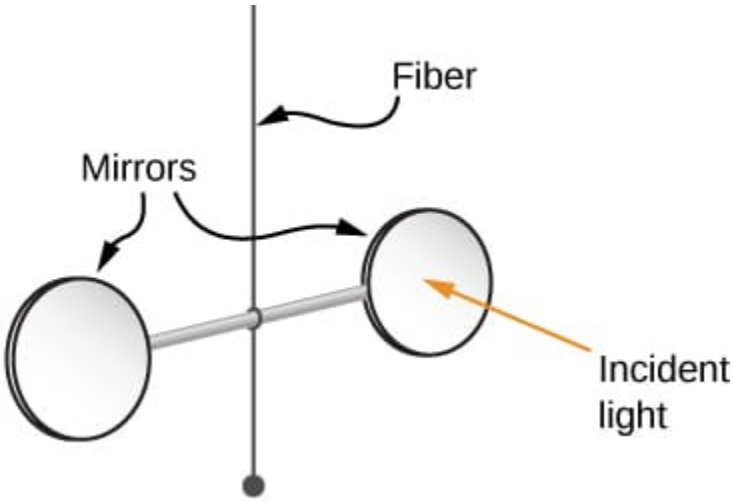


Figure 3.3.3.5: Simplified diagram of the central part of the apparatus Nichols and Hull used to precisely measure radiation pressure and confirm Maxwell's prediction.

Radiation pressure plays a role in explaining many observed astronomical phenomena, including the appearance of **comets**. Comets are basically chunks of icy material in which frozen gases and particles of rock and dust are embedded. When a comet approaches the Sun, it warms up and its surface begins to evaporate. The *coma* of the comet is the hazy area around it from the gases and dust. Some of the gases and dust form tails when they leave the comet. Notice in Figure 3.3.3.6 that a comet has *two* tails. The *ion tail* (or *gas tail* in Figure 3.3.3.6) is composed mainly of ionized gases. These ions interact electromagnetically with the solar wind, which is a continuous stream of charged particles emitted by the Sun. The force of the

solar wind on the ionized gases is strong enough that the ion tail almost always points directly away from the Sun. The second tail is composed of dust particles. Because the *dust tail* is electrically neutral, it does not interact with the solar wind. However, this tail is affected by the radiation pressure produced by the light from the Sun. Although quite small, this pressure is strong enough to cause the dust tail to be displaced from the path of the comet.

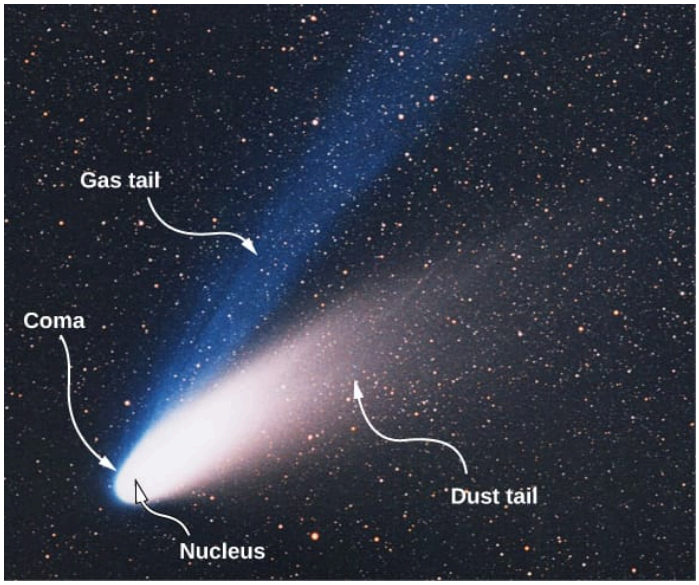


Figure 3.3.3.6: Evaporation of material being warmed by the Sun forms two tails, as shown in this photo of Comet Ison. (credit: modification of work by E. Slawik—ESO)

After Maxwell showed that light carried momentum as well as energy, a novel idea eventually emerged, initially only as science fiction. Perhaps a spacecraft with a large reflecting **solar sail** could use radiation pressure for propulsion. Such a vehicle would not have to carry fuel. It would experience a constant but small force from solar radiation, instead of the short bursts from rocket propulsion. It would accelerate slowly, but by being accelerated continuously, it would eventually reach great speeds. A spacecraft with small total mass and a sail with a large area would be necessary to obtain a usable acceleration.

The latest accomplishment on this concept is represented by the IKAROS (Interplanetary Kite-craft Accelerated by Radiation Of the Sun) project by Japan Aerospace Exploration Agency (JAXA), which was launched in May 2010. IKAROS solar sail is currently traveling through the solar system. During its six-month journey to Venus, acceleration and attitude control using the solar sail, rather than thrusters, were successfully tested. Last update from the spacecraft was in May 2014, on a ten-month orbit around the Sun. A scale model of the spacecraft is shown in Figure 3.3.3.7.

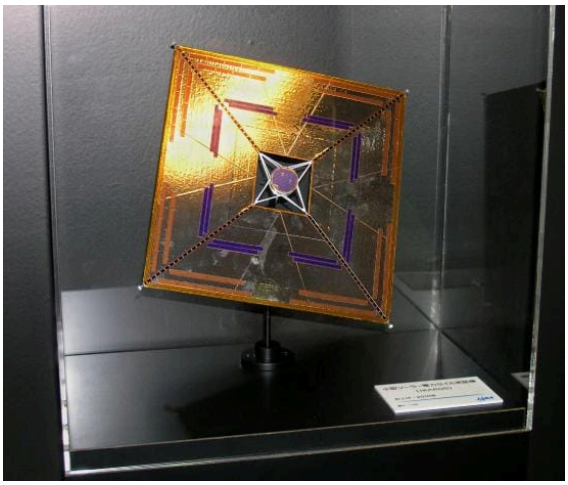


Figure 3.3.3.7: A 1:64 scale model of the IKAROS spacecraft at the 61st International Astronautical Congress in Prague, Czech Republic ("IKAROS model" By Pavel Hrdlička, Wikipedia, CC BY-SA 3.0)

Section Summary

- Electromagnetic waves are created by oscillating charges (which radiate whenever accelerated) and have the same frequency as the oscillation.
- Since the electric and magnetic fields in most electromagnetic waves are perpendicular to the direction in which the wave moves, it is ordinarily a transverse wave.
- Electromagnetic waves carry momentum and exert radiation pressure.
- The momentum of an electromagnetic wave is directly proportional to its energy.

Glossary

electric field
: a vector quantity (**E**); the lines of electric force per unit charge, moving radially outward from a positive charge and in toward a negative charge

electric field strength
: the magnitude of the electric field, denoted *E*-field

magnetic field
: a vector quantity (**B**); can be used to determine the magnetic force on a moving charged particle

magnetic field strength
: the magnitude of the magnetic field, denoted *B*-field

transverse wave
: a wave, such as an electromagnetic wave, which oscillates perpendicular to the axis along the line of travel

standing wave
: a wave that oscillates in place, with nodes where no motion happens

wavelength
: the distance from one peak to the next in a wave

amplitude
: the height, or magnitude, of an electromagnetic wave

frequency
: the number of complete wave cycles (up-down-up) passing a given point within one second (cycles/second)

resonant
: a system that displays enhanced oscillation when subjected to a periodic disturbance of the same frequency as its natural frequency

oscillate
: to fluctuate back and forth in a steady beat

radiation pressure
: pressure exerted by an electromagnetic wave on a surface

solar sail
: a spacecraft that utilizes radiation pressure due to solar radiation in its propulsion

---

1 Section 16.2: Maxwell's Equations and Electromagnetic Waves
University Physics II - Thermodynamics, Electricity, and Magnetism

---

This page titled 3.3.3: Production and Properties of Electromagnetic Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.3: Production and Properties of Electromagnetic Waves by Andrew Park is licensed CC BY 4.0.

3.3.4: The Electromagnetic Spectrum- an Overview

Learning Objectives

- List three “rules of thumb” that apply to the different frequencies along the electromagnetic spectrum.
- Explain why the higher the frequency, the shorter the wavelength of an electromagnetic wave.
- Draw a simplified electromagnetic spectrum, indicating the relative positions, frequencies, and spacing of the different types of radiation bands.

In this module we examine how electromagnetic waves are classified into categories such as radio, infrared, ultraviolet, and so on, so that we can understand some of their similarities as well as some of their differences. We will also find that there are many connections with previously discussed topics, such as wavelength and resonance. A brief overview of the production and utilization of electromagnetic waves is found in Table 3.3.4.1. Note that the vast majority of the different types of electromagnetic waves originate from atomic and/or molecular electron transitions—that is, from electrons changing their energy levels within atoms or molecules.

Table 3.3.4.1*: Electromagnetic Waves*

Type of EM wave	Production	Applications	Life sciences aspect	Issues
Radio and TV	Accelerating charges	Communications, Remote controls	MRI	Requires controls for band use
Microwaves	Accelerating charges and thermal agitation	Communications, Ovens, Radar	Deep heating	Cell phone use
Infrared	Thermal agitations and atomic/ molecular electron transitions	Thermal imaging, Heating	Absorbed by atmosphere	Greenhouse effect
Visible light	Thermal agitations and atomic/ molecular electron transitions	All pervasive	Photosynthesis, Human vision	
Ultraviolet	Thermal agitations and atomic/ molecular electron transitions	Sterilization, Cancer control	Vitamin D production	Ozone depletion, Cancer causing
X-rays	Inner atomic electron transitions and fast collisions	Medical, Security	Medical diagnosis, Cancer therapy	Cancer causing
Gamma rays	Nuclear decay	Nuclear medicine, Security	Medical diagnosis, Cancer therapy	Cancer causing, Radiation damage

CONNECTIONS: WAVES

There are many types of waves, such as water waves and even earthquakes. Among the many shared attributes of waves are propagation speed, frequency, and wavelength. These are always related by the expression $v_{\mathrm{W}} = f\lambda$. This module concentrates on EM waves, but other modules contain examples of all of these characteristics for sound waves and submicroscopic particles.

As noted before, an electromagnetic wave has a frequency and a wavelength associated with it and travels at the speed of light, or c. The relationship among these wave characteristics can be described by $v_{\mathrm{W}} = f\lambda$, where v_{W} is the propagation speed of the wave, f is the frequency, and λ is the wavelength. Here $v_{\mathrm{W}} = c$, so that for all electromagnetic waves,

$$c = f\lambda.$$

Thus, for all electromagnetic waves, the greater the frequency, the smaller the wavelength.

Figure 3.3.4.1 shows how the various types of electromagnetic waves are categorized according to their wavelengths and frequencies—that is, it shows the electromagnetic spectrum. Many of the characteristics of the various types of electromagnetic waves are related to their frequencies and wavelengths, as we shall see.

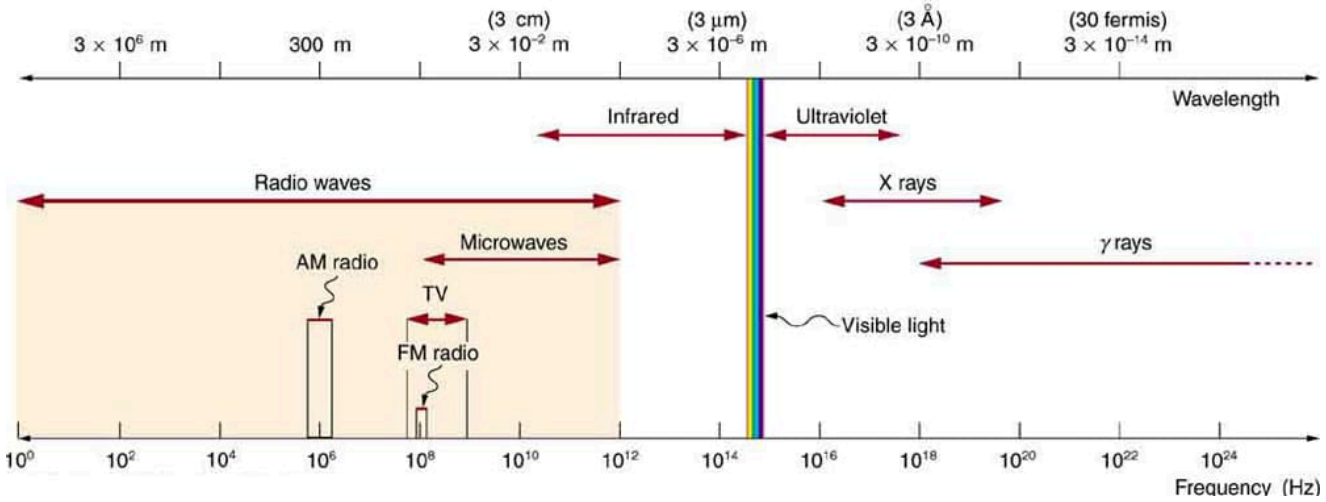


Figure 3.3.4.1: The electromagnetic spectrum, showing the major categories of electromagnetic waves. The range of frequencies and wavelengths is remarkable. The dividing line between some categories is distinct, whereas other categories overlap.

ELECTROMAGNETIC SPECTRUM: RULES OF THUMB

Three rules that apply to electromagnetic waves in general are as follows:

- High-frequency electromagnetic waves are more energetic and are more able to penetrate than low-frequency waves.
- High-frequency electromagnetic waves can carry more information per unit time than low-frequency waves.
- The shorter the wavelength of any electromagnetic wave probing a material, the smaller the detail it is possible to resolve.

Note that there are exceptions to these rules of thumb.

Transmission, Reflection, and Absorption

What happens when an electromagnetic wave impinges on a material? If the material is transparent to the particular frequency, then the wave can largely be transmitted. If the material is opaque to the frequency, then the wave can be totally reflected. The wave can also be absorbed by the material, indicating that there is some interaction between the wave and the material, such as the thermal agitation of molecules.

Of course it is possible to have partial transmission, reflection, and absorption. We normally associate these properties with visible light, but they do apply to all electromagnetic waves. What is not obvious is that something that is transparent to light may be opaque at other frequencies. For example, ordinary glass is transparent to visible light but largely opaque to ultraviolet radiation. Human skin is opaque to visible light—we cannot see through people—but transparent to X-rays.

Section Summary

- The relationship among the speed of propagation, wavelength, and frequency for any wave is given by $v_{\mathrm{W}} = f\lambda$, so that for electromagnetic waves,

$$c = f\lambda,$$

  where f is the frequency, λ is the wavelength, and c is the speed of light.
- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.

Glossary

electromagnetic spectrum
: the full range of wavelengths or frequencies of electromagnetic radiation

radio waves
: electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena

microwaves
: electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices

infrared radiation (IR)
: a region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from 0.74 μm to 300 μm

ultraviolet radiation (UV)
: electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm

visible light
: the narrow segment of the electromagnetic spectrum to which the normal human eye responds

X-ray
: invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the γ-ray range

gamma ray
: (γ ray); extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ-ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation

This page titled 3.3.4: The Electromagnetic Spectrum- an Overview is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.4: The Electromagnetic Spectrum- an Overview by Andrew Park is licensed CC BY 4.0.

3.3.5: The Electromagnetic Spectrum- Application Notes

Learning Objectives

- List and explain the characteristics and applications of different parts of the electromagnetic spectrum.

In this module, we look at the properties of different types of electromagnetic waves. Again, Figure 3.3.5.1 shows the electromagnetic spectrum. The characteristics of the various types of electromagnetic waves you will read about below are related to their frequencies and wavelengths.

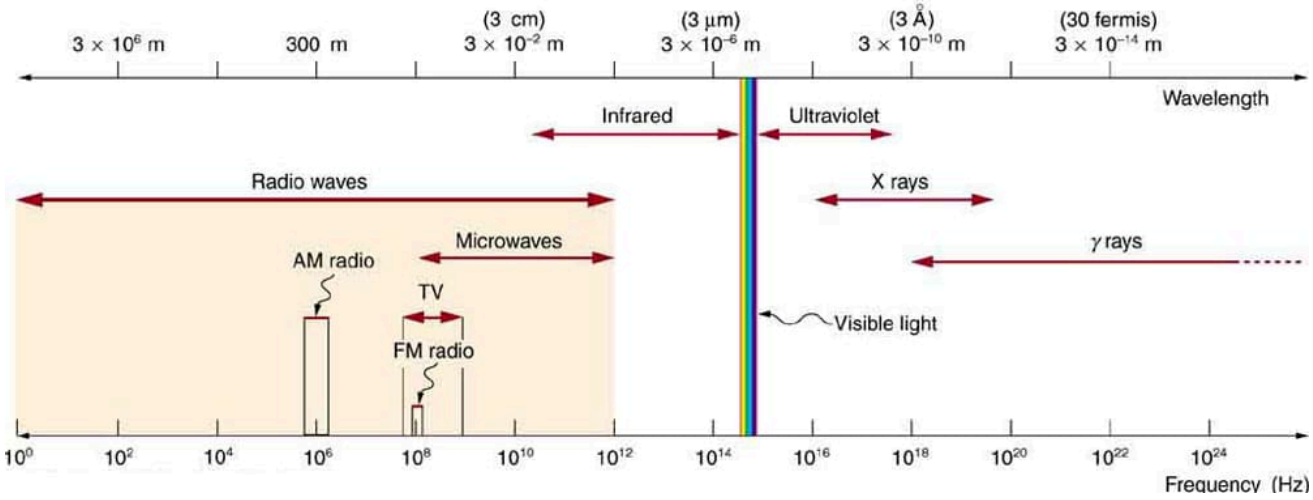


Figure 3.3.5.1: The electromagnetic spectrum, showing the major categories of electromagnetic waves. The range of frequencies and wavelengths is remarkable. The dividing line between some categories is distinct, whereas other categories overlap.

Radio and TV Waves

The broad category of **radio waves** is defined to contain any electromagnetic wave produced by currents in wires and circuits. Its name derives from their most common use as a carrier of audio information (i.e., radio). The name is applied to electromagnetic waves of similar frequencies regardless of source. Radio waves from outer space, for example, do not come from alien radio stations. They are created by many astronomical phenomena, and their study has revealed much about nature on the largest scales.

There are many uses for radio waves, and so the category is divided into many subcategories, including microwaves and those electromagnetic waves used for AM and FM radio, cellular telephones, and TV.

The lowest commonly encountered radio frequencies are produced by high-voltage AC power transmission lines at frequencies of 50 or 60 Hz. (See Figure 3.3.5.2.) These extremely long wavelength electromagnetic waves (about 6000 km!) are one means of energy loss in long-distance power transmission.



Figure 3.3.5.2: This high-voltage traction power line running to Eutingen Railway Substation in Germany radiates electromagnetic waves with very long wavelengths. (credit: Zonk43, Wikimedia Commons)

There is an ongoing controversy regarding potential health hazards associated with exposure to these electromagnetic fields (E-fields). Some people suspect that living near such transmission lines may cause a variety of illnesses, including cancer. But demographic data are either inconclusive or simply do not support the hazard theory. Recent reports that have looked at many European and American epidemiological studies have found no increase in risk for cancer due to exposure to E-fields.

Extremely low frequency (ELF) radio waves of about 1 kHz are used to communicate with submerged submarines. The ability of radio waves to penetrate salt water is related to their wavelength (much like ultrasound penetrating tissue)—the longer the wavelength, the farther they penetrate. Since salt water is a good conductor, radio waves are strongly absorbed by it, and very long wavelengths are needed to reach a submarine under the surface. (See Figure 3.3.5.3.)

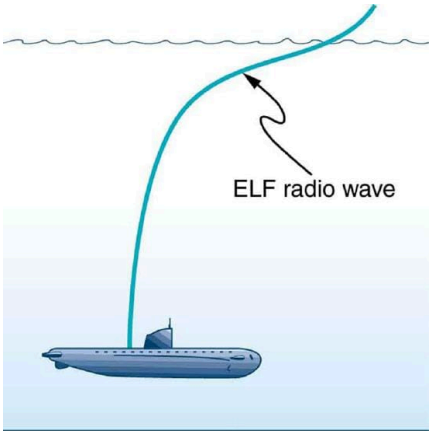


Figure 3.3.5.3: Very long wavelength radio waves are needed to reach this submarine, requiring extremely low frequency signals (ELF). Shorter wavelengths do not penetrate to any significant depth.

AM radio waves are used to carry commercial radio signals in the frequency range from 540 to 1600 kHz. The abbreviation AM stands for **amplitude modulation**, which is the method for placing information on these waves. (See Figure 3.3.5.4.) A **carrier wave** having the basic frequency of the radio station, say 1530 kHz, is varied or modulated in amplitude by an audio signal. The resulting wave has a constant frequency, but a varying amplitude.

A radio receiver tuned to have the same resonant frequency as the carrier wave can pick up the signal, while rejecting the many other frequencies impinging on its antenna. The receiver's circuitry is designed to respond to variations in amplitude of the carrier wave to replicate the original audio signal. That audio signal is amplified to drive a speaker or perhaps to be recorded.

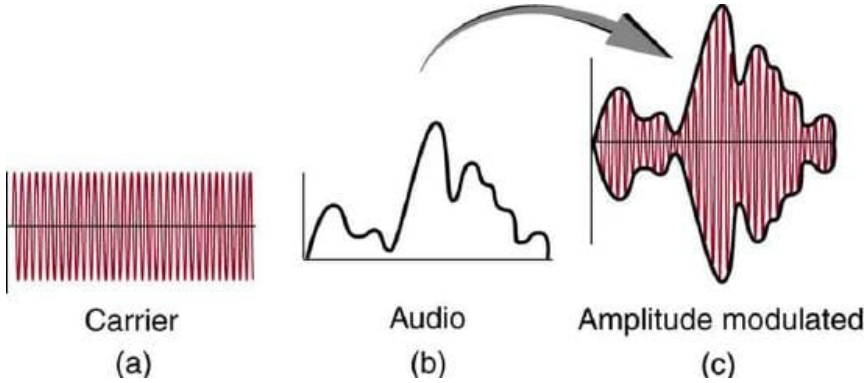


Figure 3.3.5.4: Amplitude modulation for AM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The amplitude of the carrier is modulated by the audio signal without changing its basic frequency.

FM Radio Waves

FM radio waves are also used for commercial radio transmission, but in the frequency range of 88 to 108 MHz. FM stands for **frequency modulation**, another method of carrying information. (See Figure 3.3.5.5.) Here a carrier wave having the basic frequency of the radio station, perhaps 105.1 MHz, is modulated in frequency by the audio signal, producing a wave of constant amplitude but varying frequency.

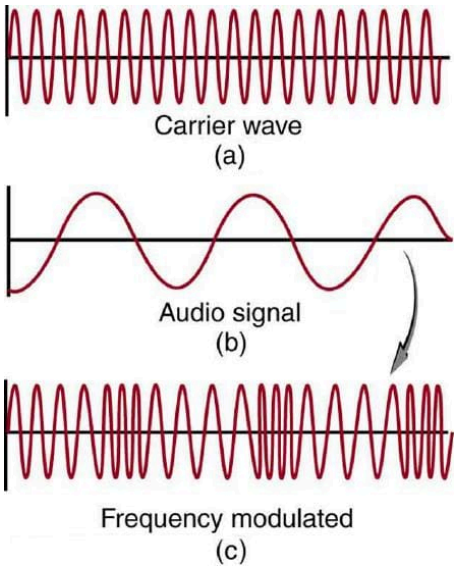


Figure 3.3.5.5: Frequency modulation for FM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The frequency of the carrier is modulated by the audio signal without changing its amplitude.

Since audible frequencies range up to 20 kHz (or 0.020 MHz) at most, the frequency of the FM radio wave can vary from the carrier by as much as 0.020 MHz. Thus the carrier frequencies of two different radio stations cannot be closer than 0.020 MHz. An FM receiver is tuned to resonate at the carrier frequency and has circuitry that responds to variations in frequency, reproducing the audio information.

FM radio is inherently less subject to noise from stray radio sources than AM radio. The reason is that amplitudes of waves add. So an AM receiver would interpret noise added onto the amplitude of its carrier wave as part of the information. An FM receiver can be made to reject amplitudes other than that of the basic carrier wave and only look for variations in frequency. It is thus easier to reject noise from FM, since noise produces a variation in amplitude.

Television is also broadcast on electromagnetic waves. Since the waves must carry a great deal of visual as well as audio information, each channel requires a larger range of frequencies than simple radio transmission. TV channels utilize frequencies in the range of 54 to 88 MHz and 174 to 222 MHz. (The entire FM radio band lies between channels 88 MHz and 174 MHz.) These TV channels are called VHF (for **very high frequency**). Other channels called UHF (for **ultra high frequency**) utilize an

even higher frequency range of 470 to 1000 MHz.

The TV video signal is AM, while the TV audio is FM. Note that these frequencies are those of free transmission with the user utilizing an old-fashioned roof antenna. Satellite dishes and cable transmission of TV occurs at significantly higher frequencies and is rapidly evolving with the use of the high-definition or HD format.

Example 3.3.5.1: Calculating Wavelengths of Radio Waves

Calculate the wavelengths of a 1530-kHz AM radio signal, a 105.1-MHz FM radio signal, and a 1.90-GHz cell phone signal.

Strategy

The relationship between wavelength and frequency is $c = f\lambda$, where $c = 3.00 \times 10^8 \text{ m/s}$ is the speed of light (the speed of light is only very slightly smaller in air than it is in a vacuum). We can rearrange this equation to find the wavelength for all three frequencies.

Solution

Rearranging gives

$$\lambda = \frac{c}{f}.$$

(a) For the $f = 1530 \text{ kHz}$ AM radio signal, then,

$$\begin{aligned}\lambda &= \frac{3.00 \times 10^8 \text{ m/s}}{1530 \times 10^3 \text{cycles/s}} \\ &= 196 \text{ m}.\end{aligned}$$

(b) For the $f = 105.1 \text{ MHz}$ FM radio signal,

$$\begin{aligned}\lambda &= \frac{3.00 \times 10^8 \text{ m/s}}{105.1 \times 10^6 \text{cycles/s}} \\ &= 2.85 \text{ m}.\end{aligned}$$

(c) And for the $f = 1.90 \text{ GHz}$ cell phone,

$$\begin{aligned}\lambda &= \frac{3.00 \times 10^8 \text{ m/s}}{1.90 \times 10^9 \text{cycles/s}} \\ &= 0.158 \text{ m}.\end{aligned}$$

Discussion

These wavelengths are consistent with the spectrum in Figure 3.3.5.1. The wavelengths are also related to other properties of these electromagnetic waves, as we shall see.

The wavelengths found in the preceding example are representative of AM, FM, and cell phones, and account for some of the differences in how they are broadcast and how well they travel. The most efficient length for a linear antenna, such as discussed in "Production of Electromagnetic Waves", is $\lambda/2$, half the wavelength of the electromagnetic wave. Thus a very large antenna is needed to efficiently broadcast typical AM radio with its carrier wavelengths on the order of hundreds of meters.

One benefit to these long AM wavelengths is that they can go over and around rather large obstacles (like buildings and hills), just as ocean waves can go around large rocks. FM and TV are best received when there is a line of sight between the broadcast antenna and receiver, and they are often sent from very tall structures. FM, TV, and mobile phone antennas themselves are much smaller than those used for AM, but they are elevated to achieve an unobstructed line of sight. (See Figure 3.3.5.6.)



Figure 3.3.5.6: (a) A large tower is used to broadcast TV signals. The actual antennas are small structures on top of the tower—they are placed at great heights to have a clear line of sight over a large broadcast area. (credit: Ozizo, Wikimedia Commons) (b) The NTT Dokomo mobile phone tower at Tokorozawa City, Japan. (credit: tokoroten, Wikimedia Commons)

Radio Wave Interference

Astronomers and astrophysicists collect signals from outer space using electromagnetic waves. A common problem for astrophysicists is the "pollution" from electromagnetic radiation pervading our surroundings from communication systems in general. Even everyday gadgets like our car keys having the facility to lock car doors remotely and being able to turn TVs on and off using remotes involve radio-wave frequencies. In order to prevent interference between all these electromagnetic signals, strict regulations are drawn up for different organizations to utilize different radio frequency bands.

One reason why we are sometimes asked to switch off our mobile phones (operating in the range of 1.9 GHz) on airplanes and in hospitals is that important communications or medical equipment often uses similar radio frequencies and their operation can be affected by frequencies used in the communication devices.

For example, radio waves used in magnetic resonance imaging (MRI) have frequencies on the order of 100 MHz, although this varies significantly depending on the strength of the magnetic field used and the nuclear type being scanned. MRI is an important medical imaging and research tool, producing highly detailed two- and three-dimensional images. Radio waves are broadcast, absorbed, and reemitted in a resonance process that is sensitive to the density of nuclei (usually protons or hydrogen nuclei).

The wavelength of 100-MHz radio waves is 3 m, yet using the sensitivity of the resonant frequency to the magnetic field strength, details smaller than a millimeter can be imaged. This is a good example of an exception to a rule of thumb (in this case, the rubric that details much smaller than the probe's wavelength cannot be detected). The intensity of the radio waves used in MRI presents little or no hazard to human health.

Microwaves

Microwaves are the highest-frequency electromagnetic waves that can be produced by currents in macroscopic circuits and devices. Microwave frequencies range from about 10^9 Hz to the highest practical LC resonance at nearly 10^{12} Hz. Since they have high frequencies, their wavelengths are short compared with those of other radio waves—hence the name "microwave."

Microwaves can also be produced by atoms and molecules. They are, for example, a component of electromagnetic radiation generated by **thermal agitation**. The thermal motion of atoms and molecules in any object at a temperature above absolute zero causes them to emit and absorb radiation.

Since it is possible to carry more information per unit time on high frequencies, microwaves are quite suitable for communications. Most satellite-transmitted information is carried on microwaves, as are land-based long-distance transmissions. A clear line of sight between transmitter and receiver is needed because of the short wavelengths involved.

Radar is a common application of microwaves that was first developed in World War II. By detecting and timing microwave echoes, radar systems can determine the distance to objects as diverse as clouds and aircraft. A Doppler shift in the radar echo can be used to determine the speed of a car or the intensity of a rainstorm. Sophisticated radar systems are used to map the

Earth and other planets, with a resolution limited by wavelength. (See Figure 3.3.5.7.) The shorter the wavelength of any probe, the smaller the detail it is possible to observe.

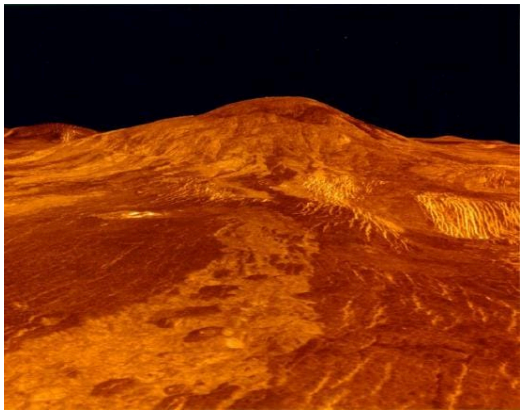


Figure 3.3.5.7: An image of Sif Mons with lava flows on Venus, based on Magellan synthetic aperture radar data combined with radar altimetry to produce a three-dimensional map of the surface. The Venusian atmosphere is opaque to visible light, but not to the microwaves that were used to create this image. (credit: NSSDC, NASA/JPL)

Heating with Microwaves

How does the ubiquitous microwave oven produce microwaves electronically, and why does food absorb them preferentially? Microwaves at a frequency of 2.45 GHz are produced by accelerating electrons. The microwaves are then used to induce an alternating electric field in the oven.

Water and some other constituents of food have a slightly negative charge at one end and a slightly positive charge at one end (called polar molecules). The range of microwave frequencies is specially selected so that the polar molecules, in trying to keep orienting themselves with the electric field, absorb these energies and increase their temperatures—called dielectric heating.

The energy thereby absorbed results in thermal agitation heating food and not the plate, which does not contain water. Hot spots in the food are related to constructive and destructive interference patterns. Rotating antennas and food turntables help spread out the hot spots.

Another use of microwaves for heating is within the human body. Microwaves will penetrate more than shorter wavelengths into tissue and so can accomplish “deep heating” (called microwave diathermy). This is used for treating muscular pains, spasms, tendonitis, and rheumatoid arthritis.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—MICROWAVE OVENS

1. Look at the door of a microwave oven. Describe the structure of the door. Why is there a metal grid on the door? How does the size of the holes in the grid compare with the wavelengths of microwaves used in microwave ovens? What is this wavelength?
2. Place a glass of water (about 250 ml) in the microwave and heat it for 30 seconds. Measure the temperature gain (the ΔT). Assuming that the power output of the oven is 1000 W, calculate the efficiency of the heat-transfer process.
3. Remove the rotating turntable or moving plate and place a cup of water in several places along a line parallel with the opening. Heat for 30 seconds and measure the ΔT for each position. Do you see cases of destructive interference?

Microwaves generated by atoms and molecules far away in time and space can be received and detected by electronic circuits. Deep space acts like a blackbody with a 2.7 K temperature, radiating most of its energy in the microwave frequency range. In 1964, Penzias and Wilson detected this radiation and eventually recognized that it was the radiation of the Big Bang’s cooled remnants.

Infrared Radiation

The microwave and infrared regions of the electromagnetic spectrum overlap (see Figure 3.3.5.1). **Infrared radiation** is generally produced by thermal motion and the vibration and rotation of atoms and molecules. Electronic transitions in atoms and molecules can also produce infrared radiation.

The range of infrared frequencies extends up to the lower limit of visible light, just below red. In fact, infrared means “below red.” Frequencies at its upper limit are too high to be produced by accelerating electrons in circuits, but small systems, such as atoms and molecules, can vibrate fast enough to produce these waves.

Water molecules rotate and vibrate particularly well at infrared frequencies, emitting and absorbing them so efficiently that the emissivity for skin is $e = 0.97$ in the infrared. Night-vision scopes can detect the infrared emitted by various warm objects, including humans, and convert it to visible light.

We can examine radiant heat transfer from a house by using a camera capable of detecting infrared radiation. Reconnaissance satellites can detect buildings, vehicles, and even individual humans by their infrared emissions, whose power radiation is proportional to the fourth power of the absolute temperature. More mundanely, we use infrared lamps, some of which are called quartz heaters, to preferentially warm us because we absorb infrared better than our surroundings.

The Sun radiates like a nearly perfect blackbody (that is, it has $e = 1$), with a 6000 K surface temperature. About half of the solar energy arriving at the Earth is in the infrared region, with most of the rest in the visible part of the spectrum, and a relatively small amount in the ultraviolet. On average, 50 percent of the incident solar energy is absorbed by the Earth.

The relatively constant temperature of the Earth is a result of the energy balance between the incoming solar radiation and the energy radiated from the Earth. Most of the infrared radiation emitted from the Earth is absorbed by CO_2 and H_2O in the atmosphere and then radiated back to Earth or into outer space. This radiation back to Earth is known as the greenhouse effect, and it maintains the surface temperature of the Earth about $40^{\circ}C$ higher than it would be if there is no absorption. Some scientists think that the increased concentration of CO_2 and other greenhouse gases in the atmosphere, resulting from increases in fossil fuel burning, has increased global average temperatures.

Visible Light

Visible light is the narrow segment of the electromagnetic spectrum to which the normal human eye responds. Visible light is produced by vibrations and rotations of atoms and molecules, as well as by electronic transitions within atoms and molecules. The receivers or detectors of light largely utilize electronic transitions. We say the atoms and molecules are excited when they absorb and relax when they emit through electronic transitions.

Figure 3.3.5.8 shows this part of the spectrum, together with the colors associated with particular pure wavelengths. We usually refer to visible light as having wavelengths of between 400 nm and 750 nm. (The retina of the eye actually responds to the lowest ultraviolet frequencies, but these do not normally reach the retina because they are absorbed by the cornea and lens of the eye.)

Red light has the lowest frequencies and longest wavelengths, while violet has the highest frequencies and shortest wavelengths. Blackbody radiation from the Sun peaks in the visible part of the spectrum but is more intense in the red than in the violet, making the Sun yellowish in appearance.

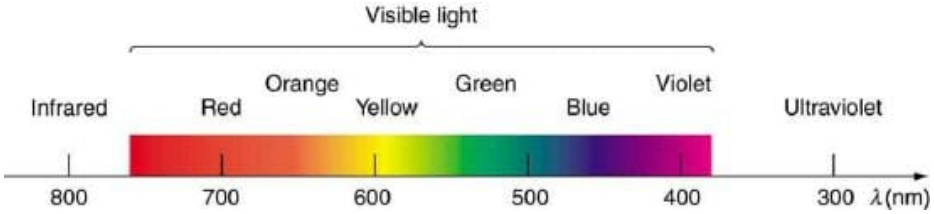


Figure 3.3.5.8: A small part of the electromagnetic spectrum that includes its visible components. The divisions between infrared, visible, and ultraviolet are not perfectly distinct, nor are those between the seven rainbow colors.

Living things—plants and animals—have evolved to utilize and respond to parts of the electromagnetic spectrum they are embedded in. Visible light is the most predominant and we enjoy the beauty of nature through visible light. Plants are more selective. Photosynthesis makes use of parts of the visible spectrum to make sugars.

Optics is the study of the behavior of visible light and other forms of electromagnetic waves. Optics falls into two distinct categories. When electromagnetic radiation, such as visible light, interacts with objects that are large compared with its wavelength, its motion can be represented by straight lines like rays. Ray optics is the study of such situations and includes lenses and mirrors.

When electromagnetic radiation interacts with objects about the same size as the wavelength or smaller, its wave nature becomes apparent. For example, observable detail is limited by the wavelength, and so visible light can never detect individual atoms, because they are so much smaller than its wavelength. Physical or wave optics is the study of such situations and includes all wave characteristics.

TAKE-HOME EXPERIMENT: COLORS THAT MATCH

When you light a match you see largely orange light; when you light a gas stove you see blue light. Why are the colors different? What other colors are present in these?

Ultraviolet Radiation

Ultraviolet means “above violet.” The electromagnetic frequencies of **ultraviolet radiation (UV)** extend upward from violet, the highest-frequency visible light. Ultraviolet is also produced by atomic and molecular motions and electronic transitions. The wavelengths of ultraviolet extend from 400 nm down to about 10 nm at its highest frequencies, which overlap with the lowest X-ray frequencies. It was recognized as early as 1801 by Johann Ritter that the solar spectrum had an invisible component beyond the violet range.

Solar UV radiation is broadly subdivided into three regions: UV-A (320–400 nm), UV-B (290–320 nm), and UV-C (220–290 nm), ranked from long to shorter wavelengths (from smaller to larger energies). Most UV-B and all UV-C is absorbed by ozone (O_3) molecules in the upper atmosphere. Consequently, 99% of the solar UV radiation reaching the Earth’s surface is UV-A.

Human Exposure to UV Radiation

It is largely exposure to UV-B that causes skin cancer. It is estimated that as many as 20% of adults will develop skin cancer over the course of their lifetime. Again, treatment is often successful if caught early. Despite very little UV-B reaching the Earth’s surface, there are substantial increases in skin-cancer rates in countries such as Australia, indicating how important it is that UV-B and UV-C continue to be absorbed by the upper atmosphere.

All UV radiation can damage collagen fibers, resulting in an acceleration of the aging process of skin and the formation of wrinkles. Because there is so little UV-B and UV-C reaching the Earth’s surface, sunburn is caused by large exposures, and skin cancer from repeated exposure. Some studies indicate a link between overexposure to the Sun when young and melanoma later in life.

The tanning response is a defense mechanism in which the body produces pigments to absorb future exposures in inert skin layers above living cells. Basically UV-B radiation excites DNA molecules, distorting the DNA helix, leading to mutations and the possible formation of cancerous cells.

Repeated exposure to UV-B may also lead to the formation of cataracts in the eyes—a cause of blindness among people living in the equatorial belt where medical treatment is limited. Cataracts, clouding in the eye’s lens and a loss of vision, are age related; 60% of those between the ages of 65 and 74 will develop cataracts. However, treatment is easy and successful, as one replaces the lens of the eye with a plastic lens. Prevention is important. Eye protection from UV is more effective with plastic sunglasses than those made of glass.

A major acute effect of extreme UV exposure is the suppression of the immune system, both locally and throughout the body.

Low-intensity ultraviolet is used to sterilize haircutting implements, implying that the energy associated with ultraviolet is deposited in a manner different from lower-frequency electromagnetic waves. (Actually this is true for all electromagnetic waves with frequencies greater than visible light.)

Flash photography is generally not allowed of precious artworks and colored prints because the UV radiation from the flash can

cause photo-degradation in the artworks. Often artworks will have an extra-thick layer of glass in front of them, which is especially designed to absorb UV radiation.

UV Light and the Ozone Layer

If all of the Sun's ultraviolet radiation reached the Earth's surface, there would be extremely grave effects on the biosphere from the severe cell damage it causes. However, the layer of ozone (O_3) in our upper atmosphere (10 to 50 km above the Earth) protects life by absorbing most of the dangerous UV radiation.

Unfortunately, today we are observing a depletion in ozone concentrations in the upper atmosphere. This depletion has led to the formation of an "ozone hole" in the upper atmosphere. The hole is more centered over the southern hemisphere, and changes with the seasons, being largest in the spring. This depletion is attributed to the breakdown of ozone molecules by refrigerant gases called chlorofluorocarbons (CFCs).

The UV radiation helps dissociate the CFC's, releasing highly reactive chlorine (Cl) atoms, which catalyze the destruction of the ozone layer. For example, the reaction of $CFCl_3$ with a photon of light $(h\upsilon)$ can be written as:

$$CFCl_3 + h\upsilon \rightarrow CFCl_2 + Cl.$$

The Cl atom then catalyzes the breakdown of ozone as follows:

$$Cl + O_3 \rightarrow ClO + O_2 \text{ and } ClO + O_3 \rightarrow Cl + 2O_2.$$

A single chlorine atom could destroy ozone molecules for up to two years before being transported down to the surface. The CFCs are relatively stable and will contribute to ozone depletion for years to come. CFCs are found in refrigerants, air conditioning systems, foams, and aerosols.

International concern over this problem led to the establishment of the "Montreal Protocol" agreement (1987) to phase out CFC production in most countries. However, developing-country participation is needed if worldwide production and elimination of CFCs is to be achieved. Probably the largest contributor to CFC emissions today is India. But the protocol seems to be working, as there are signs of an ozone recovery. (See Figure 3.3.5.9.)

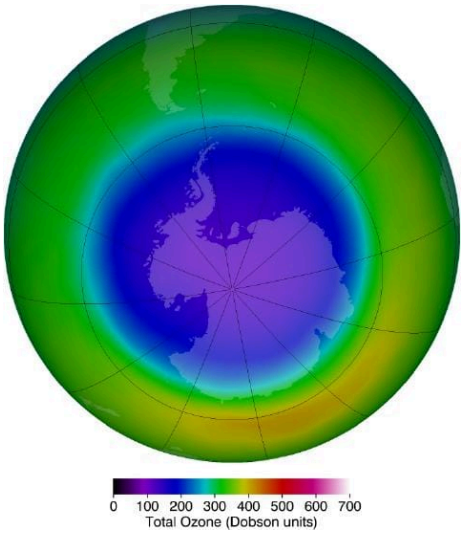


Figure 3.3.5.9: This map of ozone concentration over Antarctica in October 2011 shows severe depletion suspected to be caused by CFCs. Less dramatic but more general depletion has been observed over northern latitudes, suggesting the effect is global. With less ozone, more ultraviolet radiation from the Sun reaches the surface, causing more damage. (credit: NASA Ozone Watch)

Benefits of UV Light

Besides the adverse effects of ultraviolet radiation, there are also benefits of exposure in nature and uses in technology. Vitamin D production in the skin (epidermis) results from exposure to UVB radiation, generally from sunlight. A number of studies indicate lack of vitamin D can result in the development of a range of cancers (prostate, breast, colon), so a certain amount of UV exposure

is helpful. Lack of vitamin D is also linked to osteoporosis. Exposures (with no sunscreen) of 10 minutes a day to arms, face, and legs might be sufficient to provide the accepted dietary level. However, in the winter time north of about 37° latitude, most UVB gets blocked by the atmosphere.

UV radiation is used in the treatment of infantile jaundice and in some skin conditions. It is also used in sterilizing workspaces and tools, and killing germs in a wide range of applications. It is also used as an analytical tool to identify substances.

When exposed to ultraviolet, some substances, such as minerals, glow in characteristic visible wavelengths, a process called fluorescence. So-called black lights emit ultraviolet to cause posters and clothing to fluoresce in the visible. Ultraviolet is also used in special microscopes to detect details smaller than those observable with longer-wavelength visible-light microscopes.

X-Rays

In the 1850s, scientists (such as Faraday) began experimenting with high-voltage electrical discharges in tubes filled with rarefied gases. It was later found that these discharges created an invisible, penetrating form of very high frequency electromagnetic radiation. This radiation was called an **X-ray**, because its identity and nature were unknown.

THINGS GREAT AND SMALL: A SUBMICROSCOPIC VIEW OF X-RAY PRODUCTION

X-rays can be created in a high-voltage discharge. They are emitted in the material struck by electrons in the discharge current. There are two mechanisms by which the electrons create X-rays.

The first method is illustrated in Figure $3.3.5.10$. An electron is accelerated in an evacuated tube by a high positive voltage. The electron strikes a metal plate (e.g., copper) and produces X-rays. Since this is a high-voltage discharge, the electron gains sufficient energy to ionize the atom.

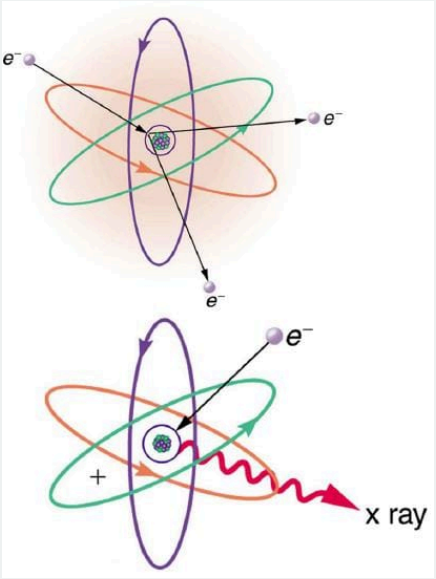


Figure $3.3.5.10$: Artist's conception of an electron ionizing an atom followed by the recapture of an electron and emission of an X-ray. An energetic electron strikes an atom and knocks an electron out of one of the orbits closest to the nucleus. Later, the atom captures another electron, and the energy released by its fall into a low orbit generates a high-energy EM wave called an X-ray.

In the case shown, an inner-shell electron (one in an orbit relatively close to and tightly bound to the nucleus) is ejected. A short time later, another electron is captured and falls into the orbit in a single great plunge. The energy released by this fall is given to an EM wave known as an X-ray. Since the orbits of the atom are unique to the type of atom, the energy of the X-ray is characteristic of the atom, hence the name characteristic X-ray.

The second method by which an energetic electron creates an X-ray when it strikes a material is illustrated in Figure $3.3.5.11$. The electron interacts with charges in the material as it penetrates. These collisions transfer kinetic energy from the electron to the electrons and atoms in the material.

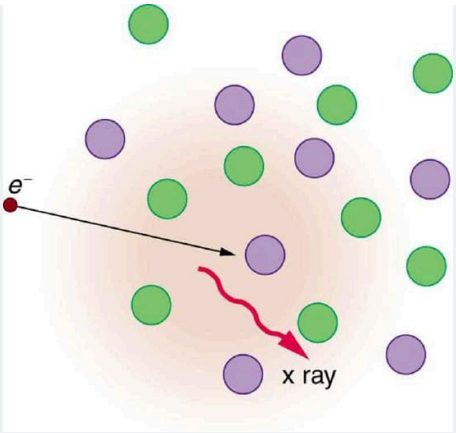


Figure 3.3.5.11: Artist's conception of an electron being slowed by collisions in a material and emitting X-ray radiation. This energetic electron makes numerous collisions with electrons and atoms in a material it penetrates. An accelerated charge radiates EM waves, a second method by which X-rays are created.

A loss of kinetic energy implies an acceleration, in this case decreasing the electron's velocity. Whenever a charge is accelerated, it radiates EM waves. Given the high energy of the electron, these EM waves can have high energy. We call them X-rays. Since the process is random, a broad spectrum of X-ray energy is emitted that is more characteristic of the electron energy than the type of material the electron encounters. Such EM radiation is called "bremsstrahlung" (German for "braking radiation").

As described above, there are two methods by which X-rays are created—both are submicroscopic processes and can be caused by high-voltage discharges. While the low-frequency end of the X-ray range overlaps with the ultraviolet, X-rays extend to much higher frequencies (and energies).

X-rays have adverse effects on living cells similar to those of ultraviolet radiation, and they have the additional liability of being more penetrating, affecting more than the surface layers of cells. Cancer and genetic defects can be induced by exposure to X-rays. Because of their effect on rapidly dividing cells, X-rays can also be used to treat and even cure cancer.

The widest use of X-rays is for imaging objects that are opaque to visible light, such as the human body or aircraft parts. In humans, the risk of cell damage is weighed carefully against the benefit of the diagnostic information obtained. However, questions have risen in recent years as to accidental overexposure of some people during CT scans—a mistake at least in part due to poor monitoring of radiation dose.

The ability of X-rays to penetrate matter depends on density, and so an X-ray image can reveal very detailed density information. Figure 3.3.5.12 shows an example of the simplest type of X-ray image, an X-ray shadow on film. The amount of information in a simple X-ray image is impressive, but more sophisticated techniques, such as CT scans, can reveal three-dimensional information with details smaller than a millimeter.

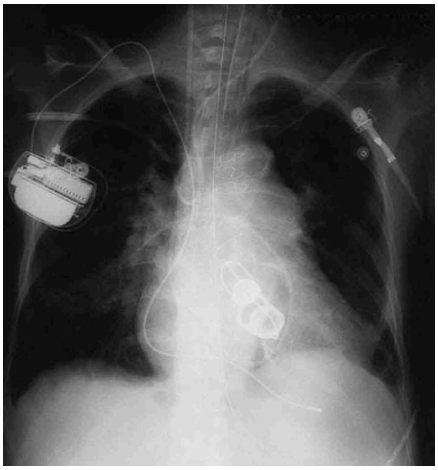


Figure 3.3.5.12: This shadow X-ray image shows many interesting features, such as artificial heart valves, a pacemaker, and the wires used to close the sternum. (credit: P. P. Urone)

The use of X-ray technology in medicine is called radiology—an established and relatively cheap tool in comparison to more sophisticated technologies. Consequently, X-rays are widely available and used extensively in medical diagnostics. During World War I, mobile X-ray units, advocated by Madame Marie Curie, were used to diagnose soldiers.

Because they can have wavelengths less than 0.01 nm, X-rays can be scattered (a process called X-ray diffraction) to detect the shape of molecules and the structure of crystals. X-ray diffraction was crucial to Crick, Watson, and Wilkins in the determination of the shape of the double-helix DNA molecule.

X-rays are also used as a precise tool for trace-metal analysis in X-ray induced fluorescence, in which the energy of the X-ray emissions are related to the specific types of elements and amounts of materials present.

Gamma Rays

Soon after nuclear radioactivity was first detected in 1896, it was found that at least three distinct types of radiation were being emitted. The most penetrating nuclear radiation was called a **gamma ray (γ ray)** (again a name given because its identity and character were unknown), and it was later found to be an extremely high frequency electromagnetic wave.

In fact, γ rays are any electromagnetic radiation emitted by a nucleus. This can be from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ-ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation.

Gamma rays have characteristics identical to X-rays of the same frequency—they differ only in source. At higher frequencies, γ rays are more penetrating and more damaging to living tissue. They have many of the same uses as X-rays, including cancer therapy. Gamma radiation from radioactive materials is used in nuclear medicine.

Figure 3.3.5.13 shows a medical image based on γ rays. Food spoilage can be greatly inhibited by exposing it to large doses of γ radiation, thereby obliterating responsible microorganisms. Damage to food cells through irradiation occurs as well, and the long-term hazards of consuming radiation-preserved food are unknown and controversial for some groups. Both X-ray and γ-ray technologies are also used in scanning luggage at airports.



Figure 3.3.5.13: This is an image of the γ rays emitted by nuclei in a compound that is concentrated in the bones and eliminated through the kidneys. Bone cancer is evidenced by nonuniform concentration in similar structures. For example, some ribs are darker than others. (credit: P. P. Urone)

Detecting Electromagnetic Waves from Space

A final note on star gazing. The entire electromagnetic spectrum is used by researchers for investigating stars, space, and time. As noted earlier, Penzias and Wilson detected microwaves to identify the background radiation originating from the Big Bang. Radio telescopes such as the Arecibo Radio Telescope in Puerto Rico and Parkes Observatory in Australia were designed to detect radio waves.

Infrared telescopes need to have their detectors cooled by liquid nitrogen to be able to gather useful signals. Since infrared radiation is predominantly from thermal agitation, if the detectors were not cooled, the vibrations of the molecules in the antenna would be stronger than the signal being collected.

The most famous of these infrared sensitive telescopes is the James Clerk Maxwell Telescope in Hawaii. The earliest telescopes, developed in the seventeenth century, were optical telescopes, collecting visible light. Telescopes in the ultraviolet, X-ray, and γ -ray regions are placed outside the atmosphere on satellites orbiting the Earth.

The Hubble Space Telescope (launched in 1990) gathers ultraviolet radiation as well as visible light. In the X-ray region, there is the Chandra X-ray Observatory (launched in 1999), and in the γ-ray region, there is the new Fermi Gamma-ray Space Telescope (launched in 2008—taking the place of the Compton Gamma Ray Observatory, 1991–2000.).

Section Summary

- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.
- Any electromagnetic wave produced by currents in wires is classified as a radio wave, the lowest frequency electromagnetic waves. Radio waves are divided into many types, depending on their applications, ranging up to microwaves at their highest frequencies.
- Infrared radiation lies below visible light in frequency and is produced by thermal motion and the vibration and rotation of atoms and molecules. Infrared's lower frequencies overlap with the highest-frequency microwaves.
- Visible light is largely produced by electronic transitions in atoms and molecules, and is defined as being detectable by the human eye. Its colors vary with frequency, from red at the lowest to violet at the highest.
- Ultraviolet radiation starts with frequencies just above violet in the visible range and is produced primarily by electronic transitions in atoms and molecules.
- X-rays are created in high-voltage discharges and by electron bombardment of metal targets. Their lowest frequencies overlap the ultraviolet range but extend to much higher values, overlapping at the high end with gamma rays.

- Gamma rays are nuclear in origin and are defined to include the highest-frequency electromagnetic radiation of any type.

Glossary

electromagnetic spectrum
: the full range of wavelengths or frequencies of electromagnetic radiation

radio waves
: electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena

microwaves
: electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices

thermal agitation
: the thermal motion of atoms and molecules in any object at a temperature above absolute zero, which causes them to emit and absorb radiation

radar
: a common application of microwaves. Radar can determine the distance to objects as diverse as clouds and aircraft, as well as determine the speed of a car or the intensity of a rainstorm

infrared radiation (IR)
: a region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from 0.74 μm to 300 μm

ultraviolet radiation (UV)
: electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm

visible light
: the narrow segment of the electromagnetic spectrum to which the normal human eye responds

amplitude modulation (AM)
: a method for placing information on electromagnetic waves by modulating the amplitude of a carrier wave with an audio signal, resulting in a wave with constant frequency but varying amplitude

extremely low frequency (ELF)
: electromagnetic radiation with wavelengths usually in the range of 0 to 300 Hz, but also about 1kHz

carrier wave
: an electromagnetic wave that carries a signal by modulation of its amplitude or frequency

frequency modulation (FM)
: a method of placing information on electromagnetic waves by modulating the frequency of a carrier wave with an audio signal, producing a wave of constant amplitude but varying frequency

TV
: video and audio signals broadcast on electromagnetic waves

very high frequency (VHF)
: TV channels utilizing frequencies in the two ranges of 54 to 88 MHz and 174 to 222 MHz

ultra-high frequency (UHF)
: TV channels in an even higher frequency range than VHF, of 470 to 1000 MHz

X-ray
: invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the γ-ray range

gamma ray

(γ ray); extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ-ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation

This page titled 3.3.5: The Electromagnetic Spectrum- Application Notes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.5: The Electromagnetic Spectrum- Application Notes by Andrew Park is licensed CC BY 4.0.

3.3.6: Reflection

Learning Objectives

- Explain reflection of light from polished and rough surfaces.

Whenever we look into a mirror, or squint at sunlight glinting from a lake, we are seeing a reflection. When you look at this page, too, you are seeing light reflected from it. Large telescopes use reflection to form an image of stars and other astronomical objects.

The law of reflection is illustrated in Figure 3.3.6.1, which also shows how the angles are measured relative to the perpendicular to the surface at the point where the light ray strikes. We expect to see reflections from smooth surfaces, but Figure 3.3.6.2 illustrates how a rough surface reflects light. Since the light strikes different parts of the surface at different angles, it is reflected in many different directions, or diffused. Diffused light is what allows us to see a sheet of paper from any angle, as illustrated in Figure 3.3.6.3. Many objects, such as people, clothing, leaves, and walls, have rough surfaces and can be seen from all sides. A mirror, on the other hand, has a smooth surface (compared with the wavelength of light) and reflects light at specific angles, as illustrated in Figure 3.3.6.4. When the moon reflects from a lake, as shown in Figure 3.3.6.5, a combination of these effects takes place.

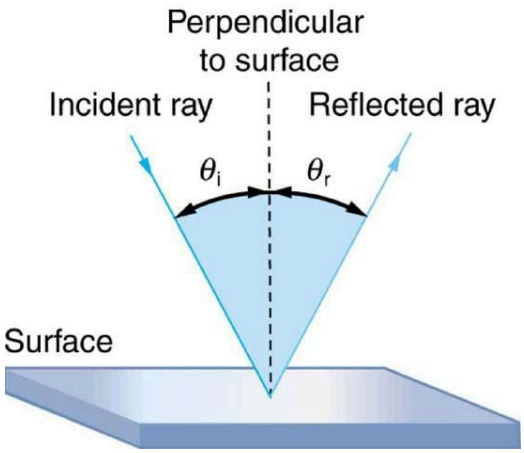


Figure 3.3.6.1: The law of reflection states that the angle of reflection equals the angle of incidence— $\theta_r = \theta_i$. The angles are measured relative to the perpendicular to the surface at the point where the ray strikes the surface.

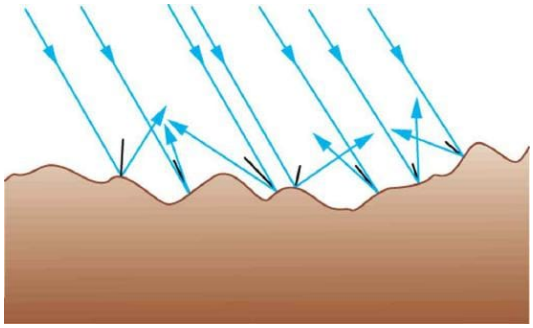


Figure 3.3.6.2: Light is diffused when it reflects from a rough surface. Here many parallel rays are incident, but they are reflected at many different angles since the surface is rough.

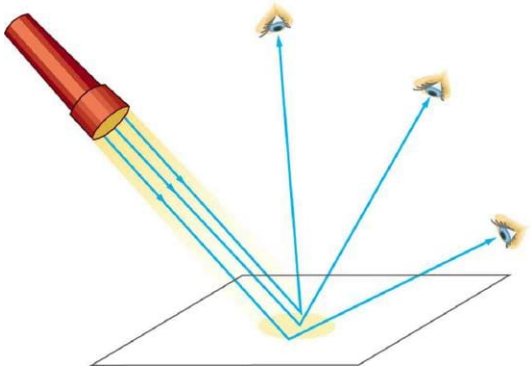


Figure 3.3.6.3: When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because its surface is rough and diffuses the light.

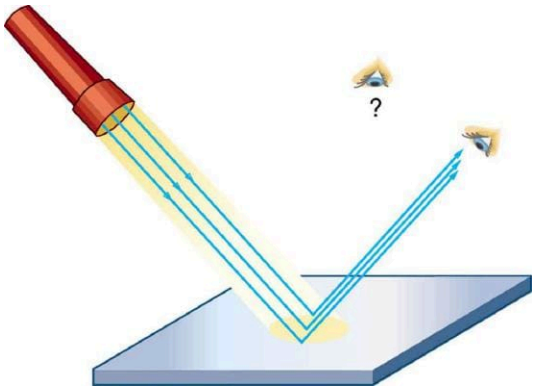


Figure 3.3.6.4: A mirror illuminated by many parallel rays reflects them in only one direction, since its surface is very smooth. Only the observer at a particular angle will see the reflected light.



Figure 3.3.6.5: Moonlight is spread out when it is reflected by the lake, since the surface is shiny but uneven. (credit: Diego Torres Silvestre, Flickr)

The law of reflection is very simple: The angle of reflection equals the angle of incidence.

THE LAW OF REFLECTION

The angle of reflection equals the angle of incidence.

When we see ourselves in a mirror, it appears that our image is actually behind the mirror. This is illustrated in Figure 3.3.6.6. We see the light coming from a direction determined by the law of reflection. The angles are such that our image is exactly the same distance behind the mirror as we stand away from the mirror. If the mirror is on the wall of a room, the images in it are all behind the mirror, which can make the room seem bigger. Although these mirror images make objects appear to be where they cannot be (like behind a solid wall), the images are not figments of our imagination. Mirror images can be photographed and videotaped

by instruments and look just as they do with our eyes (optical instruments themselves). The precise manner in which images are formed by mirrors and lenses will be treated in later sections of this chapter.

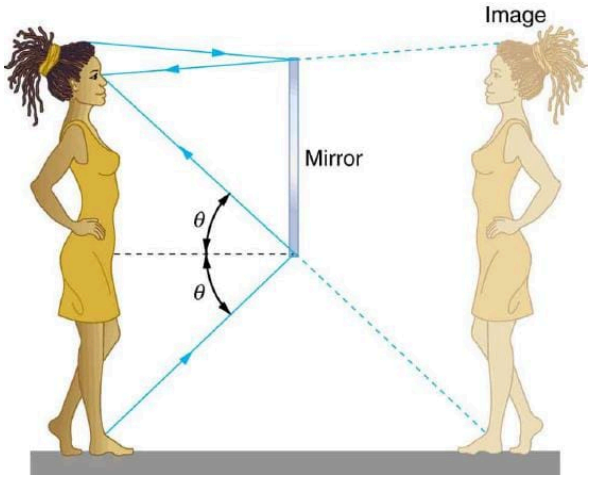


Figure 3.3.6.6: Our image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be in the direction the rays are coming from when they enter the eyes.

TAKE-HOME EXPERIMENT: LAW OF REFLECTION

Take a piece of paper and shine a flashlight at an angle at the paper, as shown in Figure 3.3.6.3. Now shine the flashlight at a mirror at an angle. Do your observations confirm the predictions in Figure 3.3.6.3 and Figure 3.3.6.4? Shine the flashlight on various surfaces and determine whether the reflected light is diffuse or not. You can choose a shiny metallic lid of a pot or your skin. Using the mirror and flashlight, can you confirm the law of reflection? You will need to draw lines on a piece of paper showing the incident and reflected rays. (This part works even better if you use a laser pencil.)

Section Summary

- The angle of reflection equals the angle of incidence.
- A mirror has a smooth surface and reflects light at specific angles.
- Light is diffused when it reflects from a rough surface.
- Mirror images can be photographed and videotaped by instruments.

Glossary

mirror
: smooth surface that reflects light at specific angles, forming an image of the person or object in front of it

law of reflection
: angle of reflection equals the angle of incidence

This page titled 3.3.6: Reflection is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.6: Reflection by Andrew Park is licensed CC BY 4.0.

3.3.7: Refraction

Learning Objectives

- Determine the index of refraction, given the speed of light in a medium.

It is easy to notice some odd things when looking into a fish tank. For example, you may see the same fish appearing to be in two different places. (See Figure 3.3.7.1.) This is because light coming from the fish to us changes direction when it leaves the tank, and in this case, it can travel two different paths to get to our eyes. The changing of a light ray's direction (loosely called bending) when it passes through variations in matter is called **refraction**. Refraction is responsible for a tremendous range of optical phenomena, from the action of lenses to voice transmission through optical fibers.

Definition: REFRACTION

The changing of a light ray's direction (loosely called bending) when it passes through variations in matter is called refraction.

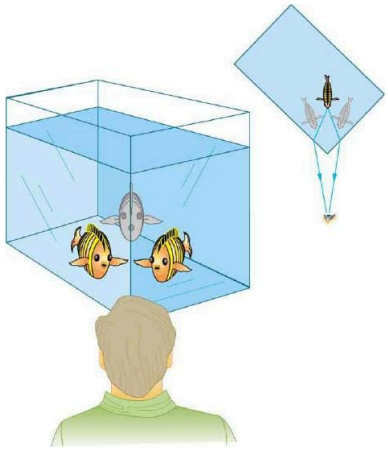


Figure 3.3.7.1: Looking at the fish tank as shown, we can see the same fish in two different locations, because light changes directions when it passes from water to air. In this case, the light can reach the observer by two different paths, and so the fish seems to be in two different places. This bending of light is called refraction and is responsible for many optical phenomena.

Why does light change direction when passing from one material (medium) to another? It is because light changes speed when going from one material to another. So before we study the law of refraction, it is useful to discuss the speed of light and how it varies in different media.

The Speed of Light

Early attempts to measure the speed of light, such as those made by Galileo, determined that light moved extremely fast, perhaps instantaneously. The first real evidence that light traveled at a finite speed came from the Danish astronomer Ole Roemer in the late 17th century. Roemer had noted that the average orbital period of one of Jupiter's moons, as measured from Earth, varied depending on whether Earth was moving toward or away from Jupiter. He correctly concluded that the apparent change in period was due to the change in distance between Earth and Jupiter and the time it took light to travel this distance. From his 1676 data, a value of the speed of light was calculated to be $2.26 \times 10^8\ \mathrm{m/s}$ (only 25% different from today's accepted value). In more recent times, physicists have measured the speed of light in numerous ways and with increasing accuracy. One particularly direct method, used in 1887 by the American physicist Albert Michelson (1852–1931), is illustrated in Figure 3.3.7.2. Light reflected from a rotating set of mirrors was reflected from a stationary mirror 35 km away and returned to the rotating mirrors. The time for the light to travel can be determined by how fast the mirrors must rotate for the light to be returned to the observer's eye.

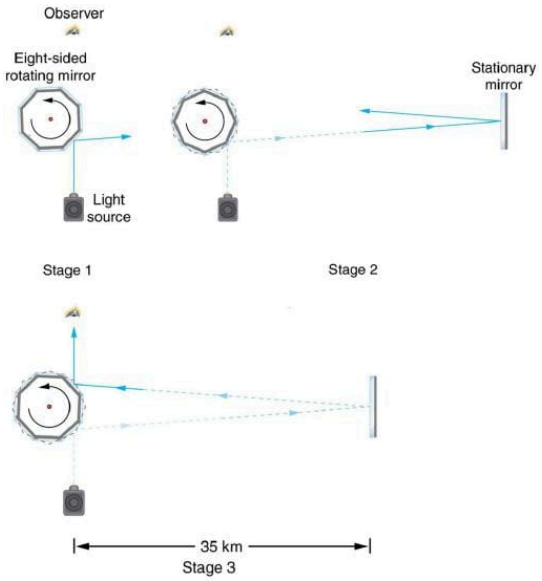


Figure 3.3.7.2: A schematic of early apparatus used by Michelson and others to determine the speed of light. As the mirrors rotate, the reflected ray is only briefly directed at the stationary mirror. The returning ray will be reflected into the observer's eye only if the next mirror has rotated into the correct position just as the ray returns. By measuring the correct rotation rate, the time for the round trip can be measured and the speed of light calculated. Michelson's calculated value of the speed of light was only 0.004% different from the value used today.

The speed of light is now known to great precision. In fact, the speed of light in a vacuum c is so important that it is accepted as one of the basic physical quantities and has the fixed value

$$c = 2.9972458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s},$$

where the approximate value of $3.00 \times 10^8 \text{ m/s}$ is used whenever three-digit accuracy is sufficient. The speed of light through matter is less than it is in a vacuum, because light interacts with atoms in a material. The speed of light depends strongly on the type of material, since its interaction with different atoms, crystal lattices, and other substructures varies. We define the **index of refraction** n of a material to be

$$n = \frac{c}{v},$$

where v is the observed speed of light in the material. Since the speed of light is always less than c in matter and equals c only in a vacuum, the index of refraction is always greater than or equal to one.

VALUE OF THE SPEED OF LIGHT

$$c = 2.9972458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$$

Definition: INDEX OF REFRACTION

$$n = \frac{c}{v}$$

That is, $n \geq 1$. Table 3.3.7.1 gives the indices of refraction for some representative substances. The values are listed for a particular wavelength of light, because they vary slightly with wavelength. (This can have important effects, such as colors produced by a prism.) Note that for gases, n is close to 1.00. This seems reasonable, since atoms in gases are widely separated and light travels at c in the vacuum between atoms. It is common to take $n = 1$ for gases unless great precision is needed. Although the speed of light v in a medium varies considerably from its value c in a vacuum, it is still a large speed.

Table 3.3.7.1: *Index of Refraction in Various Media*

Medium	*n*
Gases at 0°C, 1 atm	
Air	1.000293
Carbon dioxide	1.00045
Hydrogen	1.000139
Oxygen	1.000271
Liquids at 20°C	
Benzene	1.501
Carbon disulfide	1.628
Carbon tetrachloride	1.461
Ethanol	1.361
Glycerine	1.473
Water, fresh	1.333
Solids at 20°C	
Diamond	2.419
Fluorite	1.434
Glass, crown	1.52
Glass, flint	1.66
Ice at 20°C	1.309
Polystyrene	1.49
Plexiglas	1.51
Quartz, crystalline	1.544
Quartz, fused	1.458
Sodium chloride	1.544
Zircon	1.923

Example 3.3.7.1: Speed of Light in Matter

Calculate the speed of light in zircon, a material used in jewelry to imitate diamond.

Strategy

The speed of light in a material, v, can be calculated from the index of refraction n of the material using the equation $n = c/v$.

Solution

The equation for index of refraction states that $n = c/v$. Rearranging this to determine v gives

$$v = \frac{c}{n}.$$

The index of refraction for zircon is given as 1.923 in Table 3.3.7.1, and c is given in the equation for speed of light. Entering these values in the last expression gives

$$v = \frac{3.00 \times 10^8 \text{ m/s}}{1.923} = 1.56 \times 10^8 \text{ m/s}.$$

Discussion

This speed is slightly larger than half the speed of light in a vacuum and is still high compared with speeds we normally experience. The only substance listed in Table 3.3.7.1 that has a greater index of refraction than zircon is diamond. We shall see later that the large index of refraction for zircon makes it sparkle more than glass, but less than diamond.

Law of Refraction

Figure 3.3.7.3 shows how a ray of light changes direction when it passes from one medium to another. As before, the angles are measured relative to a perpendicular to the surface at the point where the light ray crosses it. (Some of the incident light will be reflected from the surface, but for now we will concentrate on the light that is transmitted.) The change in direction of the light ray depends on how the speed of light changes. The change in the speed of light is related to the indices of refraction of the media involved. In the situations shown in Figure 3.3.7.3, medium 2 has a greater index of refraction than medium 1. This means that the speed of light is less in medium 2 than in medium 1. Note that as shown in Figure 3.3.7.3(a), the direction of the ray moves closer to the perpendicular when it slows down. Conversely, as shown in Figure 3.3.7.3(b), the direction of the ray moves away from the perpendicular when it speeds up. The path is exactly reversible. In both cases, you can imagine what happens by thinking about pushing a lawn mower from a footpath onto grass, and vice versa. Going from the footpath to grass, the front wheels are slowed and pulled to the side as shown. This is the same change in direction as for light when it goes from a fast medium to a slow one. When going from the grass to the footpath, the front wheels can move faster and the mower changes direction as shown. This, too, is the same change in direction as for light going from slow to fast.

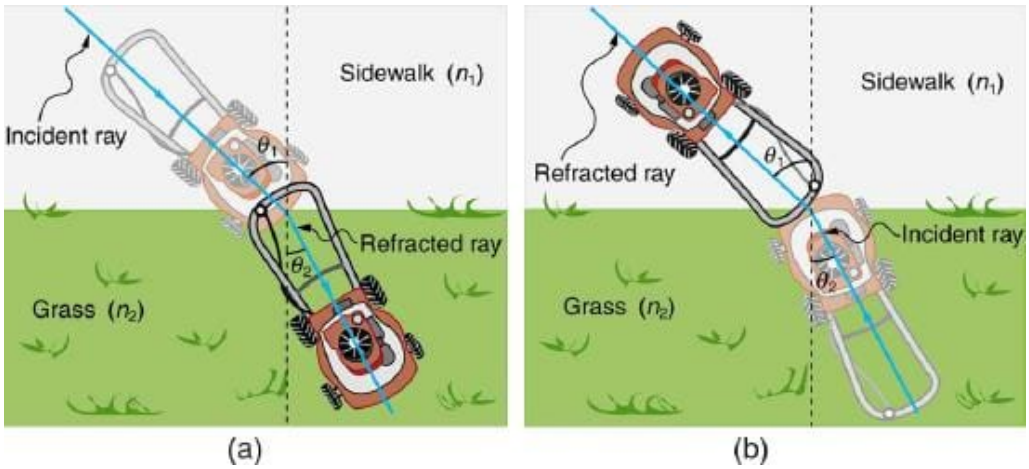


Figure 3.3.7.3: The change in direction of a light ray depends on how the speed of light changes when it crosses from one medium to another. The speed of light is greater in medium 1 than in medium 2 in the situations shown here. (a) A ray of light moves closer to the perpendicular when it slows down. This is analogous to what happens when a lawn mower goes from a footpath to grass. (b) A ray of light moves away from the perpendicular when it speeds up. This is analogous to what happens when a lawn mower goes from grass to footpath. The paths are exactly reversible.

The amount that a light ray changes its direction depends both on the incident angle and the amount that the speed changes. For a ray at a given incident angle, a large change in speed causes a large change in direction, and thus a large change in angle.

TAKE-HOME EXPERIMENT: A BROKEN PENCIL

A classic observation of refraction occurs when a pencil is placed in a glass half filled with water. Do this and observe the shape of the pencil when you look at the pencil sideways, that is, through air, glass, water. Explain your observations. Draw ray diagrams for the situation.

Section Summary

- The changing of a light ray's direction when it passes through variations in matter is called refraction.
- The speed of light in vacuum $c = 2.9972458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$
- Index of refraction $n = \frac{c}{v}$, where v is the speed of light in the material, c is the speed of light in vacuum, and n is the index of refraction.

Glossary

refraction
: changing of a light ray's direction when it passes through variations in matter

index of refraction
: for a material, the ratio of the speed of light in vacuum to that in the material

This page titled 3.3.7: Refraction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.7: Refraction by Andrew Park is licensed CC BY 4.0.

3.3.8: Dispersion- The Rainbow and Prisms

Learning Objectives

- Explain the phenomenon of dispersion and discuss its advantages and disadvantages.

Everyone enjoys the spectacle of a rainbow glimmering against a dark stormy sky. How does sunlight falling on clear drops of rain get broken into the rainbow of colors we see? The same process causes white light to be broken into colors by a clear glass prism or a diamond. (See Figure 3.3.8.1.)

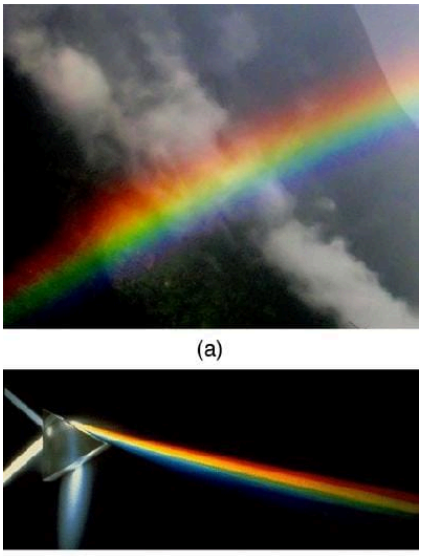


Figure 3.3.8.1: The colors of the rainbow (a) and those produced by a prism (b) are identical. (credit: Alfredo55, Wikimedia Commons; NASA)

We see about six colors in a rainbow—red, orange, yellow, green, blue, and violet; sometimes indigo is listed, too. Those colors are associated with different wavelengths of light, as shown in Figure 3.3.8.2. When our eye receives pure-wavelength light, we tend to see only one of the six colors, depending on wavelength. The thousands of other hues we can sense in other situations are our eye's response to various mixtures of wavelengths. White light, in particular, is a fairly uniform mixture of all visible wavelengths. Sunlight, considered to be white, actually appears to be a bit yellow because of its mixture of wavelengths, but it does contain all visible wavelengths. The sequence of colors in rainbows is the same sequence as the colors plotted versus wavelength in Figure 3.3.8.2. What this implies is that white light is spread out according to wavelength in a rainbow. **Dispersion** is defined as the spreading of white light into its full spectrum of wavelengths. More technically, dispersion occurs whenever there is a process that changes the direction of light in a manner that depends on wavelength. Dispersion, as a general phenomenon, can occur for any type of wave and always involves wavelength-dependent processes.

Definition: DISPERSION

Dispersion is defined to be the spreading of white light into its full spectrum of wavelengths.

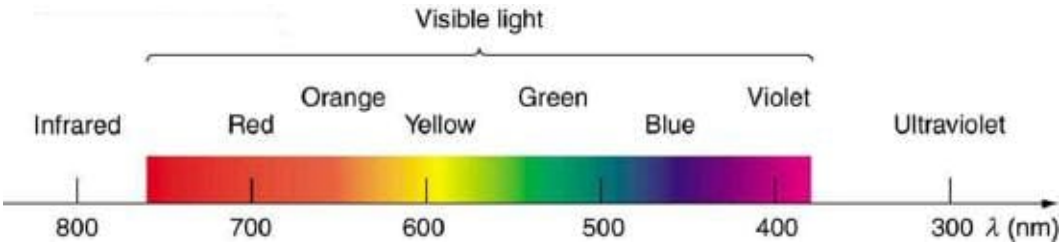


Figure 3.3.8.2: Even though rainbows are associated with seven colors, the rainbow is a continuous distribution of colors according to wavelengths.

Refraction is responsible for dispersion in rainbows and many other situations. The angle of refraction depends on the index of refraction. We know that the index of refraction n depends on the medium. But for a given medium, n also depends on wavelength. (See Table 3.3.8.1. Note that, for a given medium, n increases as wavelength decreases and is greatest for violet light. Thus violet light is bent more than red light, as shown for a prism in Figure 3.3.8.3(b), and the light is dispersed into the same sequence of wavelengths as seen in Figure 3.3.8.1 and Figure 3.3.8.2.

MAKING CONNECTIONS: DISPERSION

Any type of wave can exhibit dispersion. Sound waves, all types of electromagnetic waves, and water waves can be dispersed according to wavelength. Dispersion occurs whenever the speed of propagation depends on wavelength, thus separating and spreading out various wavelengths. Dispersion may require special circumstances and can result in spectacular displays such as in the production of a rainbow. This is also true for sound, since all frequencies ordinarily travel at the same speed. If you listen to sound through a long tube, such as a vacuum cleaner hose, you can easily hear it is dispersed by interaction with the tube. Dispersion, in fact, can reveal a great deal about what the wave has encountered that disperses its wavelengths. The dispersion of electromagnetic radiation from outer space, for example, has revealed much about what exists between the stars—the so-called empty space.

Table 3.3.8.1*: Index of Refraction n in Selected Media at Various Wavelengths*

Medium	Red (660 nm)	Orange (610 nm)	Yellow (580 nm)	Green (550 nm)	Blue (470 nm)	Violet (410 nm)
Water	1.331	1.332	1.333	1.335	1.338	1.342
Diamond	2.410	2.415	2.417	2.426	2.444	2.458
Glass, crown	1.512	1.514	1.518	1.519	1.524	1.530
Glass, flint	1.662	1.665	1.667	1.674	1.684	1.698
Polystyrene	1.488	1.490	1.492	1.493	1.499	1.506
Quartz, fused	1.455	1.456	1.458	1.459	1.462	1.468

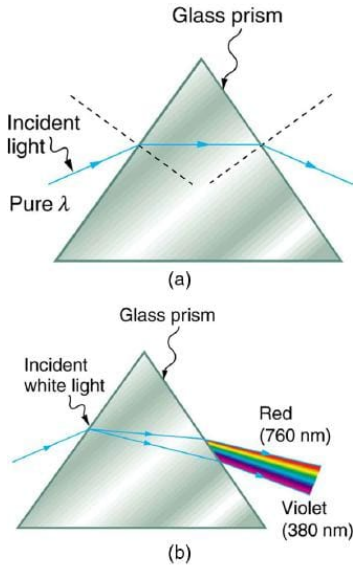


Figure 3.3.8.3: (a) A pure wavelength of light falls onto a prism and is refracted at both surfaces. (b) White light is dispersed by the prism (shown exaggerated). Since the index of refraction varies with wavelength, the angles of refraction vary with wavelength. A sequence of red to violet is produced, because the index of refraction increases steadily with decreasing wavelength.

Rainbows are produced by a combination of refraction and reflection. You may have noticed that you see a rainbow only when you look away from the sun. Light enters a drop of water and is reflected from the back of the drop, as shown in Figure 3.3.8.4. The light is refracted both as it enters and as it leaves the drop. Since the index of refraction of water varies with wavelength, the light is dispersed, and a rainbow is observed, as shown in Figure 3.3.8.5(a). (There is no dispersion caused by reflection at the back surface, since the law of reflection does not depend on wavelength.) The actual rainbow of colors seen by an observer depends on the myriad of rays being refracted and reflected toward the observer's eyes from numerous drops of water. The effect is most spectacular when the background is dark, as in stormy weather, but can also be observed in waterfalls and lawn sprinklers. The arc of a rainbow comes from the need to be looking at a specific angle relative to the direction of the sun, as illustrated in Figure 3.3.8.5(b). (If there are two reflections of light within the water drop, another "secondary" rainbow is produced. This rare event produces an arc that lies above the primary rainbow arc—see Figure 3.3.8.5(c).)

Definition: RAINBOWS

Rainbows are produced by a combination of refraction and reflection.

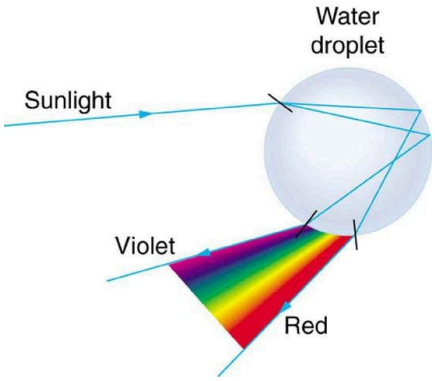


Figure 3.3.8.4: Part of the light falling on this water drop enters and is reflected from the back of the drop. This light is refracted and dispersed both as it enters and as it leaves the drop.

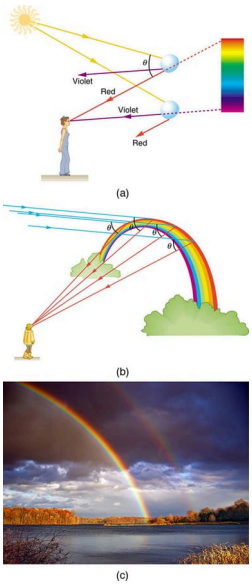


Figure 3.3.8.5: (a) Different colors emerge in different directions, and so you must look at different locations to see the various colors of a rainbow. (b) The arc of a rainbow results from the fact that a line between the observer and any point on the arc must make the correct angle with the parallel rays of sunlight to receive the refracted rays. (c) Double rainbow. (credit: Nicholas, Wikimedia Commons)

Dispersion may produce beautiful rainbows, but it can cause problems in optical systems. White light used to transmit messages in a fiber is dispersed, spreading out in time and eventually overlapping with other messages. Since a laser produces a nearly pure wavelength, its light experiences little dispersion, an advantage over white light for transmission of information. In contrast, dispersion of electromagnetic waves coming to us from outer space can be used to determine the amount of matter they pass through. As with many phenomena, dispersion can be useful or a nuisance, depending on the situation and our human goals.

Section Summary

- The spreading of white light into its full spectrum of wavelengths is called dispersion.
- Rainbows are produced by a combination of refraction and reflection and involve the dispersion of sunlight into a continuous distribution of colors.
- Dispersion produces beautiful rainbows but also causes problems in certain optical systems.

Glossary

dispersion
: spreading of white light into its full spectrum of wavelengths

rainbow
: dispersion of sunlight into a continuous distribution of colors according to wavelength, produced by the refraction and reflection of sunlight by water droplets in the sky

This page titled 3.3.8: Dispersion- The Rainbow and Prisms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.8: Dispersion- The Rainbow and Prisms by Andrew Park is licensed CC BY 4.0.

3.3.9: Polarization

Learning Objectives

- Discuss the meaning of polarization.
- Discuss the property of optical activity of certain materials.

Polaroid sunglasses are familiar to most of us. They have a special ability to cut the glare of light reflected from water or glass (see Figure 3.3.9.1). Polaroids have this ability because of a wave characteristic of light called polarization. What is polarization? How is it produced? What are some of its uses? The answers to these questions are related to the wave character of light.



Figure 3.3.9.1: These two photographs of a river show the effect of a polarizing filter in reducing glare in light reflected from the surface of water. Part (b) of this figure was taken with a polarizing filter and part (a) was not. As a result, the reflection of clouds and sky observed in part (a) is not observed in part (b). Polarizing sunglasses are particularly useful on snow and water. (credit: Amithshs, Wikimedia Commons)

Light is one type of electromagnetic (EM) wave. As noted earlier, EM waves are *transverse waves* consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation (see Figure 3.3.9.2). There are specific directions for the oscillations of the electric and magnetic fields. **Polarization** is the attribute that a wave's oscillations have a definite direction relative to the direction of propagation of the wave. (This is not the same type of polarization as that discussed for the separation of charges.) Waves having such a direction are said to be **polarized**. For an EM wave, we define the **direction of polarization** to be the direction parallel to the electric field. Thus we can think of the electric field arrows as showing the direction of polarization, as in Figure 3.3.9.2.

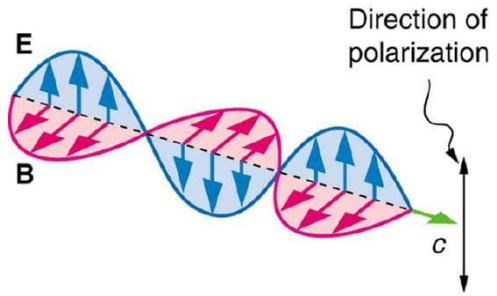


Figure 3.3.9.2: An EM wave, such as light, is a transverse wave. The electric and magnetic fields are perpendicular to the direction of propagation.

To examine this further, consider the transverse waves in the ropes shown in Figure 3.3.9.3. The oscillations in one rope are in a vertical plane and are said to be **vertically polarized**. Those in the other rope are in a horizontal plane and are **horizontally polarized**. If a vertical slit is placed on the first rope, the waves pass through. However, a vertical slit blocks the horizontally polarized waves. For EM waves, the direction of the electric field is analogous to the disturbances on the ropes.

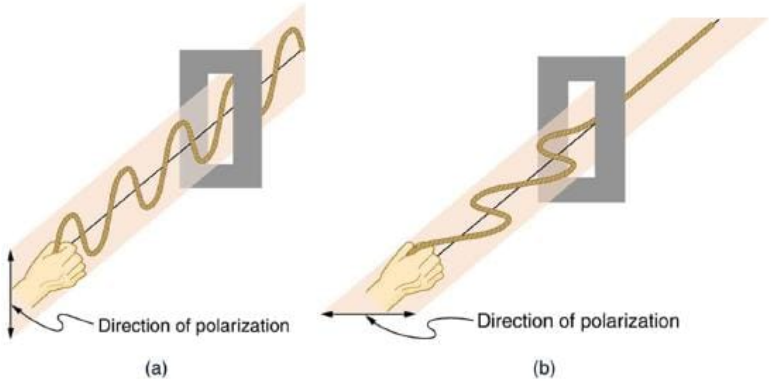


Figure 3.3.9.3: The transverse oscillations in one rope are in a vertical plane, and those in the other rope are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally polarized waves.

The Sun and many other light sources produce waves that are randomly polarized (see Figure 3.3.9.4). Such light is said to be **unpolarized** because it is composed of many waves with all possible directions of polarization. Polaroid materials, invented by the founder of Polaroid Corporation, Edwin Land, act as a *polarizing* slit for light, allowing only polarization in one direction to pass through. Polarizing filters are composed of long molecules aligned in one direction. Thinking of the molecules as many slits, analogous to those for the oscillating ropes, we can understand why only light with a specific polarization can get through. The **axis of a polarizing filter** is the direction along which the filter passes the electric field of an EM wave (see Figure 3.3.9.5).

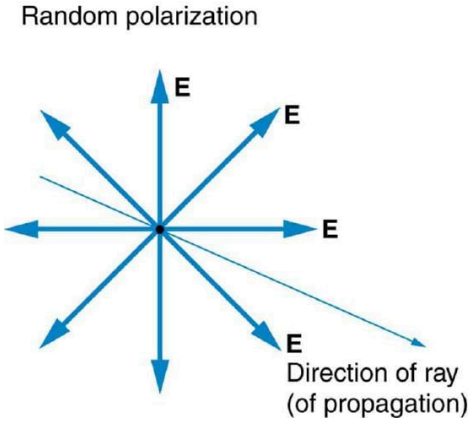


Figure 3.3.9.4: The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. Since the light is unpolarized, the arrows point in all directions.

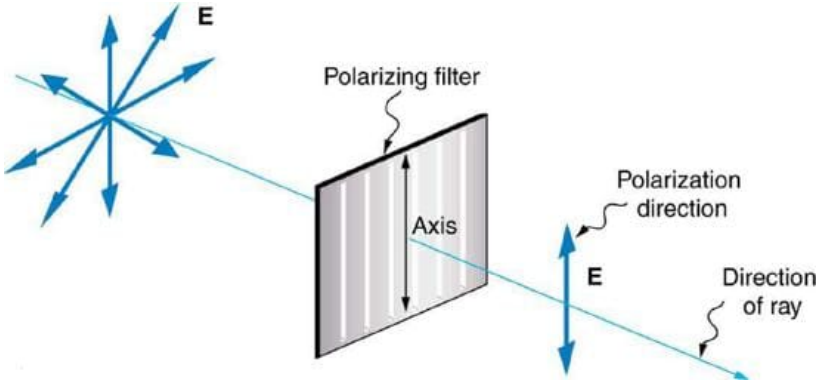


Figure 3.3.9.5: A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM wave is defined to be the direction of its electric field.

Figure 3.3.9.6 shows the effect of two polarizing filters on originally unpolarized light. The first filter polarizes the light along its axis. When the axes of the first and second filters are aligned (parallel), then all of the polarized light passed by the first filter is also passed by the second. If the second polarizing filter is rotated, only the component of the light parallel to the second filter's axis is passed. When the axes are perpendicular, no light is passed by the second.

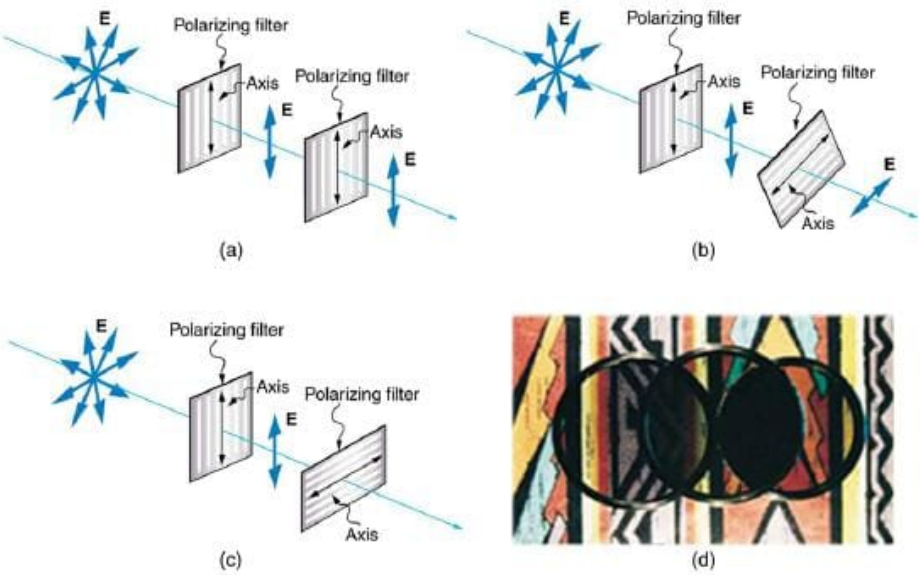


Figure 3.3.9.6: The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second is rotated, only part of the light is passed. (c) When the second is perpendicular to the first, no light is passed. (d) In this photograph, a polarizing filter is placed above two others. Its axis is perpendicular to the filter on the right (dark area) and parallel to the filter on the left (lighter area). (credit: P.P. Urone)

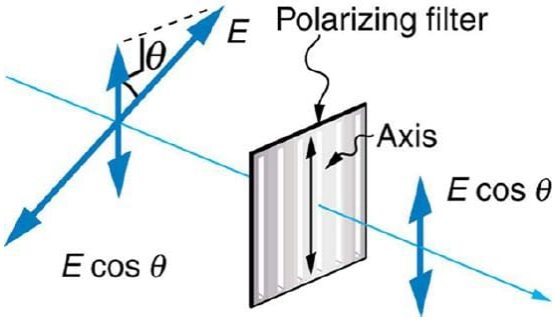


Figure 3.3.9.7: A polarizing filter transmits only the component of the wave parallel to its axis, $E\cos\theta$, reducing the intensity of any light not polarized parallel to its axis.

Polarization by Reflection

By now you can probably guess that Polaroid sunglasses cut the glare in reflected light because that light is polarized. You can check this for yourself by holding Polaroid sunglasses in front of you and rotating them while looking at light reflected from water or glass. As you rotate the sunglasses, you will notice the light gets bright and dim, but not completely black. This implies the reflected light is partially polarized and cannot be completely blocked by a polarizing filter.

Figure 3.3.9.8 illustrates what happens when unpolarized light is reflected from a surface. Vertically polarized light is preferentially refracted at the surface, so that *the reflected light is left more horizontally polarized*. The reasons for this phenomenon are beyond the scope of this text, but a convenient mnemonic for remembering this is to imagine the polarization direction to be like an arrow. Vertical polarization would be like an arrow perpendicular to the surface and would be more likely to stick and not be reflected. Horizontal polarization is like an arrow bouncing on its side and would be more likely to be reflected. Sunglasses with vertical axes would then block more reflected light than unpolarized light from other sources.

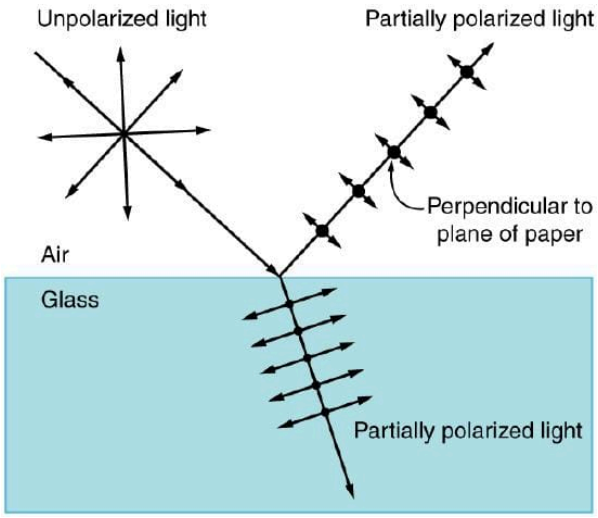


Figure 3.3.9.8: Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking on their sides bouncing off, whereas arrows striking on their tips go into the surface.

THINGS GREAT AND SMALL: ATOMIC EXPLANATION OF POLARIZING FILTERS

Polarizing filters have a polarization axis that acts as a slit. This slit passes electromagnetic waves (often visible light) that have an electric field parallel to the axis. This is accomplished with long molecules aligned perpendicular to the axis as shown in Figure 3.3.9.9.

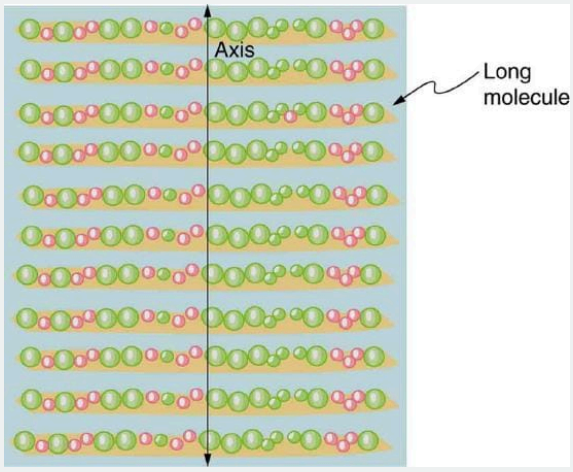


Figure 3.3.9.9: Long molecules are aligned perpendicular to the axis of a polarizing filter. The component of the electric field in an EM wave perpendicular to these molecules passes through the filter, while the component parallel to the molecules is absorbed.

Figure 3.3.9.10 illustrates how the component of the electric field parallel to the long molecules is absorbed. An electromagnetic wave is composed of oscillating electric and magnetic fields. The electric field is strong compared with the magnetic field and is more effective in exerting force on charges in the molecules. The most affected charged particles are the electrons in the molecules, since electron masses are small. If the electron is forced to oscillate, it can absorb energy from the EM wave. This reduces the fields in the wave and, hence, reduces its intensity. In long molecules, electrons can more easily oscillate parallel to the molecule than in the perpendicular direction. The electrons are bound to the molecule and are more restricted in their movement perpendicular to the molecule. Thus, the electrons can absorb EM waves that have a component of their electric field parallel to the molecule. The electrons are much less responsive to electric fields perpendicular to the molecule and will allow those fields to pass. Thus the axis of the polarizing filter is perpendicular to the length of the molecule.

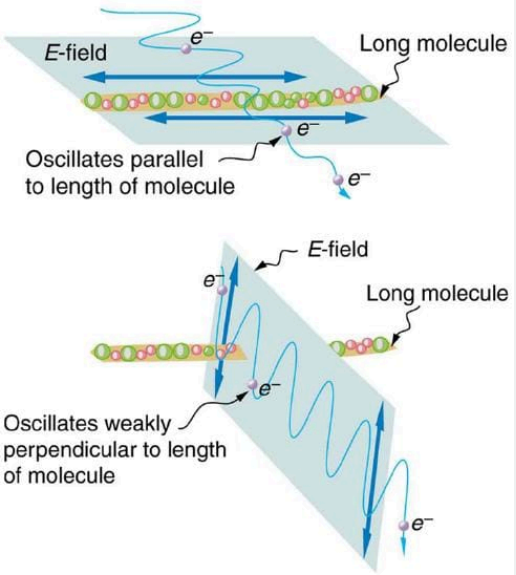


Figure 3.3.9.10: Artist's conception of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and reduces the intensity of the component of the EM wave that is parallel to the molecule.

Polarization by Scattering

If you hold your Polaroid sunglasses in front of you and rotate them while looking at blue sky, you will see the sky get bright and dim. This is a clear indication that light scattered by air is partially polarized. Figure 3.3.9.11 helps illustrate how this happens. Since light is a transverse EM wave, it vibrates the electrons of air molecules perpendicular to the direction it is traveling. The electrons then radiate like small antennae. Since they are oscillating perpendicular to the direction of the light ray, they produce EM radiation that is polarized perpendicular to the direction of the ray. When viewing the light along a line perpendicular to the original ray, as in Figure 3.3.9.11, there can be no polarization in the scattered light parallel to the original ray, because that would require the original ray to be a longitudinal wave. Along other directions, a component of the other polarization can be projected along the line of sight, and the scattered light will only be partially polarized. Furthermore, multiple scattering can bring light to your eyes from other directions and can contain different polarizations.

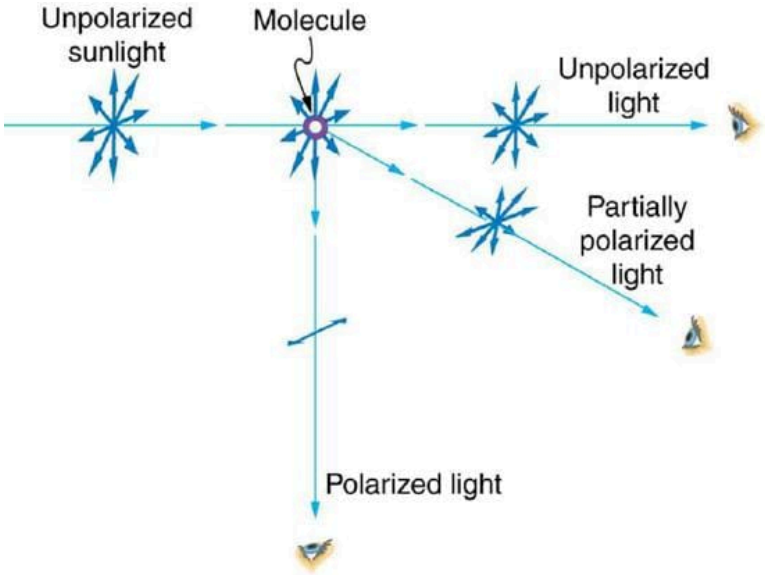


Figure 3.3.9.11: Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

Photographs of the sky can be darkened by polarizing filters, a trick used by many photographers to make clouds brighter by

contrast. Scattering from other particles, such as smoke or dust, can also polarize light. Detecting polarization in scattered EM waves can be a useful analytical tool in determining the scattering source.

There is a range of optical effects used in sunglasses. Besides being Polaroid, other sunglasses have colored pigments embedded in them, while others use non-reflective or even reflective coatings. A recent development is photochromic lenses, which darken in the sunlight and become clear indoors. Photochromic lenses are embedded with organic microcrystalline molecules that change their properties when exposed to UV in sunlight, but become clear in artificial lighting with no UV.

TAKE-HOME EXPERIMENT: POLARIZATION

Find Polaroid sunglasses and rotate one while holding the other still and look at different surfaces and objects. Explain your observations. What is the difference in angle from when you see a maximum intensity to when you see a minimum intensity? Find a reflective glass surface and do the same. At what angle does the glass need to be oriented to give minimum glare?

Liquid Crystals and Other Polarization Effects in Materials

While you are undoubtedly aware of liquid crystal displays (LCDs) found in watches, calculators, computer screens, cellphones, flat screen televisions, and other myriad places, you may not be aware that they are based on polarization. Liquid crystals are so named because their molecules can be aligned even though they are in a liquid. Liquid crystals have the property that they can rotate the polarization of light passing through them by 90°. Furthermore, this property can be turned off by the application of a voltage, as illustrated in Figure $3.3.9.12$. It is possible to manipulate this characteristic quickly and in small well-defined regions to create the contrast patterns we see in so many LCD devices.

In flat screen LCD televisions, there is a large light at the back of the TV. The light travels to the front screen through millions of tiny units called pixels (picture elements). One of these is shown in Figure $3.3.9.1$ (a) and (b). Each unit has three cells, with red, blue, or green filters, each controlled independently. When the voltage across a liquid crystal is switched off, the liquid crystal passes the light through the particular filter. One can vary the picture contrast by varying the strength of the voltage applied to the liquid crystal.

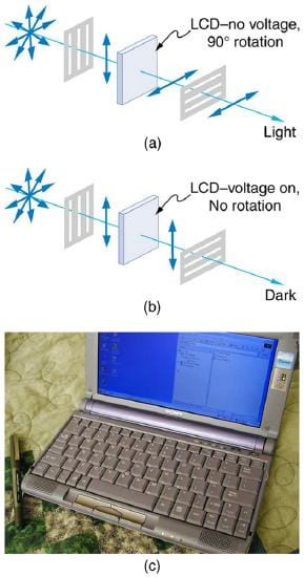


Figure $3.3.9.12$: (a) Polarized light is rotated 90° by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the original polarization direction. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making the region dark in comparison with its surroundings. (c) LCDs can be made color specific, small, and fast enough to use in laptop computers and TVs. (credit: Jon Sullivan)

Many crystals and solutions rotate the plane of polarization of light passing through them. Such substances are said to be **optically active**. Examples include sugar water, insulin, and collagen (see Figure $3.3.9.13$). In addition to depending on the type of substance, the amount and direction of rotation depends on a number of factors. Among these is the concentration of the substance, the distance the light travels through it, and the wavelength of light. Optical activity is due to the asymmetric shape

of molecules in the substance, such as being helical. Measurements of the rotation of polarized light passing through substances can thus be used to measure concentrations, a standard technique for sugars. It can also give information on the shapes of molecules, such as proteins, and factors that affect their shapes, such as temperature and pH.

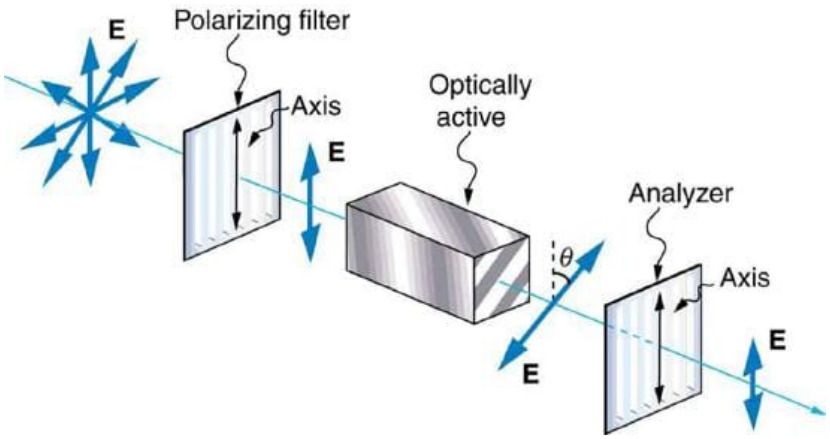


Figure 3.3.9.13: Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

Glass and plastic become optically active when stressed; the greater the stress, the greater the effect. Optical stress analysis on complicated shapes can be performed by making plastic models of them and observing them through crossed filters, as seen in Figure 3.3.9.14. It is apparent that the effect depends on wavelength as well as stress. The wavelength dependence is sometimes also used for artistic purposes.

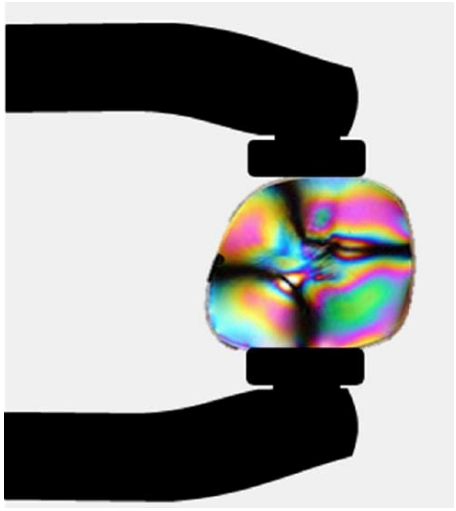


Figure 3.3.9.14: Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: Infopro, Wikimedia Commons)

Another interesting phenomenon associated with polarized light is the ability of some crystals to split an unpolarized beam of light into two. Such crystals are said to be **birefringent** (see Figure 3.3.9.15). Each of the separated rays has a specific polarization. One behaves normally and is called the ordinary ray, whereas the other does not obey Snell's law and is called the extraordinary ray. Birefringent crystals can be used to produce polarized beams from unpolarized light. Some birefringent materials preferentially absorb one of the polarizations. These materials are called dichroic and can produce polarization by this preferential absorption. This is fundamentally how polarizing filters and other polarizers work. The interested reader is invited to further pursue the numerous properties of materials related to polarization.

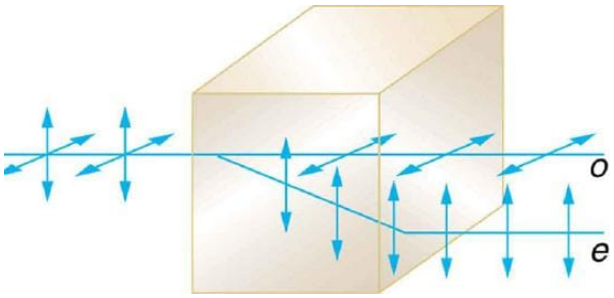


Figure 3.3.9.15: Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two. The ordinary ray behaves as expected, but the extraordinary ray does not obey Snell's law.

Section Summary

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave.
- EM waves are transverse waves that may be polarized.
- The direction of polarization is defined to be the direction parallel to the electric field of the EM wave.
- Unpolarized light is composed of many rays having random polarization directions.
- Light can be polarized by passing it through a polarizing filter or other polarizing material. The intensity of polarized light after passing through a polarizing filter depends on the direction of polarization and the axis of the filter.
- Polarization is also produced by reflection.
- Polarization can also be produced by scattering.
- There are a number of types of optically active substances that rotate the direction of polarization of light passing through them.

Glossary

axis of a polarizing filter
: the direction along which the filter passes the electric field of an EM wave

birefringent
: crystals that split an unpolarized beam of light into two beams

direction of polarization
: the direction parallel to the electric field for EM waves

horizontally polarized
: the oscillations are in a horizontal plane

optically active
: substances that rotate the plane of polarization of light passing through them

polarization
: the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave

polarized
: waves having the electric and magnetic field oscillations in a definite direction

unpolarized
: waves that are randomly polarized

vertically polarized
: the oscillations are in a vertical plane

This page titled 3.3.9: Polarization is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.11: Polarization by Andrew Park is licensed CC BY 4.0.

3.3.10: Image Formation by Lenses

Learning Objectives

- List the rules for ray tracking for thin lenses.
- Illustrate the formation of images using the technique of ray tracking.
- Determine power of a lens given the focal length.

Lenses are found in a huge array of optical instruments, ranging from a simple magnifying glass to the eye to a camera's zoom lens. In this section, we will use the law of refraction to explore the properties of lenses and how they form images.

The word *lens* derives from the Latin word for a lentil bean, the shape of which is similar to the convex lens in Figure 3.3.10.1. The convex lens shown has been shaped so that all light rays that enter it parallel to its axis cross one another at a single point on the opposite side of the lens. (The axis is defined to be a line normal to the lens at its center, as shown in Figure 3.3.10.1.) Such a lens is called a **converging (or convex) lens** for the converging effect it has on light rays. An expanded view of the path of one ray through the lens is shown, to illustrate how the ray changes direction both as it enters and as it leaves the lens. Since the index of refraction of the lens is greater than that of air, the ray moves towards the perpendicular as it enters and away from the perpendicular as it leaves. (This is in accordance with the law of refraction.) Due to the lens's shape, light is thus bent toward the axis at both surfaces. The point at which the rays cross is defined to be the **focal point** F of the lens. The distance from the center of the lens to its focal point is defined to be the **focal length** f of the lens. Figure 3.3.10.2 shows how a converging lens, such as that in a magnifying glass, can converge the nearly parallel light rays from the sun to a small spot.

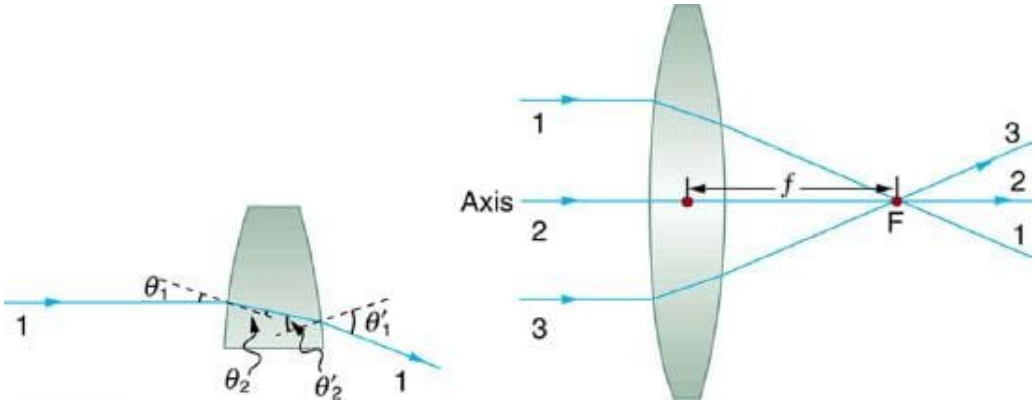


Figure 3.3.10.1: Rays of light entering a converging lens parallel to its axis converge at its focal point F. (Ray 2 lies on the axis of the lens.) The distance from the center of the lens to the focal point is the lens's focal length f. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Definition: CONVERGING OR CONVEX LENS

The lens in which light rays that enter it parallel to its axis cross one another at a single point on the opposite side with a converging effect is called converging lens.

Definition: FOCAL POINT F

The point at which the light rays cross is called the focal point F of the lens.

Definition: FOCAL LENGTH f

The distance from the center of the lens to its focal point is called focal length f.



Figure 3.3.10.2: Sunlight focused by a converging magnifying glass can burn paper. Light rays from the sun are nearly parallel and cross at the focal point of the lens. The more powerful the lens, the closer to the lens the rays will cross.

The greater effect a lens has on light rays, the more powerful it is said to be. For example, a powerful converging lens will focus parallel light rays closer to itself and will have a smaller focal length than a weak lens. The light will also focus into a smaller and more intense spot for a more powerful lens. The **power** P of a lens is defined to be the inverse of its focal length. In equation form, this is

$$P = \frac{1}{f}.$$

Definition: POWER P

The **power** P of a lens is defined to be the inverse of its focal length. In equation form, this is

$$P = \frac{1}{f}.$$

where f is the focal length of the lens, which must be given in meters (and not cm or mm). The power of a lens P has the unit diopters (D), provided that the focal length is given in meters. That is, $1\ \mathrm{D} = 1/\mathrm{m}$, or $1\ \mathrm{m}^{-1}$. (Note that this power (optical power, actually) is not the same as power in watts. It is a concept related to the effect of optical devices on light.) Optometrists prescribe common spectacles and contact lenses in units of diopters.

Example 3.3.10.1: What is the Power of a Common Magnifying Glass?

Suppose you take a magnifying glass out on a sunny day and you find that it concentrates sunlight to a small spot 8.00 cm away from the lens. What are the focal length and power of the lens?

Strategy

The situation here is the same as those shown in Figure 3.3.10.1 and Figure 3.3.10.2. The Sun is so far away that the Sun's rays are nearly parallel when they reach Earth. The magnifying glass is a convex (or converging) lens, focusing the nearly parallel rays of sunlight. Thus the focal length of the lens is the distance from the lens to the spot, and its power is the inverse of this distance (in m).

Solution

The focal length of the lens is the distance from the center of the lens to the spot, given to be 8.00 cm. Thus,

$$f = 8.00\ \mathrm{cm}.$$

To find the power of the lens, we must first convert the focal length to meters; then, we substitute this value into the equation for power. This gives

$$P = \frac{1}{f} = \frac{1}{0.0800 \text{ m}} = 12.5 \text{ D}.$$

Discussion

This is a relatively powerful lens. The power of a lens in diopters should not be confused with the familiar concept of power in watts. It is an unfortunate fact that the word “power” is used for two completely different concepts. If you examine a prescription for eyeglasses, you will note lens powers given in diopters. If you examine the label on a motor, you will note energy consumption rate given as a power in watts.

Figure 3.3.10.3 shows a concave lens and the effect it has on rays of light that enter it parallel to its axis (the path taken by ray 2 in the figure is the axis of the lens). The concave lens is a **diverging lens**, because it causes the light rays to bend away (diverge) from its axis. In this case, the lens has been shaped so that all light rays entering it parallel to its axis appear to originate from the same point, F, defined to be the focal point of a diverging lens. The distance from the center of the lens to the focal point is again called the focal length f of the lens. Note that the focal length and power of a diverging lens are defined to be negative. For example, if the distance to F in Figure 3.3.10.3 is 5.00 cm, then the focal length is $f = -5.00 \text{ cm}$ and the power of the lens is $P = -20 \text{ D}$. An expanded view of the path of one ray through the lens is shown in the figure to illustrate how the shape of the lens, together with the law of refraction, causes the ray to follow its particular path and be diverged.

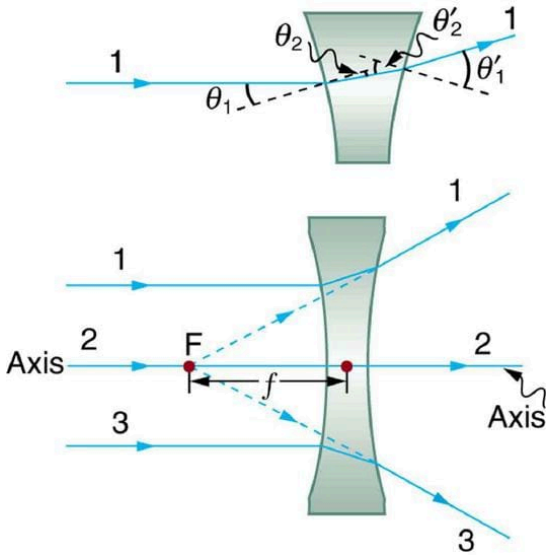


Figure 3.3.10.3: Rays of light entering a diverging lens parallel to its axis are diverged, and all appear to originate at its focal point F. The dashed lines are not rays—they indicate the directions from which the rays appear to come. The focal length f of a diverging lens is negative. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Definition: DIVERGING LENS

A lens that causes the light rays to bend away from its axis is called a diverging lens.

As noted in the initial discussion of the law of refraction, the paths of light rays are exactly reversible. This means that the direction of the arrows could be reversed for all of the rays in Figure 3.3.10.1 and Figure 3.3.10.3. For example, if a point light source is placed at the focal point of a convex lens, as shown in Figure 3.3.10.4, parallel light rays emerge from the other side.

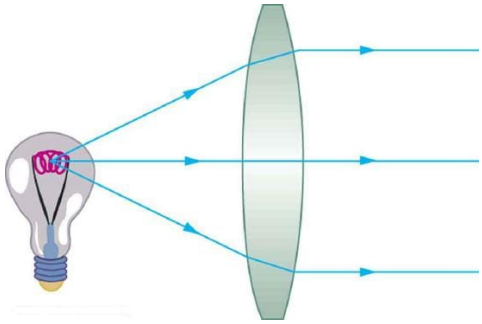


Figure 3.3.10.4: A small light source, like a light bulb filament, placed at the focal point of a convex lens, results in parallel rays of light emerging from the other side. The paths are exactly the reverse of those shown in Figure 3.3.10.1. This technique is used in lighthouses and sometimes in traffic lights to produce a directional beam of light from a source that emits light in all directions.

Ray Tracing and Thin Lenses

Ray tracing is the technique of determining or following (tracing) the paths that light rays take. For rays passing through matter, the law of refraction is used to trace the paths. Here we use ray tracing to help us understand the action of lenses in situations ranging from forming images on film to magnifying small print to correcting nearsightedness. While ray tracing for complicated lenses, such as those found in sophisticated cameras, may require computer techniques, there is a set of simple rules for tracing rays through thin lenses. A **thin lens** is defined to be one whose thickness allows rays to refract, as illustrated in Figure 3.3.10.1 , but does not allow properties such as dispersion and aberrations. An ideal thin lens has two refracting surfaces but the lens is thin enough to assume that light rays bend only once. A thin symmetrical lens has two focal points, one on either side and both at the same distance from the lens. (See Figure 3.3.10.5.) Another important characteristic of a thin lens is that light rays through its center are deflected by a negligible amount, as seen in Figure 3.3.10.6.

Definition: THIN LENS

A thin lens is defined to be one whose thickness allows rays to refract but does not allow properties such as dispersion and aberrations.

TAKE-HOME EXPERIMENT: A VISIT TO THE OPTICIAN

Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they act like thin lenses.

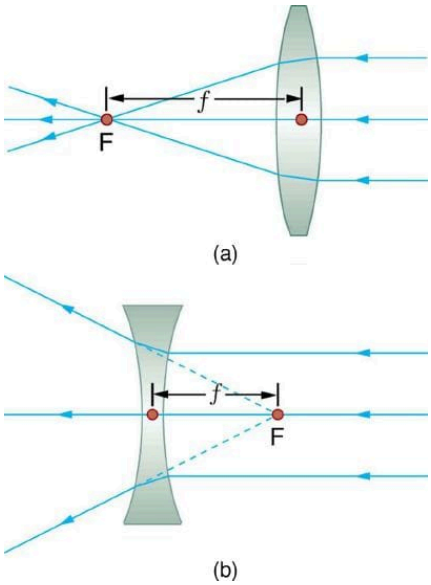


Figure 3.3.10.5: Thin lenses have the same focal length on either side. (a) Parallel light rays entering a converging lens from the right cross at its focal point on the left. (b) Parallel light rays entering a diverging lens from the right seem to come from the focal point on the right.

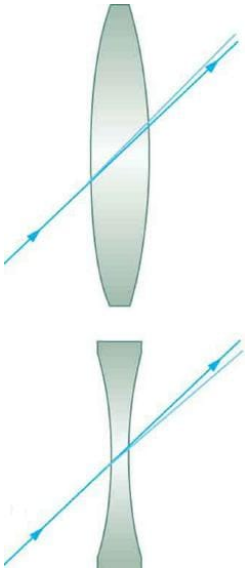


Figure 3.3.10.6: The light ray through the center of a thin lens is deflected by a negligible amount and is assumed to emerge parallel to its original path (shown as a shaded line).

Using paper, pencil, and a straight edge, ray tracing can accurately describe the operation of a lens. The rules for ray tracing for thin lenses are based on the illustrations already discussed:

- **Principal Ray 1**: a ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side (see rays 1 and 3 in Figure 3.3.10.1); a ray entering a diverging lens parallel to its axis seems to come from the focal point F (see rays 1 and 3 in Figure 3.3.10.3)
- **Principal Ray 2**: a ray passing through the center of either a converging or a diverging lens does not change direction (see Figure 3.3.10.6, and see ray 2 in Figure 3.3.10.1 and Figure 3.3.10.3)
- **Principal Ray 3**: a ray entering a converging lens through its focal point exits parallel to its axis (the reverse of rays 1 and 3 in Figure 3.3.10.1); a ray that enters a diverging lens by heading toward the focal point on the opposite side exits parallel to the axis (the reverse of rays 1 and 3 in Figure 3.3.10.3). *The third principal ray is optional and may be used to verify the accuracy of image location.*

Image Formation by Thin Lenses

In some circumstances, a lens forms an obvious image, such as when a movie projector casts an image onto a screen. In other cases, the image is less obvious. Where, for example, is the image formed by eyeglasses? We will use ray tracing for thin lenses to illustrate how they form images, and we will develop equations to describe the image formation quantitatively.

Consider an object some distance away from a converging lens, as shown in Figure 3.3.10.7. To find the location and size of the image formed, we trace the paths of principal rays originating from one point on the object, in this case the top of the person's head. The figure shows the three principal rays from the top of the object as described above. (Note that there are many light rays leaving this point going in many directions, but we concentrate on the three principal rays which can be traced by simple rules.) The first ray is one that enters the lens parallel to its axis and passes through the focal point on the other side (Principal Ray 1). The second ray passes through the center of the lens without changing direction (Principal Ray 2). The third ray passes through the nearer focal point on its way into the lens and leaves the lens parallel to its axis (Principal Ray 3). If correctly drawn, all three rays cross at the same point on the other side of the lens. The image of the top of the person's head is located at this point. *All* rays, including those that are not principal rays, that come from the same point on the top of the person's head are refracted in such a way as to cross at the point shown; the principal rays are what we use to *find* this point. Rays from another point on the object, such as her belt buckle, will also cross at another common point, forming a complete image, as shown. Although three rays are traced in Figure 3.3.10.7, only two are necessary to locate the image. Before applying ray tracing to other situations, let us consider the example shown in Figure 3.3.10.7 in more detail.

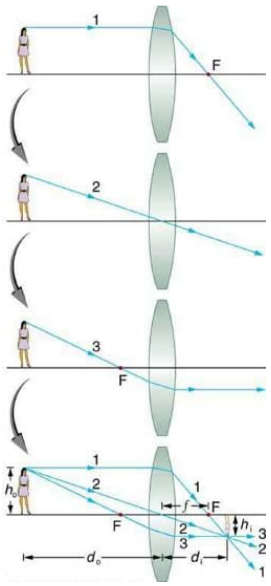


Figure 3.3.10.7: Ray tracing is used to locate the image formed by a lens. Rays originating from the same point on the object are traced—the three chosen rays each follow one of the rules for ray tracing, so that their paths are easy to determine. The image is located at the point where the rays cross. In this case, a real image—one that can be projected on a screen—is formed.

The image formed in Figure 3.3.10.7 is a **real image**, meaning that it can be projected. That is, light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye, for example. Figure 3.3.10.8 shows how such an image would be projected onto film by a camera lens. This figure also shows how a real image is projected onto the retina by the lens of an eye. Note that the image is there whether it is projected onto a screen or not.

Definition: REAL IMAGE

The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.

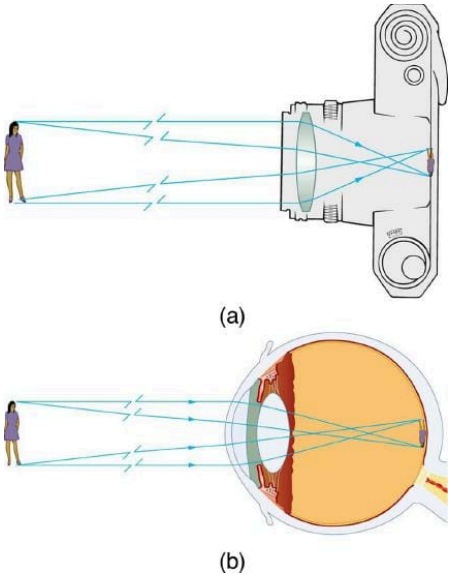


Figure 3.3.10.8: Real images can be projected. (a) A real image of the person is projected onto film. (b) The converging nature of the multiple surfaces that make up the eye result in the projection of a real image on the retina.

Several important distances appear in Figure 3.3.10.7. We define d_{o} to be the object distance, the distance of an object from the center of a lens. **Image distance** d_{i} is defined to be the distance of the image from the center of a lens. The height of the object and height of the image are given the symbols h_{o} and h_{i}, respectively. Images that appear upright relative to the object have heights that are positive and those that are inverted have negative heights. Using the rules of ray tracing and making a scale drawing with paper and pencil, like that in Figure 3.3.10.8, we can accurately describe the location and size of an image. But the real benefit of ray tracing is in visualizing how images are formed in a variety of situations. To obtain numerical information, we use a pair of equations that can be derived from a geometric analysis of ray tracing for thin lenses. The **thin lens equations** are

$$\frac{1}{d_{\mathrm{o}}} + \frac{1}{d_{\mathrm{i}}} = \frac{1}{f}$$

and

$$\frac{h_{\mathrm{i}}}{h_{\mathrm{o}}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = m.$$

We define the ratio of image height to object height ($h_{\mathrm{i}}/h_{\mathrm{o}}$) to be the **magnification** m. (The minus sign in the equation above will be discussed shortly.) The thin lens equations are broadly applicable to all situations involving thin lenses (and “thin” mirrors, as we will see later). We will explore many features of image formation in the following worked examples.

Definition: IMAGE DISTANCE

The distance of the image from the center of the lens is called image distance.

THIN LENS EQUATIONS AND MAGNIFICATION

$$\frac{1}{d_{\mathrm{o}}} + \frac{1}{d_{\mathrm{i}}} = \frac{1}{f}$$

$$\frac{h_{\mathrm{i}}}{h_{\mathrm{o}}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = m$$

Example 3.3.10.2: Finding the Image of a Light Bulb Filament by Ray Tracing and by the Thin Lens Equations

A clear glass light bulb is placed 0.750 m from a convex lens having a 0.500 m focal length, as shown in Figure 3.3.10.9. Use ray tracing to get an approximate location for the image. Then use the thin lens equations to calculate (a) the location of the image and (b) its magnification. Verify that ray tracing and the thin lens equations produce consistent results.

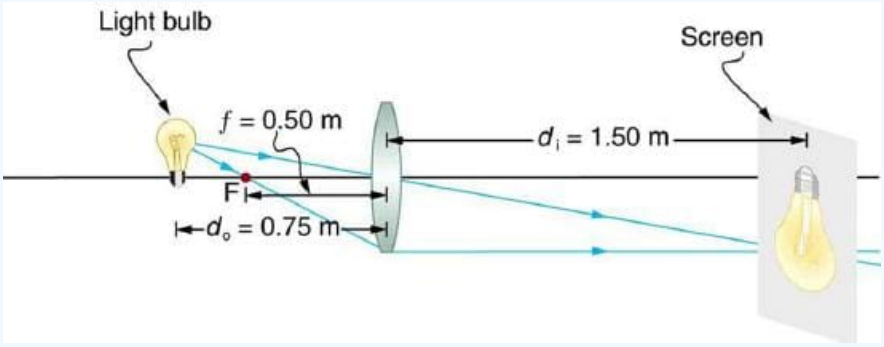


Figure 3.3.10.9: A light bulb placed 0.750 m from a lens having a 0.500 m focal length produces a real image on a poster board as discussed in the example above. Ray tracing predicts the image location and size.

Strategy and Concept

Since the object is placed farther away from a converging lens than the focal length of the lens, this situation is analogous to those illustrated in Figure 3.3.10.7 and Figure 3.3.10.8. Ray tracing to scale should produce similar results for di. Numerical solutions for d_i and m can be obtained using the thin lens equations, noting that $d_\mathrm{o} = 0.750\ \mathrm{m}$ and $f = 0.500\ \mathrm{m}$.

Solutions (Ray tracing)

The ray tracing to scale in Figure 3.3.10.9 shows two rays from a point on the bulb's filament crossing about 1.50 m on the far side of the lens. Thus the image distance d_i is about 1.50 m. Similarly, the image height based on ray tracing is greater than the object height by about a factor of 2, and the image is inverted. Thus m is about –2. The minus sign indicates that the image is inverted.

The thin lens equations can be used to find d_i from the given information:

$$\frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}} = \frac{1}{f}.$$

Rearranging to isolate d_i gives

$$\frac{1}{d_\mathrm{i}} = \frac{1}{f} - \frac{1}{d_\mathrm{o}}.$$

Entering known quantities gives a value for $1/d_\mathrm{i}$:

$$\frac{1}{d_\mathrm{i}} = \frac{1}{0.500\ \mathrm{m}} - \frac{1}{0.750\ \mathrm{m}} = \frac{0.667}{\mathrm{m}}.$$

This must be inverted to find d_i:

$$d_\mathrm{i} = \frac{\mathrm{m}}{0.667} = 1.50\ \mathrm{m}.$$

Note that another way to find d_i is to rearrange the equation:

$$\frac{1}{d_\mathrm{i}} = \frac{1}{f} - \frac{1}{d_\mathrm{o}}.$$

This yields the equation for the image distance as:

$$d_{\mathrm{i}} = \frac{f d_{\mathrm{o}}}{d_{\mathrm{o}} - f}.$$

Note that there is no inverting here.

The thin lens equations can be used to find the magnification m, since both d_{i} and d_{o} are known. Entering their values gives

$$m = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = -\frac{1.50\ \mathrm{m}}{0.750\ \mathrm{m}} = -2.00.$$

Discussion

Note that the minus sign causes the magnification to be negative when the image is inverted. Ray tracing and the use of the thin lens equations produce consistent results. The thin lens equations give the most precise results, being limited only by the accuracy of the given information. Ray tracing is limited by the accuracy with which you can draw, but it is highly useful both conceptually and visually.

Real images, such as the one considered in the previous example, are formed by converging lenses whenever an object is farther from the lens than its focal length. This is true for movie projectors, cameras, and the eye. We shall refer to these as *case 1* images. A case 1 image is formed when $d_0 > f$ and f is positive, as in Figure 3.3.10.10(a). (A summary of the three cases or types of image formation appears at the end of this section.)

A different type of image is formed when an object, such as a person's face, is held close to a convex lens. The image is upright and larger than the object, as seen in Figure 3.3.10.10(b), and so the lens is called a magnifier. If you slowly pull the magnifier away from the face, you will see that the magnification steadily increases until the image begins to blur. Pulling the magnifier even farther away produces an inverted image as seen in Figure 3.3.10.10(a). The distance at which the image blurs, and beyond which it inverts, is the focal length of the lens. To use a convex lens as a magnifier, the object must be closer to the converging lens than its focal length. This is called a *case 2* image. A case 2 image is formed when $d_{\mathrm{o}} < f$ and f is positive.



(a)



(b)

Figure 3.3.10.10: (a) When a converging lens is held farther away from the face than the lens's focal length, an inverted image is formed. This is a case 1 image. Note that the image is in focus but the face is not, because the image is much closer to the camera taking this photograph than the face. (credit: DaMongMan, Flickr) (b) A magnified image of a face is produced by placing it closer to the converging lens than its focal length. This is a case 2 image. (credit: Casey Fleser, Flickr)

Figure 3.3.10.11 uses ray tracing to show how an image is formed when an object is held closer to a converging lens than its focal length. Rays coming from a common point on the object continue to diverge after passing through the lens, but all appear to originate from a point at the location of the image. The image is on the same side of the lens as the object and is farther away from the lens than the object. This image, like all case 2 images, cannot be projected and, hence, is called a **virtual image**. Light

rays only appear to originate at a virtual image; they do not actually pass through that location in space. A screen placed at the location of a virtual image will receive only diffuse light from the object, not focused rays from the lens. Additionally, a screen placed on the opposite side of the lens will receive rays that are still diverging, and so no image will be projected on it. We can see the magnified image with our eyes, because the lens of the eye converges the rays into a real image projected on our retina. Finally, we note that a virtual image is upright and larger than the object, meaning that the magnification is positive and greater than 1.

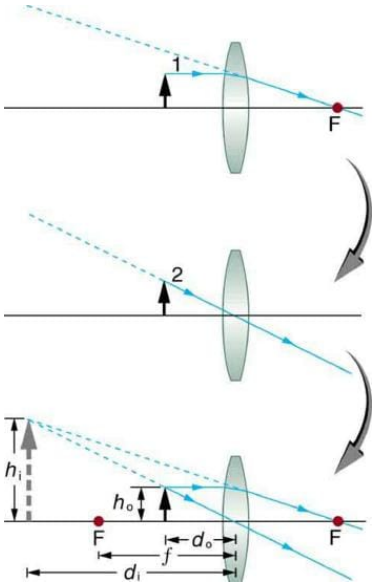


Figure 3.3.10.11: Ray tracing predicts the image location and size for an object held closer to a converging lens than its focal length. Ray 1 enters parallel to the axis and exits through the focal point on the opposite side, while ray 2 passes through the center of the lens without changing path. The two rays continue to diverge on the other side of the lens, but both appear to come from a common point, locating the upright, magnified, virtual image. This is a case 2 image.

Definition: VIRTUAL IMAGE

An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Example 3.3.10.3: Image Produced by a Magnifying Glass

Suppose the book page in Figure 3.3.10.11(a) is held 7.50 cm from a convex lens of focal length 10.0 cm, such as a typical magnifying glass might have. What magnification is produced?

Strategy and Concept

We are given that $d_o = 7.50\ \text{cm}$ and $f = 10.0\ \text{cm}$, so we have a situation where the object is placed closer to the lens than its focal length. We therefore expect to get a case 2 virtual image with a positive magnification that is greater than 1. Ray tracing produces an image like that shown in Figure 3.3.10.11, but we will use the thin lens equations to get numerical solutions in this example.

Solution

To find the magnification m, we try to use magnification equation, $m = -d_i/d_o$. We do not have a value for d_i, so that we must first find the location of the image using lens equation. (The procedure is the same as followed in the preceding example, where d_o and f were known.) Rearranging the magnification equation to isolate d_i gives

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}.$$

Entering known values, we obtain a value for $1/d_i$:

$$\frac{1}{d_\text{i}} = \frac{1}{10.0\ \text{cm}} - \frac{1}{7.50\ \text{cm}} = \frac{-0.0333}{\text{cm}}.$$

This must be inverted to find d_i:

$$d_\text{i} = -\frac{\text{cm}}{0.0333} = -30.0\ \text{cm}.$$

Now the thin lens equation can be used to find the magnification m, since both d_i and d_o are known. Entering their values gives

$$m = -\frac{d_\text{i}}{d_\text{o}} = -\frac{-30.0\ \text{cm}}{7.5\ \text{cm}} = 4.00.$$

Discussion

A number of results in this example are true of all case 2 images, as well as being consistent with Figure 3.3.10.11. Magnification is indeed positive (as predicted), meaning the image is upright. The magnification is also greater than 1, meaning that the image is larger than the object—in this case, by a factor of 3. Note that the image distance is negative. This means the image is on the same side of the lens as the object. Thus the image cannot be projected and is virtual. (Negative values of d_i occur for virtual images.) The image is farther from the lens than the object, since the image distance is greater in magnitude than the object distance. The location of the image is not obvious when you look through a magnifier. In fact, since the image is bigger than the object, you may think the image is closer than the object. But the image is farther away, a fact that is useful in correcting farsightedness, as we shall see in a later section.

A third type of image is formed by a diverging or concave lens. Try looking through eyeglasses meant to correct nearsightedness. (See Figure 3.3.10.12.) You will see an image that is upright but smaller than the object. This means that the magnification is positive but less than 1. The ray diagram in Figure 3.3.10.13 shows that the image is on the same side of the lens as the object and, hence, cannot be projected—it is a virtual image. Note that the image is closer to the lens than the object. This is a *case 3* image, formed for any object by a negative focal length or diverging lens.



Figure 3.3.10.12: A car viewed through a concave or diverging lens looks upright. This is a case 3 image. (credit: Daniel Oines, Flickr)

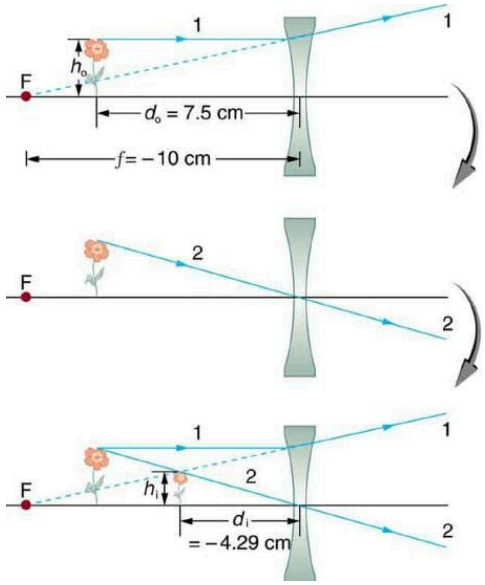


Figure 3.3.10.13: Ray tracing predicts the image location and size for a concave or diverging lens. Ray 1 enters parallel to the axis and is bent so that it appears to originate from the focal point. Ray 2 passes through the center of the lens without changing path. The two rays appear to come from a common point, locating the upright image. This is a case 3 image, which is closer to the lens than the object and smaller in height.

Example 3.3.10.4: Image Produced by a Concave Lens

Suppose an object such as a book page is held 7.50 cm from a concave lens of focal length –10.0 cm. Such a lens could be used in eyeglasses to correct pronounced nearsightedness. What magnification is produced?

Strategy and Concept

This example is identical to the preceding one, except that the focal length is negative for a concave or diverging lens. The method of solution is thus the same, but the results are different in important ways.

Solution

To find the magnification m, we must first find the image distance d_i using thin lens equation

$$\frac{1}{d_\mathrm{i}} = \frac{1}{f} - \frac{1}{d_\mathrm{o}},$$

or its alternative rearrangement

$$d_i = \frac{f d_\mathrm{o}}{d_\mathrm{o} - f}.$$

We are given that $f = -10.0\ \mathrm{cm}$ and $d_\mathrm{o} = 7.50\ \mathrm{cm}$. Entering these yields a value for $1/d_\mathrm{i}$:

$$\frac{1}{d_\mathrm{i}} = \frac{1}{-10.0\ \mathrm{cm}} - \frac{1}{7.50\ \mathrm{cm}} = \frac{-0.2333}{\mathrm{cm}}.$$

This must be inverted to find d_i:

$$d_\mathrm{i} = -\frac{\mathrm{cm}}{0.2333} = -4.29\ \mathrm{cm}.$$

Or

$$d_\mathrm{i} = \frac{(7.5)(-10)}{(7.5 - (-10))} = -75/17.5 = -4.29\ \mathrm{cm}.$$

Now the magnification equation can be used to find the magnification m, since both d_i and d_o are known. Entering their values gives

$$m = -\frac{d_\mathrm{i}}{d_\mathrm{o}} = -\frac{-4.29\ \mathrm{cm}}{7.50\ \mathrm{cm}} = 0.571.$$

Discussion

A number of results in this example are true of all case 3 images, as well as being consistent with Figure 3.3.10.13. Magnification is positive (as predicted), meaning the image is upright. The magnification is also less than 1, meaning the image is smaller than the object—in this case, a little over half its size. The image distance is negative, meaning the image is on the same side of the lens as the object. (The image is virtual.) The image is closer to the lens than the object, since the image distance is smaller in magnitude than the object distance. The location of the image is not obvious when you look through a concave lens. In fact, since the image is smaller than the object, you may think it is farther away. But the image is closer than the object, a fact that is useful in correcting nearsightedness, as we shall see in a later section.

Table 3.3.10.1 summarizes the three types of images formed by single thin lenses. These are referred to as case 1, 2, and 3 images. Convex (converging) lenses can form either real or virtual images (cases 1 and 2, respectively), whereas concave (diverging) lenses can form only virtual images (always case 3). Real images are always inverted, but they can be either larger or smaller than the object. For example, a slide projector forms an image larger than the slide, whereas a camera makes an image smaller than the object being photographed. Virtual images are always upright and cannot be projected. Virtual images are larger than the object only in case 2, where a convex lens is used. The virtual image produced by a concave lens is always smaller than the object—a case 3 image. We can see and photograph virtual images only by using an additional lens to form a real image.

Table 3.3.10.1*: Three Types of Images Formed By Thin Lenses*

Type	Formed when	Image type	d_i	m
Case 1	f positive, $d_o > f$	real	positive	negative
Case 2	f positive, $d_o < f$	virtual	negative	positive, $m > 1$
Case 3	f negative	virtual	negative	positive, $m < 1$

TAKE-HOME EXPERIMENT: CONCENTRATING SUNLIGHT

Find several lenses and determine whether they are converging or diverging. In general those that are thicker near the edges are diverging and those that are thicker near the center are converging. On a bright sunny day take the converging lenses outside and try focusing the sunlight onto a piece of paper. Determine the focal lengths of the lenses. Be careful because the paper may start to burn, depending on the type of lens you have selected.

Section Summary

- Light rays entering a converging lens parallel to its axis cross one another at a single point on the opposite side.
- For a converging lens, the focal point is the point at which converging light rays cross; for a diverging lens, the focal point is the point from which diverging light rays appear to originate.
- The distance from the center of the lens to its focal point is called the focal length f.
- Power P of a lens is defined to be the inverse of its focal length, $P = \frac{1}{f}$.
- A lens that causes the light rays to bend away from its axis is called a diverging lens.
- Ray tracing is the technique of graphically determining the paths that light rays take.
- The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.
- Thin lens equations are $\frac{1}{d_o} + \frac{1}{d_\mathrm{i}} = \frac{1}{f}$ and $\frac{h_\mathrm{i}}{h_\mathrm{o}} = -\frac{d_\mathrm{i}}{d_\mathrm{o}} = m$ (magnification).
- The distance of the image from the center of the lens is called image distance.
- An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Glossary

converging lens
: a convex lens in which light rays that enter it parallel to its axis converge at a single point on the opposite side

diverging lens
: a concave lens in which light rays that enter it parallel to its axis bend away (diverge) from its axis

focal point
: for a converging lens or mirror, the point at which converging light rays cross; for a diverging lens or mirror, the point from which diverging light rays appear to originate

focal length
: distance from the center of a lens or curved mirror to its focal point

magnification
: ratio of image height to object height

power
: inverse of focal length

real image
: image that can be projected

virtual image
: image that cannot be projected

This page titled 3.3.10: Image Formation by Lenses is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.9: Image Formation by Lenses by Andrew Park is licensed CC BY 4.0.

3.3.11: Image Formation by Mirrors

Learning Objectives

- Illustrate image formation in a flat mirror.
- Explain with ray diagrams the formation of an image using spherical mirrors.
- Determine focal length and magnification given radius of curvature, distance of object and image.

We only have to look as far as the nearest bathroom to find an example of an image formed by a mirror. Images in flat mirrors are the same size as the object and are located behind the mirror. Like lenses, mirrors can form a variety of images. For example, dental mirrors may produce a magnified image, just as makeup mirrors do. Security mirrors in shops, on the other hand, form images that are smaller than the object. We will use the law of reflection to understand how mirrors form images, and we will find that mirror images are analogous to those formed by lenses.

Figure 3.3.11.1 helps illustrate how a flat mirror forms an image. Two rays are shown emerging from the same point, striking the mirror, and being reflected into the observer's eye. The rays can diverge slightly, and both still get into the eye. If the rays are extrapolated backward, they seem to originate from a common point behind the mirror, locating the image. (The paths of the reflected rays into the eye are the same as if they had come directly from that point behind the mirror.) Using the law of reflection—the angle of reflection equals the angle of incidence—we can see that the image and object are the same distance from the mirror. This is a virtual image, since it cannot be projected—the rays only appear to originate from a common point behind the mirror. Obviously, if you walk behind the mirror, you cannot see the image, since the rays do not go there. But in front of the mirror, the rays behave exactly as if they had come from behind the mirror, so that is where the image is situated.

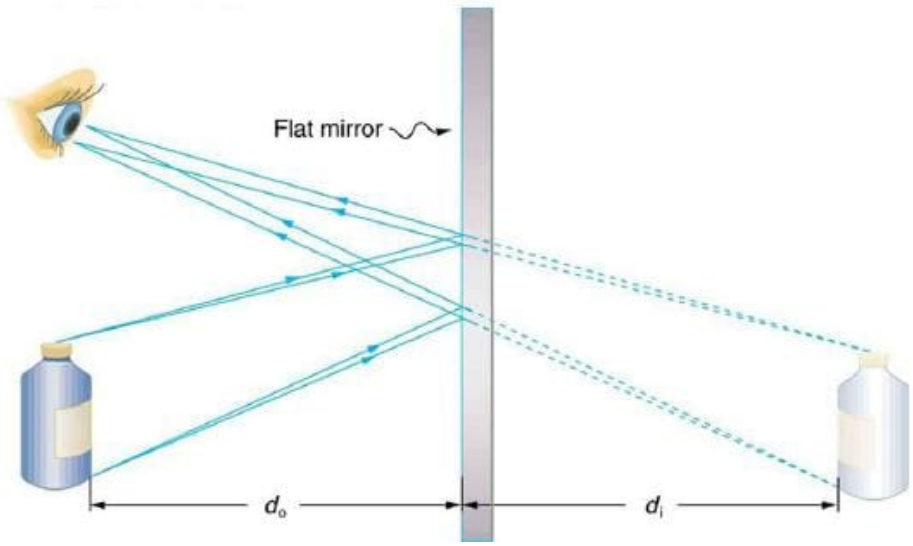


Figure 3.3.11.1: Two sets of rays from common points on an object are reflected by a flat mirror into the eye of an observer. The reflected rays seem to originate from behind the mirror, locating the virtual image.

Now let us consider the focal length of a mirror—for example, the concave spherical mirrors in Figure 3.3.11.2. Rays of light that strike the surface follow the law of reflection. For a mirror that is large compared with its radius of curvature, as in Figure 3.3.11.1 (a), we see that the reflected rays do not cross at the same point, and the mirror does not have a well-defined focal point. If the mirror had the shape of a parabola, the rays would all cross at a single point, and the mirror would have a well-defined focal point. But parabolic mirrors are much more expensive to make than spherical mirrors. The solution is to use a mirror that is small compared with its radius of curvature, as shown in Figure 3.3.11.2(b). (This is the mirror equivalent of the thin lens approximation.) To a very good approximation, this mirror has a well-defined focal point at F that is the focal distance f from the center of the mirror. The focal length f of a concave mirror is positive, since it is a converging mirror.

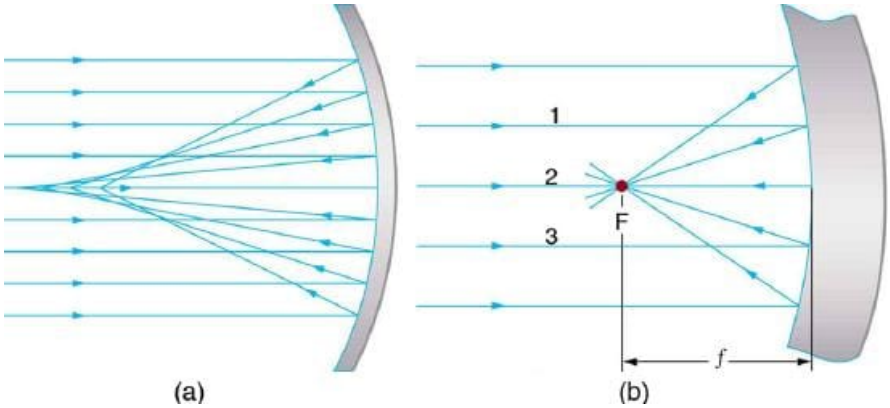


Figure 3.3.11.2: (a) Parallel rays reflected from a large spherical mirror do not all cross at a common point. (b) If a spherical mirror is small compared with its radius of curvature, parallel rays are focused to a common point. The distance of the focal point from the center of the mirror is its focal length f. Since this mirror is converging, it has a positive focal length.

Just as for lenses, the shorter the focal length, the more powerful the mirror; thus, $P = 1/f$ for a mirror, too. A more strongly curved mirror has a shorter focal length and a greater power. Using the law of reflection and some simple trigonometry, it can be shown that the focal length is half the radius of curvature, or

$$f = \frac{R}{2},$$

where R is the radius of curvature of a spherical mirror. The smaller the radius of curvature, the smaller the focal length and, thus, the more powerful the mirror.

The convex mirror shown in Figure 3.3.11.3 also has a focal point. Parallel rays of light reflected from the mirror seem to originate from the point F at the focal distance f behind the mirror. The focal length and power of a convex mirror are negative, since it is a diverging mirror.

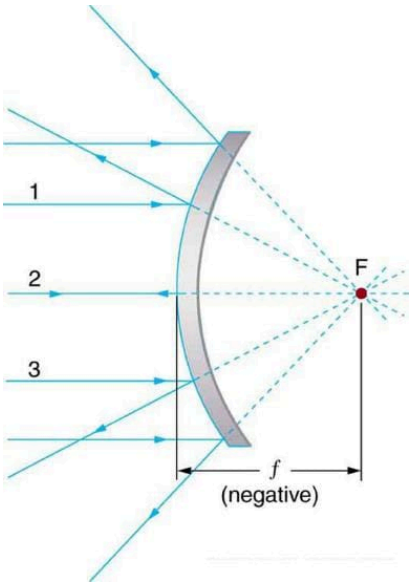


Figure 3.3.11.3: Parallel rays of light reflected from a convex spherical mirror (small in size compared with its radius of curvature) seem to originate from a well-defined focal point at the focal distance f behind the mirror. Convex mirrors diverge light rays and, thus, have a negative focal length.

Ray tracing is as useful for mirrors as for lenses. The rules for ray tracing for mirrors are based on the illustrations just discussed:

1. **Principal Ray 1**: a ray approaching a concave converging mirror parallel to its axis is reflected through the focal point F of the mirror on the same side. (See rays 1 and 3 in Figure 3.3.11.2(b).); a ray approaching a convex diverging mirror parallel to its axis is reflected so that it seems to come from the focal point F behind the mirror. (See rays 1 and 3 in Figure 3.3.11.3.)
2. **Principal Ray 2**: any ray striking the center of a mirror is followed by applying the law of reflection; it makes the same angle

with the axis when leaving as when approaching. (See ray 2 in Figure 3.3.11.4.)

3. **Principal Ray 3**: a ray approaching a concave converging mirror through its focal point is reflected parallel to its axis. (The reverse of rays 1 and 3 in Figure 3.3.11.2.); a ray approaching a convex diverging mirror by heading toward its focal point on the opposite side is reflected parallel to the axis. (The reverse of rays 1 and 3 in Figure 3.3.11.3.) *The third principal ray is optional and may be used to verify the accuracy of image location.*

We will use ray tracing to illustrate how images are formed by mirrors, and we can use ray tracing quantitatively to obtain numerical information. But since we assume each mirror is small compared with its radius of curvature, we can use the thin lens equations for mirrors just as we did for lenses.

Consider the situation shown in Figure 3.3.11.4, concave spherical mirror reflection, in which an object is placed farther from a concave (converging) mirror than its focal length. That is, f is positive and $d_o > f$, so that we may expect an image similar to the case 1 real image formed by a converging lens. Ray tracing in Figure 3.3.11.4 shows that the rays from a common point on the object all cross at a point on the same side of the mirror as the object. Thus a real image can be projected onto a screen placed at this location. The image distance is positive, and the image is inverted, so its magnification is negative. This is a *case 1 image for mirrors*. It differs from the case 1 image for lenses only in that the image is on the same side of the mirror as the object. It is otherwise identical.

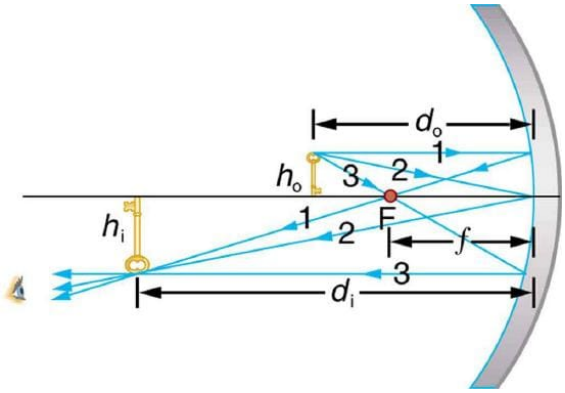


Figure 3.3.11.4: A case 1 image for a mirror. An object is farther from the converging mirror than its focal length. Rays from a common point on the object are traced using the rules in the text. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 goes through the focal point on the way toward the mirror. All three rays cross at the same point after being reflected, locating the inverted real image. Although three rays are shown, only two of the three are needed to locate the image and determine its height.

Example 3.3.11.1: A Concave Reflector

Electric room heaters use a concave mirror to reflect infrared (IR) radiation from hot coils. Note that IR follows the same law of reflection as visible light. Given that the mirror has a radius of curvature of 50.0 cm and produces an image of the coils 3.00 m away from the mirror, where are the coils?

Strategy and Concept

We are given that the concave mirror projects a real image of the coils at an image distance $d_i = 3.00\ \text{m}$. The coils are the object, and we are asked to find their location—that is, to find the object distance d_o. We are also given the radius of curvature of the mirror, so that its focal length is $f = R/2 = 25.0\ \text{cm}$ (positive since the mirror is concave or converging). Assuming the mirror is small compared with its radius of curvature, we can use the thin lens equations, to solve this problem.

Solution

Since d_i and f are known, thin lens equation can be used to find d_o:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}.$$

Rearranging to isolate d_o gives

$$\frac{1}{d_o} = \frac{1}{f} - \frac{1}{d_i}.$$

Entering known quantities gives a value for $1/d_o$:

$$\frac{1}{d_o} = \frac{1}{0.250\ \text{m}} - \frac{1}{3.00\ \text{m}} = \frac{3.667}{\text{m}}.$$

This must be inverted to find d_o:

$$d_o = \frac{1\ \text{m}}{3.667} = 27.3\ \text{cm}.$$

Discussion

Note that the object (the filament) is farther from the mirror than the mirror's focal length. This is a case 1 image ($d_o > f$ and f positive), consistent with the fact that a real image is formed. You will get the most concentrated thermal energy directly in front of the mirror and 3.00 m away from it. Generally, this is not desirable, since it could cause burns. Usually, you want the rays to emerge parallel, and this is accomplished by having the filament at the focal point of the mirror.

Note that the filament here is not much farther from the mirror than its focal length and that the image produced is considerably farther away. This is exactly analogous to a slide projector. Placing a slide only slightly farther away from the projector lens than its focal length produces an image significantly farther away. As the object gets closer to the focal distance, the image gets farther away. In fact, as the object distance approaches the focal length, the image distance approaches infinity and the rays are sent out parallel to one another.

Example 3.3.11.2: Solar Electric Generating System

One of the solar technologies used today for generating electricity is a device (called a parabolic trough or concentrating collector) that concentrates the sunlight onto a blackened pipe that contains a fluid. This heated fluid is pumped to a heat exchanger, where its heat energy is transferred to another system that is used to generate steam—and so generate electricity through a conventional steam cycle. Figure 3.3.11.5 shows such a working system in southern California. Concave mirrors are used to concentrate the sunlight onto the pipe. The mirror has the approximate shape of a section of a cylinder. For the problem, assume that the mirror is exactly one-quarter of a full cylinder.

a. If we wish to place the fluid-carrying pipe 40.0 cm from the concave mirror at the mirror's focal point, what will be the radius of curvature of the mirror?
b. Per meter of pipe, what will be the amount of sunlight concentrated onto the pipe, assuming the insolation (incident solar radiation) is $0.900\ \text{kW/m}^2$?
c. If the fluid-carrying pipe has a 2.00-cm diameter, what will be the temperature increase of the fluid per meter of pipe over a period of one minute? Assume all the solar radiation incident on the reflector is absorbed by the pipe, and that the fluid is mineral oil.

Strategy

To solve an *Integrated Concept Problem* we must first identify the physical principles involved. Part (a) is related to the current topic. Part (b) involves a little math, primarily geometry. Part (c) requires an understanding of heat and density.

Solution to (a)

To a good approximation for a concave or semi-spherical surface, the point where the parallel rays from the sun converge will be at the focal point, so $R = 2f = 80.0\ \text{cm}$.

Solution to (b)

The insolation is $900\ \text{W/m}^2$. We must find the cross-sectional area A of the concave mirror, since the power delivered is $900\ \text{W/m}^2 \times \text{A}$. The mirror in this case is a quarter-section of a cylinder, so the area for a length L of the mirror is $\text{A} = \frac{1}{4}(2\pi R)\text{L}$. The area for a length of 1.00 m is then

$$\mathrm{A} = \frac{\pi}{2} R(1.00\ \mathrm{m}) = \frac{(3.14)}{2}(0.800\ \mathrm{m})(1.00\ \mathrm{m}) = 1.26\ \mathrm{m}^2.$$

The insolation on the 1.00-m length of pipe is then

$$\left(9.00 \times 10^2 \frac{\mathrm{W}}{\mathrm{m}^2}\right)\left(1.26\ \mathrm{m}^2\right) = 1130\ \mathrm{W}.$$

Solution to (c)

The increase in temperature is given by $Q = mc\Delta T$. The mass m of the mineral oil in the one-meter section of pipe is

$$\begin{aligned} m = \rho \mathrm{V} &= \rho\pi \left(\frac{d}{2}\right)^2 (1.00\ \mathrm{m}) \\ &= \left(8.00 \times 10^2\ \mathrm{kg/m}^3\right)(3.14)(0.0100\ \mathrm{m})^2(1.00\ \mathrm{m}) \\ &= 0.251\ \mathrm{kg}. \end{aligned}$$

Therefore, the increase in temperature in one minute is

$$\begin{aligned} \Delta T &= Q/mc \\ &= \frac{(1130\ \mathrm{W})(60.0\ \mathrm{s})}{(0.251\ \mathrm{kg})\,(1670\ \mathrm{J \cdot kg/^\circ C})} \\ &= 162^\circ \mathrm{C}. \end{aligned}$$

Discussion for (c)

An array of such pipes in the California desert can provide a thermal output of 250 MW on a sunny day, with fluids reaching temperatures as high as $400^\circ \mathrm{C}$. We are considering only one meter of pipe here, and ignoring heat losses along the pipe.



Figure 3.3.11.5: Parabolic trough collectors are used to generate electricity in southern California. (credit: kjkolb, Wikimedia Commons)

What happens if an object is closer to a concave mirror than its focal length? This is analogous to a case 2 image for lenses ($d_\mathrm{o} < f$ and f positive), which is a magnifier. In fact, this is how makeup mirrors act as magnifiers. Figure 3.3.11.6(a) uses ray tracing to locate the image of an object placed close to a concave mirror. Rays from a common point on the object are reflected in such a manner that they appear to be coming from behind the mirror, meaning that the image is virtual and cannot be projected. As with a magnifying glass, the image is upright and larger than the object. This is a *case 2 image for mirrors* and is exactly analogous to that for lenses.

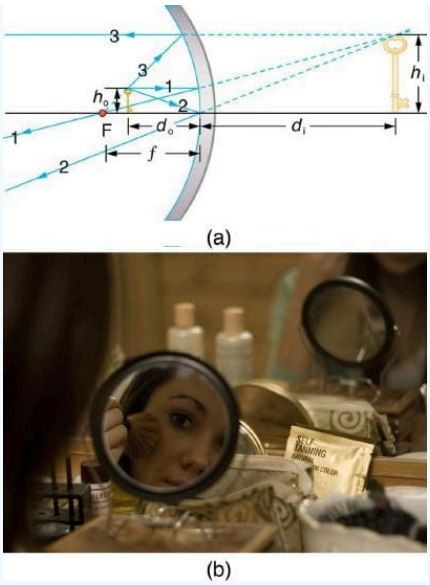


Figure 3.3.11.6: (a) Case 2 images for mirrors are formed when a converging mirror has an object closer to it than its focal length. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches the mirror as if it came from the focal point. (b) A magnifying mirror showing the reflection. (credit: Mike Melrose, Flickr)

All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be larger than the object. (b) Makeup mirrors are perhaps the most common use of a concave mirror to produce a larger, upright image.

A convex mirror is a diverging mirror (f is negative) and forms only one type of image. It is a *case 3* image—one that is upright and smaller than the object, just as for diverging lenses. Figure 3.3.11.7(a) uses ray tracing to illustrate the location and size of the case 3 image for mirrors. Since the image is behind the mirror, it cannot be projected and is thus a virtual image. It is also seen to be smaller than the object.

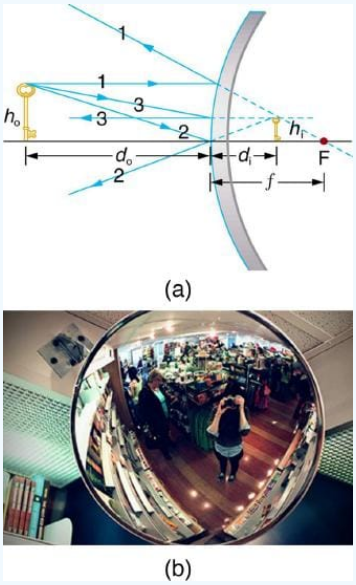


Figure 3.3.11.7: Case 3 images for mirrors are formed by any convex mirror. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches toward the focal point. All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be smaller than the object. (b) Security mirrors are convex, producing a smaller, upright image. Because the image is smaller, a larger area is imaged compared to what would be observed for a flat mirror (and hence security is improved). (credit: Laura D'Alessandro, Flickr)

Example 3.3.11.3: Image in a Convex Mirror

A keratometer is a device used to measure the curvature of the cornea, particularly for fitting contact lenses. Light is reflected from the cornea, which acts like a convex mirror, and the keratometer measures the magnification of the image. The smaller the magnification, the smaller the radius of curvature of the cornea. If the light source is 12.0 cm from the cornea and the image's magnification is 0.0320, what is the cornea's radius of curvature?

Strategy

If we can find the focal length of the convex mirror formed by the cornea, we can find its radius of curvature (the radius of curvature is twice the focal length of a spherical mirror). We are given that the object distance is $d_{\text{o}} = 12.0 \text{ cm}$ and that $m = 0.0320$. We first solve for the image distance d_{i}, and then for f.

Solution

$m = -d_{\text{i}}/d_{\text{o}}$. Solving this expression for d_{i} gives

$$d_{\text{i}} = -md_{\text{o}}.$$

Entering known values yields

$$d_{\text{i}} = -(0.0320)(12.0 \text{ cm}) = -0.384 \text{ cm}.$$

$$\frac{1}{f} = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}}$$

Substituting known values,

$$\frac{1}{f} = \frac{1}{12.0 \text{ cm}} + \frac{1}{-0.384 \text{ cm}} = \frac{-2.52}{\text{cm}}.$$

This must be inverted to find f:

$$f = \frac{\text{cm}}{-2.52} = -0.400 \text{ cm}.$$

The radius of curvature is twice the focal length, so that

$$R = 2|f| = 0.800 \text{ cm}.$$

Discussion

Although the focal length f of a convex mirror is defined to be negative, we take the absolute value to give us a positive value for R. The radius of curvature found here is reasonable for a cornea. The distance from cornea to retina in an adult eye is about 2.0 cm. In practice, many corneas are not spherical, complicating the job of fitting contact lenses. Note that the image distance here is negative, consistent with the fact that the image is behind the mirror, where it cannot be projected. In this section's Problems and Exercises, you will show that for a fixed object distance, the smaller the radius of curvature, the smaller the magnification.

The three types of images formed by mirrors (cases 1, 2, and 3) are exactly analogous to those formed by lenses, as summarized in the table at the end of "Image Formation by Lenses". It is easiest to concentrate on only three types of images—then remember that concave mirrors act like convex lenses, whereas convex mirrors act like concave lenses.

TAKE-HOME EXPERIMENT: CONCAVE MIRRORS CLOSE TO HOME

Find a flashlight and identify the curved mirror used in it. Find another flashlight and shine the first flashlight onto the second one, which is turned off. Estimate the focal length of the mirror. You might try shining a flashlight on the curved mirror behind the headlight of a car, keeping the headlight switched off, and determine its focal length.

Section Summary

- The characteristics of an image formed by a flat mirror are: (a) The image and object are the same distance from the mirror, (b) The image is a virtual image, and (c) The image is situated behind the mirror.
- Image length is half the radius of curvature.

$$f = \frac{R}{2}$$

- A convex mirror is a diverging mirror and forms only one type of image, namely a virtual image.

Glossary

converging mirror
: a concave mirror in which light rays that strike it parallel to its axis converge at one or more points along the axis

diverging mirror
: a convex mirror in which light rays that strike it parallel to its axis bend away (diverge) from its axis

law of reflection
: angle of reflection equals the angle of incidence

This page titled 3.3.11: Image Formation by Mirrors is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.10: Image Formation by Mirrors by Andrew Park is licensed CC BY 4.0.

3.3.12: Light (Exercises)

Conceptual Questions

11.4: The Electromagnetic Spectrum- Application Notes

1. If you live in a region that has a particular TV station, you can sometimes pick up some of its audio portion on your FM radio receiver. Explain how this is possible. Does it imply that TV audio is broadcast as FM?

2. Explain why people who have the lens of their eye removed because of cataracts are able to see low-frequency ultraviolet.

3. How do fluorescent soap residues make clothing look “brighter and whiter” in outdoor light? Would this be effective in candlelight?

4. Give an example of resonance in the reception of electromagnetic waves.

5. Illustrate that the size of details of an object that can be detected with electromagnetic waves is related to their wavelength, by comparing details observable with two different types (for example, radar and visible light or infrared and X-rays).

6. Why don’t buildings block radio waves as completely as they do visible light?

7. Make a list of some everyday objects and decide whether they are transparent or opaque to each of the types of electromagnetic waves.

8. Your friend says that more patterns and colors can be seen on the wings of birds if viewed in ultraviolet light. Would you agree with your friend? Explain your answer.

9. The rate at which information can be transmitted on an electromagnetic wave is proportional to the frequency of the wave. Is this consistent with the fact that laser telephone transmission at visible frequencies carries far more conversations per optical fiber than conventional electronic transmission in a wire? What is the implication for ELF radio communication with submarines?

10. Give an example of energy carried by an electromagnetic wave.

11. In an MRI scan, a higher magnetic field requires higher frequency radio waves to resonate with the nuclear type whose density and location is being imaged. What effect does going to a larger magnetic field have on the most efficient antenna to broadcast those radio waves? Does it favor a smaller or larger antenna?

12. Laser vision correction often uses an excimer laser that produces 193-nm electromagnetic radiation. This wavelength is extremely strongly absorbed by the cornea and ablates it in a manner that reshapes the cornea to correct vision defects. Explain how the strong absorption helps concentrate the energy in a thin layer and thus give greater accuracy in shaping the cornea. Also explain how this strong absorption limits damage to the lens and retina of the eye.

11.5: Reflection

13. Using the law of reflection, explain how powder takes the shine off of a person’s nose. What is the name of the optical effect?

11.6: Refraction

14. Diffusion by reflection from a rough surface is described in this chapter. Light can also be diffused by refraction. Describe how this occurs in a specific situation, such as light interacting with crushed ice.

15. Why is the index of refraction always greater than or equal to 1?

16. Does the fact that the light flash from lightning reaches you before its sound prove that the speed of light is extremely large or simply that it is greater than the speed of sound? Discuss how you could use this effect to get an estimate of the speed of light.

17. Will light change direction toward or away from the perpendicular when it goes from air to water? Water to glass? Glass to air?

18. Explain why an object in water always appears to be at a depth shallower than it actually is? Why do people sometimes sustain neck and spinal injuries when diving into unfamiliar ponds or waters?

19. Explain why a person's legs appear very short when wading in a pool. Justify your explanation with a ray diagram showing the path of rays from the feet to the eye of an observer who is out of the water.

20. Why is the front surface of a thermometer curved as shown?

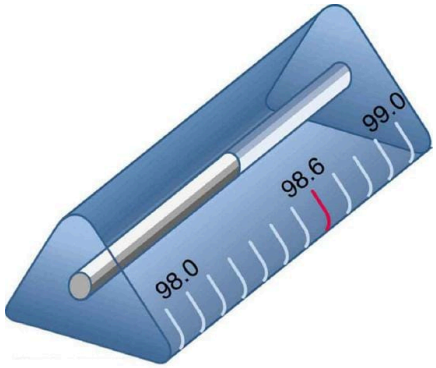


Figure 3.3.12.1: The curved surface of the thermometer serves a purpose.

21. Suppose light were incident from air onto a material that had a negative index of refraction, say –1.3; where does the refracted light ray go?

11.8: Image Formation by Lenses

22. It can be argued that a flat piece of glass, such as in a window, is like a lens with an infinite focal length. If so, where does it form an image? That is, how are d_{i} and d_{o} related?

23. You can often see a reflection when looking at a sheet of glass, particularly if it is darker on the other side. Explain why you can often see a double image in such circumstances.

24. When you focus a camera, you adjust the distance of the lens from the film. If the camera lens acts like a thin lens, why can it not be a fixed distance from the film for both near and distant objects?

25. A thin lens has two focal points, one on either side, at equal distances from its center, and should behave the same for light entering from either side. Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they are thin lenses.

26. Will the focal length of a lens change when it is submerged in water? Explain.

11.9: Image Formation by Mirrors

27. What are the differences between real and virtual images? How can you tell (by looking) whether an image formed by a single lens or mirror is real or virtual?

28. Can you see a virtual image? Can you photograph one? Can one be projected onto a screen with additional lenses or mirrors? Explain your responses.

29. Is it necessary to project a real image onto a screen for it to exist?

30. At what distance is an image *always* located—at d_o, d_i, or f?

31. Under what circumstances will an image be located at the focal point of a lens or mirror?

32. What is meant by a negative magnification? What is meant by a magnification that is less than 1 in magnitude?

33. Can a case 1 image be larger than the object even though its magnification is always negative? Explain.

34. Figure 3.3.12.2 shows a light bulb between two mirrors. One mirror produces a beam of light with parallel rays; the other keeps light from escaping without being put into the beam. Where is the filament of the light in relation to the focal point or radius of curvature of each mirror?

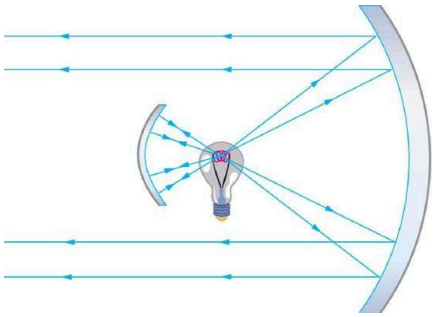


Figure 3.3.12.2: The two mirrors trap most of the bulb's light and form a directional beam as in a headlight.

35. Devise an arrangement of mirrors allowing you to see the back of your head. What is the minimum number of mirrors needed for this task?

36. If you wish to see your entire body in a flat mirror (from head to toe), how tall should the mirror be? Does its size depend upon your distance away from the mirror? Provide a sketch.

37. It can be argued that a flat mirror has an infinite focal length. If so, where does it form an image? That is, how are d_i and d_o related?

38. Why are diverging mirrors often used for rear-view mirrors in vehicles? What is the main disadvantage of using such a mirror compared with a flat one?

11.10: Polarization

39. Under what circumstances is the phase of light changed by reflection? Is the phase related to polarization?

40. Can a sound wave in air be polarized? Explain.

41. No light passes through two perfect polarizing filters with perpendicular axes. However, if a third polarizing filter is placed between the original two, some light can pass. Why is this? Under what circumstances does most of the light pass?

42. Explain what happens to the energy carried by light that it is dimmed by passing it through two crossed polarizing filters.

43. When particles scattering light are much smaller than its wavelength, the amount of scattering is proportional to $1/\lambda^4$. Does this mean there is more scattering for small λ than large λ? How does this relate to the fact that the sky is blue?

44. Using the information given in the preceding question, explain why sunsets are red.

Problems & Exercises

11.6: Refraction

45. What is the speed of light in water? In glycerine?

Solution

$2.25 \times 10^8 \text{ m/s}$ in water

$2.04 \times 10^8 \text{ m/s}$ in glycerine

46. What is the speed of light in air? In crown glass?

47. Calculate the index of refraction for a medium in which the speed of light is $2.012 \times 10^8 \text{ m/s}$, and identify the most likely substance based on Table 12.6.1.

Solution

1.490, polystyrene

48. In what substance in Table 12.6.1 is the speed of light $2.290 \times 10^8 \text{ m/s}$?

49. There was a major collision of an asteroid with the Moon in medieval times. It was described by monks at Canterbury Cathedral in England as a red glow on and around the Moon. How long after the asteroid hit the Moon, which is $3.84 \times 10^5 \text{ km}$ away, would the light first arrive on Earth?

Solution

1.28 s

50. Components of some computers communicate with each other through optical fibers having an index of refraction $n = 1.55$. What time in nanoseconds is required for a signal to travel 0.200 m through such a fiber?

Solution

1.03 ns

51. On the Moon's surface, lunar astronauts placed a corner reflector, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. What percent correction is needed to account for the delay in time due to the slowing of light in Earth's atmosphere? Assume the distance to the Moon is precisely $3.84 \times 10^8 \text{ m}$, and Earth's atmosphere (which varies in density with altitude) is equivalent to a layer 30.0 km thick with a constant index of refraction $n = 1.000293$.

11.7: Dispersion- The Rainbow and Prisms

52. (a) What is the ratio of the speed of red light to violet light in diamond, based on Table 12.7.1?

(b) What is this ratio in polystyrene?

(c) Which is more dispersive?

This page titled 3.3.12: Light (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 12.12: Light (Exercises) by Andrew Park is licensed CC BY 4.0.

3.4: THERMAL PHYSICS

3.4.1: Introduction to Thermal Physics

3.4.2: Temperature

3.4.3: The Ideal Gas Law

3.4.4: Heat

3.4.5: Heat Transfer Methods

3.4.6: Temperature Change and Heat Capacity

3.4.7: Phase Change and Latent Heat

3.4.8: The First Law of Thermodynamics

3.4.9: The First Law of Thermodynamics and Heat Engine Processes

3.4.10: Introduction to the Second Law of Thermodynamics- Heat Engines and Their Efficiency

3.4.11: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated

3.4.12: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy

3.4.13: Statistical Interpretation of Entropy and the Second Law of Thermodynamics- The Underlying Explanation

3.4.14: Applications of Thermodynamics- Heat Pumps and Refrigerators

3.4.15: Thermal Physics (Exercises)

This page titled 3.4: Thermal Physics is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

3.4.1: Introduction to Thermal Physics



Figure 3.4.1.1: The welder's gloves and helmet protect him from the electric arc that transfers enough thermal energy to melt the rod, spray sparks, and burn the retina of an unprotected eye. The thermal energy can be felt on exposed skin a few meters away, and its light can be seen for kilometers. (credit: Kevin S. O'Brien/U.S. Navy)

Heat is something familiar to each of us. We feel the warmth of the summer Sun, the chill of a clear summer night, the heat of coffee after a winter stroll, and the cooling effect of our sweat. Heat transfer is maintained by temperature differences. Manifestations of **heat transfer**—the movement of heat energy from one place or material to another—are apparent throughout the universe. Heat from beneath Earth's surface is brought to the surface in flows of incandescent lava. The Sun warms Earth's surface and is the source of much of the energy we find on it. Rising levels of atmospheric carbon dioxide threaten to trap more of the Sun's energy, perhaps fundamentally altering the ecosphere. In space, supernovas explode, briefly radiating more heat than an entire galaxy does.

What is heat? How do we define it? How is it related to temperature? What are heat's effects? How is it related to other forms of energy and to work? We will find that, in spite of the richness of the phenomena, there is a small set of underlying physical principles that unite the subjects and tie them to other fields.

Most notably, heat transfer can be used to do work. It can also be converted to any other form of energy. A car engine, for example, burns fuel for heat transfer into a gas. Work is done by the gas as it exerts a force through a distance, converting its energy into a variety of other forms—into the car's kinetic or gravitational potential energy; into electrical energy to run the spark plugs, radio, and lights; and back into stored energy in the car's battery. But most of the heat transfer produced from burning fuel in the engine does not do work on the gas-the engine is quite inefficient.

Basic physical laws govern how heat transfer for doing work takes place and place insurmountable limits onto its efficiency. This chapter will explore these laws as well as some applications and concepts associated with them. These topics are part of thermodynamics—the study of heat transfer and its relationship to doing work.

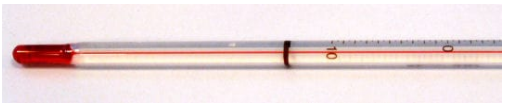


Figure 3.4.1.2: In a typical thermometer like this one, the alcohol, with a red dye, expands more rapidly than the glass containing it. When the thermometer's temperature increases, the liquid from the bulb is forced into the narrow tube, producing a large change in the length of the column for a small change in temperature. (credit: Chemical Engineer, Wikimedia Commons)

This page titled 3.4.1: Introduction to Thermal Physics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.1: Introduction to Thermal Physics by Andrew Park is licensed CC BY 4.0.

3.4.2: Temperature

Learning Objectives

- Define temperature.
- Convert temperatures between the Celsius, Fahrenheit, and Kelvin scales.
- Define thermal equilibrium.
- State the zeroth law of thermodynamics.

The concept of temperature has evolved from the common concepts of hot and cold. Human perception of what feels hot or cold is a relative one. For example, if you place one hand in hot water and the other in cold water, and then place both hands in tepid water, the tepid water will feel cool to the hand that was in hot water, and warm to the one that was in cold water. The scientific definition of temperature is less ambiguous than your senses of hot and cold. **Temperature** is operationally defined to be what we measure with a thermometer. (Many physical quantities are defined solely in terms of how they are measured. We shall see later how temperature is related to the kinetic energies of atoms and molecules, a more physical explanation.) Two accurate thermometers, one placed in hot water and the other in cold water, will show the hot water to have a higher temperature. If they are then placed in the tepid water, both will give identical readings (within measurement uncertainties). In this section, we discuss temperature, its measurement by thermometers, and its relationship to thermal equilibrium. Again, temperature is the quantity measured by a thermometer.

MISCONCEPTION ALERT: HUMAN PERCEPTION VS. REALITY

On a cold winter morning, the wood on a porch feels warmer than the metal of your bike. The wood and bicycle are in thermal equilibrium with the outside air, and are thus the same temperature. They *feel* different because of the difference in the way that they conduct heat away from your skin. The metal conducts heat away from your body faster than the wood does. This is just one example demonstrating that the human sense of hot and cold is not determined by temperature alone.

Another factor that affects our perception of temperature is humidity. Most people feel much hotter on hot, humid days than on hot, dry days. This is because on humid days, sweat does not evaporate from the skin as efficiently as it does on dry days. It is the evaporation of sweat (or water from a sprinkler or pool) that cools us off.

Any physical property that depends on temperature, and whose response to temperature is reproducible, can be used as the basis of a thermometer. Because many physical properties depend on temperature, the variety of thermometers is remarkable. For example, volume increases with temperature for most substances. This property is the basis for the common alcohol thermometer, the old mercury thermometer, and the bimetallic strip (Figure 3.4.2.1). Other properties used to measure temperature include electrical resistance and color, as shown in Figure 3.4.2.2, and the emission of infrared radiation, as shown in Figure 3.4.2.3.

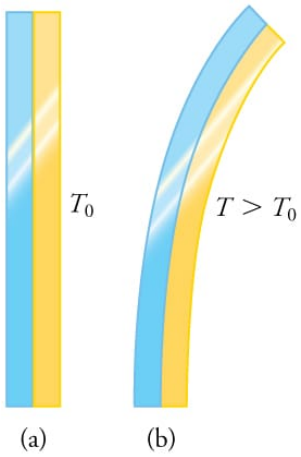


Figure 3.4.2.1: The curvature of a bimetallic strip depends on temperature. (a) The strip is straight at the starting temperature, where its two components have the same length. (b) At a higher temperature, this strip bends to the right, because the metal on the left has expanded more than the metal on the right.



Figure 3.4.2.2: Each of the six squares on this plastic (liquid crystal) thermometer contains a film of a different heat-sensitive liquid crystal material. Below 95°F, all six squares are black. When the plastic thermometer is exposed to temperature that increases to 95°F, the first liquid crystal square changes color. When the temperature increases above 96.8°F the second liquid crystal square also changes color, and so forth. (credit: Arkrishna, Wikimedia Commons)

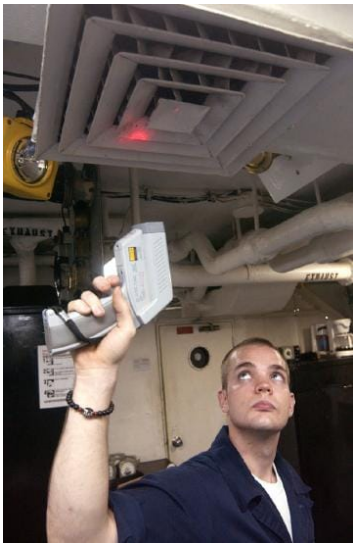


Figure 3.4.2.3: Fireman Jason Ormand uses a pyrometer to check the temperature of an aircraft carrier's ventilation system. Infrared radiation (whose emission varies with temperature) from the vent is measured and a temperature readout is quickly produced. Infrared measurements are also frequently used as a measure of body temperature. These modern thermometers, placed in the ear canal, are more accurate than alcohol thermometers placed under the tongue or in the armpit. (credit: Lamel J. Hinton/U.S. Navy)

Temperature Scales

Thermometers are used to measure temperature according to well-defined scales of measurement, which use pre-defined reference points to help compare quantities. The three most common temperature scales are the Fahrenheit, Celsius, and Kelvin scales. A temperature scale can be created by identifying two easily reproducible temperatures. The freezing and boiling temperatures of water at standard atmospheric pressure are commonly used.

The **Celsius** scale (which replaced the slightly different *centigrade* scale) has the freezing point of water at 0°C and the boiling point at 100°C. Its unit is the **degree Celsius**(°C). On the **Fahrenheit** scale (still the most frequently used in the United States),

the freezing point of water is at 32°F and the boiling point is at 212°F. The unit of temperature on this scale is the **degree Fahrenheit**(°F). Note that a temperature difference of one degree Celsius is greater than a temperature difference of one degree Fahrenheit. Only 100 Celsius degrees span the same range as 180 Fahrenheit degrees, thus one degree on the Celsius scale is 1.8 times larger than one degree on the Fahrenheit scale $180/100 = 9/5$.

The **Kelvin** scale is the temperature scale that is commonly used in science. It is an *absolute temperature* scale defined to have 0 K at the lowest possible temperature, called **absolute zero**. The official temperature unit on this scale is the *kelvin*, which is abbreviated K, and is not accompanied by a degree sign. The freezing and boiling points of water are 273.15 K and 373.15 K, respectively. Thus, the magnitude of temperature differences is the same in units of kelvins and degrees Celsius. Unlike other temperature scales, the Kelvin scale is an absolute scale. It is used extensively in scientific work because a number of physical quantities, such as the volume of an ideal gas, are directly related to absolute temperature. The kelvin is the SI unit used in scientific work.

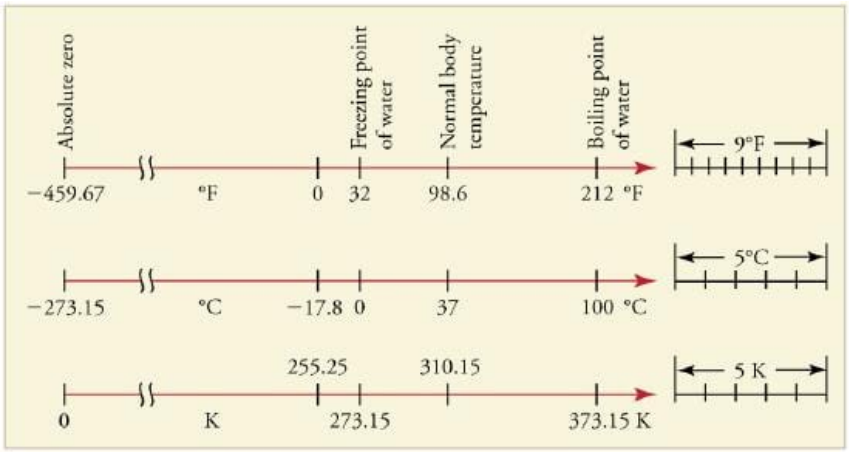


Figure 3.4.2.4: Relationships between the Fahrenheit, Celsius, and Kelvin temperature scales, rounded to the nearest degree. The relative sizes of the scales are also shown.

The relationships between the three common temperature scales is shown in Figure 3.4.2.4.

Temperature Ranges in the Universe

Figure 3.4.2.6 shows the wide range of temperatures found in the universe. Human beings have been known to survive with body temperatures within a small range, from 24°C to 44°C (75°F to 111°F). The average normal body temperature is usually given as 37.0°C (98.6°F), and variations in this temperature can indicate a medical condition: a fever, an infection, a tumor, or circulatory problems (see Figure 3.4.2.5).

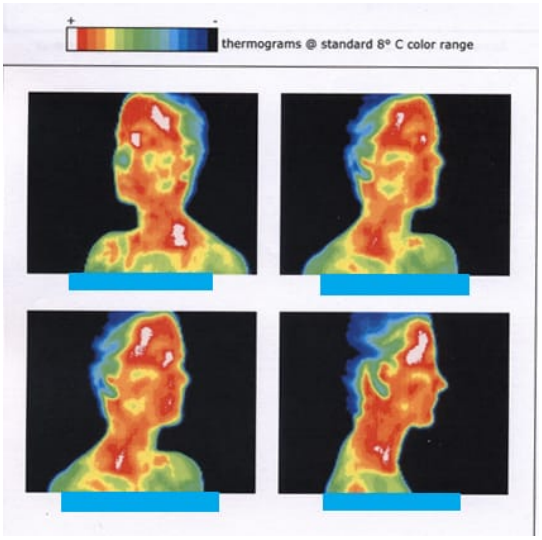


Figure 3.4.2.5: This image of radiation from a person's body (an infrared thermograph) shows the location of temperature abnormalities in the upper body. Dark blue corresponds to cold areas and red to white corresponds to hot areas. An elevated temperature might be an indication of malignant tissue (a cancerous tumor in the breast, for example), while a depressed temperature might be due to a decline in blood flow from a clot. In this case, the abnormalities are caused by a condition called hyperhidrosis. (credit: Porcelina81, Wikimedia Commons)

The lowest temperatures ever recorded have been measured during laboratory experiments: 4.5×10^{-10} K at the Massachusetts Institute of Technology (USA), and 1.0×10^{-10} K at Helsinki University of Technology (Finland). In comparison, the coldest recorded place on Earth's surface is Vostok, Antarctica at 183 K (–89°C), and the coldest place (outside the lab) known in the universe is the Boomerang Nebula, with a temperature of 1 K.

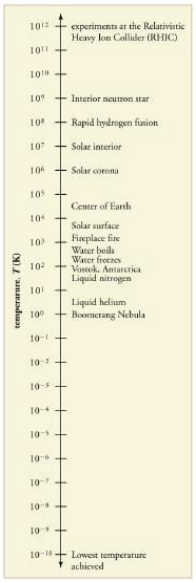


Figure 3.4.2.6: Each increment on this logarithmic scale indicates an increase by a factor of ten, and thus illustrates the tremendous range of temperatures in nature. Note that zero on a logarithmic scale would occur off the bottom of the page at infinity.

MAKING CONNECTIONS: ABSOLUTE ZERO

What is absolute zero? Absolute zero is the temperature at which all molecular motion has ceased. The concept of absolute zero arises from the behavior of gases. Figure 3.4.2.7 shows how the pressure of gases at a constant volume decreases as temperature decreases. Various scientists have noted that the pressures of gases extrapolate to zero at the same temperature, –273.15°C. This extrapolation implies that there is a lowest temperature. This temperature is called *absolute zero*. Today we know that most gases first liquefy and then freeze, and it is not actually possible to reach absolute zero. The numerical value of absolute zero temperature is –273.15°C or 0 K.

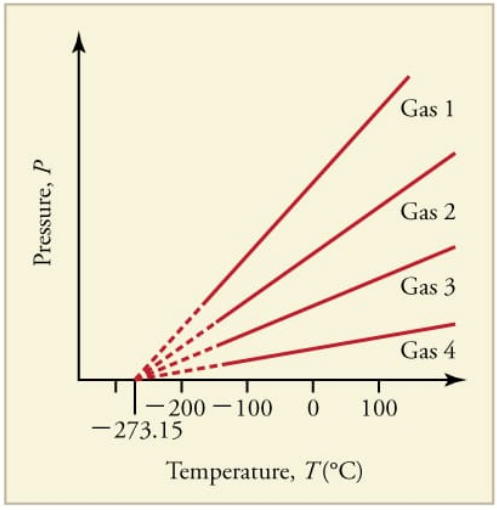


Figure 3.4.2.7: Graph of pressure versus temperature for various gases kept at a constant volume. Note that all of the graphs extrapolate to zero pressure at the same temperature.

Thermal Equilibrium and the Zeroth Law of Thermodynamics

Thermometers actually take their *own* temperature, not the temperature of the object they are measuring. This raises the question of how we can be certain that a thermometer measures the temperature of the object with which it is in contact. It is based on the fact that any two systems placed in *thermal contact* (meaning heat transfer can occur between them) will reach the same temperature. That is, heat will flow from the hotter object to the cooler one until they have exactly the same temperature. The objects are then in **thermal equilibrium**, and no further changes will occur. The systems interact and change because their temperatures differ, and the changes stop once their temperatures are the same. Thus, if enough time is allowed for this transfer of heat to run its course, the temperature a thermometer registers *does* represent the system with which it is in thermal equilibrium. Thermal equilibrium is established when two bodies are in contact with each other and can freely exchange energy.

Furthermore, experimentation has shown that if two systems, A and B, are in thermal equilibrium with each another, and B is in thermal equilibrium with a third system C, then A is also in thermal equilibrium with C. This conclusion may seem obvious, because all three have the same temperature, but it is basic to thermodynamics. It is called the **zeroth law of thermodynamics**.

THE ZEROTH LAW OF THERMODYNAMICS

If two systems, A and B, are in thermal equilibrium with each other, and B is in thermal equilibrium with a third system, C, then A is also in thermal equilibrium with C.

This law was postulated in the 1930s, after the first and second laws of thermodynamics had been developed and named. It is called the *zeroth law* because it comes logically before the first and second laws. An example of this law in action is seen in babies in incubators: babies in incubators normally have very few clothes on, so to an observer they look as if they may not be warm enough. However, the temperature of the air, the cot, and the baby is the same, because they are in thermal equilibrium, which is accomplished by maintaining air temperature to keep the baby comfortable.

Exercise 3.4.2.1

Does the temperature of a body depend on its size?

Answer available at the end of this section.

Answer available at the end of this section.

Section Summary

- Temperature is the quantity measured by a thermometer.
- Temperature is related to the average kinetic energy of atoms and molecules in a system.
- Absolute zero is the temperature at which there is no molecular motion.
- There are three main temperature scales: Celsius, Fahrenheit, and Kelvin.
- Temperatures on one scale can be converted to temperatures on another scale using the following equations:

$$T_{\circ}\mathrm{F} = \frac{9}{5}T_{\mathrm{C}}\mathrm{C} + 32$$
$$T_{\mathrm{o}}\mathrm{C} = \frac{5}{9}\left(T_{\mathrm{o}} - 32\right)$$
$$T_{\mathrm{K}} = T_{\mathrm{o}}\mathrm{C} + 273.15$$
$$T_{\mathrm{o}}\mathrm{C} = T_{\mathrm{K}} - 273.15$$

- Systems are in thermal equilibrium when they have the same temperature.
- Thermal equilibrium occurs when two bodies are in contact with each other and can freely exchange energy.
- The zeroth law of thermodynamics states that when two systems, A and B, are in thermal equilibrium with each other, and B is in thermal equilibrium with a third system, C, then A is also in thermal equilibrium with C.

Glossary

temperature
: the quantity measured by a thermometer

Celsius scale
: temperature scale in which the freezing point of water is 0°C and the boiling point of water is 100°C

degree Celsius
: unit on the Celsius temperature scale

Fahrenheit scale
: temperature scale in which the freezing point of water is 32°F and the boiling point of water is 212°F

degree Fahrenheit
: unit on the Fahrenheit temperature scale

Kelvin scale
: temperature scale in which 0 K is the lowest possible temperature, representing absolute zero

absolute zero
: the lowest possible temperature; the temperature at which all molecular motion ceases

thermal equilibrium
: the condition in which heat no longer flows between two objects that are in contact; the two objects have the same temperature

zeroth law of thermodynamics
: law that states that if two objects are in thermal equilibrium, and a third object is in thermal equilibrium with one of those objects, it is also in thermal equilibrium with the other object

Answers

Exercise 3.4.2.1

Exercise 3.4.2.1

No, the system can be divided into smaller parts each of which is at the same temperature. We say that the temperature is an *intensive* quantity. Intensive quantities are independent of size.

This page titled 3.4.2: Temperature is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.2: Temperature by Andrew Park is licensed CC BY 4.0.

3.4.3: The Ideal Gas Law

Learning Objectives

- State and explain the ideal gas law using Boltzmann's constant
- Use the ideal gas law to calculate pressure change, temperature change, volume change, or the number of molecules in a given volume.



Figure 3.4.3.1: The air inside this hot air balloon flying over Putrajaya, Malaysia, is hotter than the ambient air. As a result, the balloon experiences a buoyant force pushing it upward. (credit: Kevin Poh, Flickr)

In this section, we continue to explore the thermal behavior of gases. In particular, we examine the characteristics of atoms and molecules that compose gases. (Most gases, for example nitrogen, N_2, and oxygen, O_2, are composed of two or more atoms. We will primarily use the term “molecule” in discussing a gas because the term can also be applied to monatomic gases, such as helium.)

Gases are easily compressed. Gases expand and contract very rapidly with temperature changes. In addition, you will note that most gases expand at the *same* rate, or have the same β. This raises the question as to why gases should all act in nearly the same way, when liquids and solids have widely varying expansion rates.

The answer lies in the large separation of atoms and molecules in gases, compared to their sizes, as illustrated in Figure 3.4.3.2 . Because atoms and molecules have large separations, forces between them can be ignored, except when they collide with each other during collisions. The motion of atoms and molecules (at temperatures well above the boiling temperature) is fast, such that the gas occupies all of the accessible volume and the expansion of gases is rapid. In contrast, in liquids and solids, atoms and molecules are closer together and are quite sensitive to the forces between them.

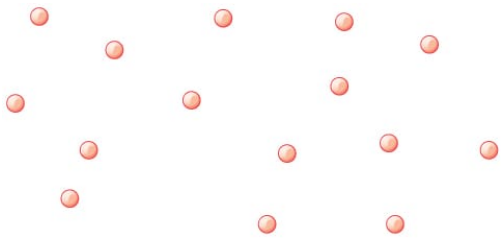


Figure 3.4.3.2: Atoms and molecules in a gas are typically widely separated, as shown. Because the forces between them are quite weak at these distances, the properties of a gas depend more on the number of atoms per unit volume and on temperature than on the type of atom.

To get some idea of how pressure, temperature, and volume of a gas are related to one another, consider what happens when you pump air into an initially deflated tire. The tire's volume first increases in direct proportion to the amount of air injected, without much increase in the tire pressure. Once the tire has expanded to nearly its full size, the walls limit volume expansion. If we continue to pump air into it, the pressure increases. The pressure will further increase when the car is driven and the tires move. Most manufacturers specify optimal tire pressure for cold tires. (See Figure 3.4.3.3.)

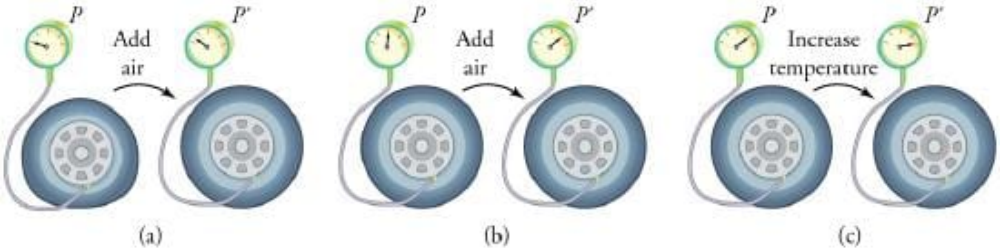


Figure 3.4.3.3: (a) When air is pumped into a deflated tire, its volume first increases without much increase in pressure. (b) When the tire is filled to a certain point, the tire walls resist further expansion and the pressure increases with more air. (c) Once the tire is inflated, its pressure increases with temperature.

At room temperatures, collisions between atoms and molecules can be ignored. In this case, the gas is called an ideal gas, in which case the relationship between the pressure, volume, and temperature is given by the equation of state called the ideal gas law.

IDEAL GAS LAW

The **ideal gas law** states that

$$PV = NkT,$$

where P is the absolute pressure of a gas, V is the volume it occupies, N is the number of atoms and molecules in the gas, and T is its absolute temperature. The constant k is called the **Boltzmann constant** in honor of Austrian physicist Ludwig Boltzmann (1844–1906) and has the value

$$k = 1.38 \times 10^{-23}\ \text{J/K}.$$

The ideal gas law can be derived from basic principles, but was originally deduced from experimental measurements of Charles' law (that volume occupied by a gas is proportional to temperature at a fixed pressure) and from Boyle's law (that for a fixed temperature, the product PV is a constant). In the ideal gas model, the volume occupied by its atoms and molecules is a negligible fraction of V. The ideal gas law describes the behavior of real gases under most conditions. (Note, for example, that N is the total number of atoms and molecules, independent of the type of gas.)

Let us see how the ideal gas law is consistent with the behavior of filling the tire when it is pumped slowly and the temperature is constant. At first, the pressure P is essentially equal to atmospheric pressure, and the volume V increases in direct proportion to the number of atoms and molecules N put into the tire. Once the volume of the tire is constant, the equation $PV = NkT$ predicts that the pressure should increase in proportion to *the number N of atoms and molecules*.

Example 3.4.3.1: Calculating Pressure Changes Due to Temperature Changes: Tire Pressure

Suppose your bicycle tire is fully inflated, with an absolute pressure of $7.00 \times 10^5\ \text{Pa}$ (a gauge pressure of just under $90.0\ \text{lb/in}^2$) at a temperature of 18.0°C. What is the pressure after its temperature has risen to 35.0°C? Assume that there are no appreciable leaks or changes in volume.

Strategy

The pressure in the tire is changing only because of changes in temperature. First we need to identify what we know and what we want to know, and then identify an equation to solve for the unknown.

We know the initial pressure $P_0 = 7.00 \times 10^5$ Pa, the initial temperature $T_0 = 18.0^\circ$C, and the final temperature $T_\mathrm{f} = 35.0^\circ$C. We must find the final pressure P_f. How can we use the equation $PV = NkT$? At first, it may seem that not enough information is given, because the volume V and number of atoms N are not specified. What we can do is use the equation twice: $P_0V_0 = NkT_0$ and $P_\mathrm{f}V_\mathrm{f} = NkT_\mathrm{f}$. If we divide $P_\mathrm{f}V_\mathrm{f}$ by P_0V_0 we can come up with an equation that allows us to solve for Pf.

$$\frac{P_\mathrm{f}V_\mathrm{f}}{P_0V_0} = \frac{N_\mathrm{f}kT_\mathrm{f}}{N_0kT_0}$$

Since the volume is constant, V_f and V_0 are the same and they cancel out. The same is true for N_f and N_0, and k, which is a constant. Therefore,

$$\frac{P_\mathrm{f}}{P_0} = \frac{T_\mathrm{f}}{T_0}.$$

We can then rearrange this to solve for P_f:

$$P_\mathrm{f} = P_0\frac{T_\mathrm{f}}{T_0},$$

where the temperature must be in units of kelvins, because T_0 and T_f are absolute temperatures.

Solution

1. Convert temperatures from Celsius to Kelvin.

$$T_0 = (18.0 + 273)\mathrm{K} = 291\ \mathrm{K}$$
$$T_\mathrm{f} = (35.0 + 273)\mathrm{K} = 308\ \mathrm{K}$$

2. Substitute the known values into the equation.

$$P_\mathrm{f} = P_0\frac{T_\mathrm{f}}{T_0} = 7.00 \times 10^5\ \mathrm{Pa}\left(\frac{308\ \mathrm{K}}{291\ \mathrm{K}}\right) = 7.41 \times 10^5\ \mathrm{Pa}$$

Discussion

The final temperature is about 6% greater than the original temperature, so the final pressure is about 6% greater as well. Note that *absolute* pressure and *absolute* temperature must be used in the ideal gas law.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT—REFRIGERATING A BALLOON

Inflate a balloon at room temperature. Leave the inflated balloon in the refrigerator overnight. What happens to the balloon, and why?

Example 3.4.3.2: Calculating the Number of Molecules in a Cubic Meter of Gas

How many molecules are in a typical object, such as gas in a tire or water in a drink? We can use the ideal gas law to give us an idea of how large N typically is.

Calculate the number of molecules in a cubic meter of gas at standard temperature and pressure (STP), which is defined to be 0°C and atmospheric pressure.

Strategy

Because pressure, volume, and temperature are all specified, we can use the ideal gas law $PV = NkT$, to find N.

Solution

1. Identify the knowns.

$$T = 0^{\circ}\text{C} = 273\ \text{K}$$
$$P = 1.01 \times 10^{5}\ \text{Pa}$$
$$V = 1.00\ \text{m}^3$$
$$k = 1.38 \times 10^{-23}\ \text{J/K}$$

2. Identify the unknown: number of molecules, N.

3. Rearrange the ideal gas law to solve for N.

$$PV = NkT$$
$$N = \frac{PV}{kT}$$

4. Substitute the known values into the equation and solve for N.

$$N = \frac{PV}{kT} = \frac{(1.01 \times 10^{5}\ \text{Pa})\,(1.00\ \text{m}^3)}{(1.38 \times 10^{-23}\ \text{J/K})\,(273\ \text{K})} = 2.68 \times 10^{25}\ \text{molecules}$$

Discussion

The calculated number, 2.68×10^{25}, is certainly very large. You might say that the volume of a cubic meter is also large ($1\ \text{m}^3 = 1000$), but even in a small volume of $1\ \text{cm}^3$, which is about size of a thimble ($1\ \text{cm}^3 = 10^{-6}\ \text{m}^3$), a gas at STP has 2.68×10^{19} molecules in it (still a very large number). Once again, note that N is the same for all types or mixtures of gases.

Section Summary

- The ideal gas law relates the pressure and volume of a gas to the number of gas molecules and the temperature of the gas.
- The ideal gas law can be written in terms of the number of molecules of gas:

  $$PV = NkT,$$

  where P is pressure, V is volume, T is temperature, N is number of molecules, and k is the Boltzmann constant

  $$k = 1.38 \times 10^{-23}\ \text{J/K}.$$

- The ideal gas law is generally valid at temperatures well above the boiling temperature.

Glossary

ideal gas law
: the physical law that relates the pressure and volume of a gas to the number of gas molecules or number of moles of gas and the temperature of the gas

Boltzmann constant
: k , a physical constant that relates energy to temperature; k=1.38×10^{-23} J/K

This page titled 3.4.3: The Ideal Gas Law is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.3: The Ideal Gas Law by Andrew Park is licensed CC BY 4.0.

3.4.4: Heat

Learning Objectives

- Define heat as transfer of energy.

We have defined work as force times distance and learned that work done on an object changes its kinetic energy. We have also seen that temperature is proportional to the (average) kinetic energy of atoms and molecules. We say that a thermal system has a certain internal energy: its internal energy is higher if the temperature is higher. If two objects at different temperatures are brought in contact with each other, energy is transferred from the hotter to the colder object until equilibrium is reached and the bodies reach thermal equilibrium (i.e., they are at the same temperature). No work is done by either object, because no force acts through a distance. The transfer of energy is caused by the temperature difference, and ceases once the temperatures are equal. These observations lead to the following definition of **heat**: Heat is the spontaneous transfer of energy due to a temperature difference.

Heat is often confused with temperature. For example, we may say the heat was unbearable, when we actually mean that the temperature was high. Heat is a form of energy, whereas temperature is not. The misconception arises because we are sensitive to the flow of heat, rather than the temperature.

Owing to the fact that heat is a form of energy, it has the SI unit of *joule* (J). The *calorie* (cal) is a common unit of energy, defined as the energy needed to change the temperature of 1.00 g of water by 1.00°C —specifically, between 14.5°C and 15.5°C, since there is a slight temperature dependence. Perhaps the most common unit of heat is the **kilocalorie** (kcal), which is the energy needed to change the temperature of 1.00 kg of water by 1.00°C. Since mass is most often specified in kilograms, kilocalorie is commonly used. Food calories (given the notation Cal, and sometimes called “big calorie”) are actually kilocalories (1kilocalorie=1000 calories), a fact not easily determined from package labeling.

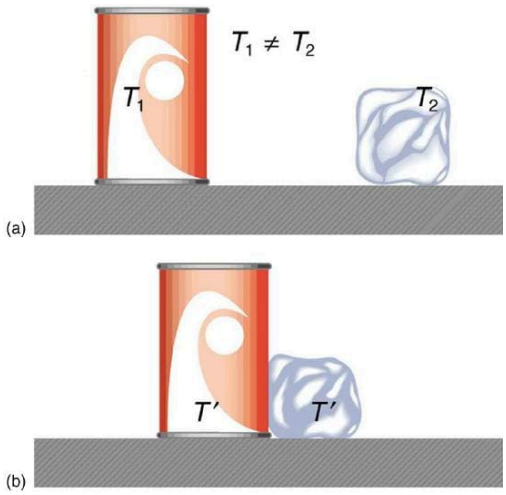


Figure 3.4.4.1: In figure (a) the soft drink and the ice have different temperatures, T_1 and T_2, and are not in thermal equilibrium. In figure (b), when the soft drink and ice are allowed to interact, energy is transferred until they reach the same temperature T', achieving equilibrium. Heat transfer occurs due to the difference in temperatures. In fact, since the soft drink and ice are both in contact with the surrounding air and bench, the equilibrium temperature will be the same for both.

Mechanical Equivalent of Heat

It is also possible to change the temperature of a substance by doing work. Work can transfer energy into or out of a system. This realization helped establish the fact that heat is a form of energy. James Prescott Joule (1818–1889) performed many experiments to establish the **mechanical equivalent of heat**—*the work needed to produce the same effects as heat transfer*. In terms of the units used for these two terms, the best modern value for this equivalence is

$$1.000 \text{ kcal} = 4186 \text{ J}.$$

We consider this equation as the conversion between two different units of energy.

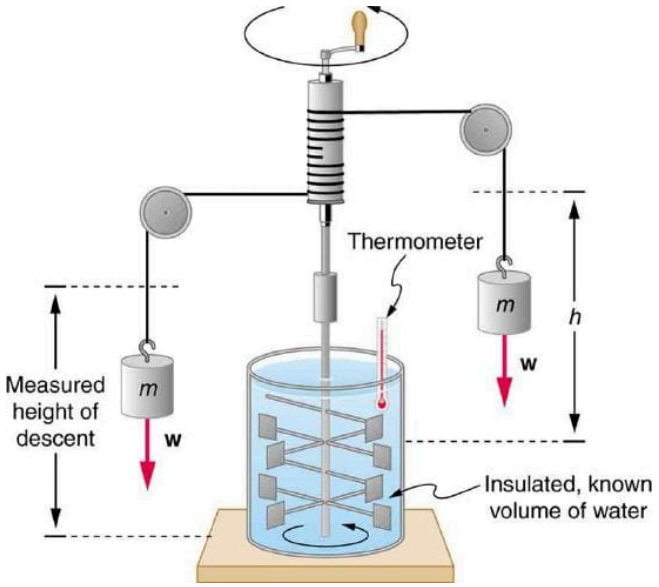


Figure 3.4.4.2: Schematic depiction of Joule's experiment that established the equivalence of heat and work.

The figure above shows one of Joule's most famous experimental setups for demonstrating the mechanical equivalent of heat. It demonstrated that work and heat can produce the same effects, and helped establish the principle of conservation of energy. Gravitational potential energy (PE) (work done by the gravitational force) is converted into kinetic energy (KE), and then randomized by viscosity and turbulence into increased average kinetic energy of atoms and molecules in the system, producing a temperature increase. His contributions to the field of thermodynamics were so significant that the SI unit of energy was named after him.

Heat added or removed from a system changes its internal energy and thus its temperature. Such a temperature increase is observed while cooking. However, adding heat does not necessarily increase the temperature. An example is melting of ice; that is, when a substance changes from one phase to another. Work done on the system or by the system can also change the internal energy of the system. Joule demonstrated that the temperature of a system can be increased by stirring. If an ice cube is rubbed against a rough surface, work is done by the frictional force. A system has a well-defined internal energy, but we cannot say that it has a certain "heat content" or "work content". We use the phrase "heat transfer" to emphasize its nature.

Exercise 3.4.4.1

Two samples (A and B) of the same substance are kept in a lab. Someone adds 10 kilojoules (kJ) of heat to one sample, while 10 kJ of work is done on the other sample. How can you tell to which sample the heat was added?

Answer available at the end of this section.

Summary

- Heat and work are the two distinct methods of energy transfer.
- Heat is energy transferred solely due to a temperature difference.
- Any energy unit can be used for heat transfer, and the most common are kilocalorie (kcal) and joule (J).
- Kilocalorie is defined to be the energy needed to change the temperature of 1.00 kg of water between 14.5°C and 15.5°C.
- The mechanical equivalent of this heat transfer is 1.00 kcal=4186 J.

Glossary

heat
: the spontaneous transfer of energy due to a temperature difference

kilocalorie
: $1 \text{ kilocalorie } = 1000 \text{ calories}$

mechanical equivalent of heat
: the work needed to produce the same effects as heat transfer

Answers

Exercise 3.4.4.1

Heat and work both change the internal energy of the substance. However, the properties of the sample only depend on the internal energy so that it is impossible to tell whether heat was added to sample A or B.

This page titled 3.4.4: Heat is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.4: Heat by Andrew Park is licensed CC BY 4.0.

3.4.5: Heat Transfer Methods

Learning Objectives

- Discuss the three different methods of heat transfer.

Equally as interesting as the effects of heat transfer on a system are the methods by which this occurs. Whenever there is a temperature difference, heat transfer occurs. Heat transfer may occur rapidly, such as through a cooking pan, or slowly, such as through the walls of a picnic ice chest. We can control rates of heat transfer by choosing materials (such as thick wool clothing for the winter), controlling air movement (such as the use of weather stripping around doors), or by choice of color (such as a white roof to reflect summer sunlight). So many processes involve heat transfer, so that it is hard to imagine a situation where no heat transfer occurs. Yet every process involving heat transfer takes place by only three methods:

1. **Conduction** is heat transfer through stationary matter by physical contact. (The matter is stationary on a macroscopic scale—we know there is thermal motion of the atoms and molecules at any temperature above absolute zero.) Heat transferred between the electric burner of a stove and the bottom of a pan is transferred by conduction.
2. **Convection** is the heat transfer by the macroscopic movement of a fluid. This type of transfer takes place in a forced-air furnace and in weather systems, for example.
3. Heat transfer by **radiation** occurs when microwaves, infrared radiation, visible light, or another form of electromagnetic radiation is emitted or absorbed. An obvious example is the warming of the Earth by the Sun. A less obvious example is thermal radiation from the human body.

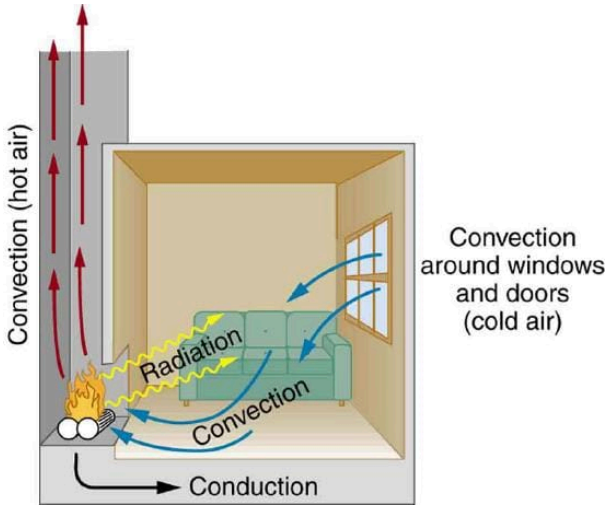


Figure 3.4.5.1: In a fireplace, heat transfer occurs by all three methods: conduction, convection, and radiation. Radiation is responsible for most of the heat transferred into the room. Heat transfer also occurs through conduction into the room, but at a much slower rate. Heat transfer by convection also occurs through cold air entering the room around windows and hot air leaving the room by rising up the chimney.

Figure 3.4.5.1 illustrates these three methods of heat transfer occurring in one system. While we skip detailed discussion of methods of heat transfer, the one thing all three methods-conduction, convection, and radiation-share is they are driven by temperature difference.

Exercise 3.4.5.1

Name an example from daily life (different from the text) for each mechanism of heat transfer.

Answer available at the end of this section.

Answer available at the end of this section.

Summary

- Heat is transferred by three different methods: conduction, convection, and radiation.

Glossary

conduction
: heat transfer through stationary matter by physical contact

convection
: heat transfer by the macroscopic movement of fluid

radiation
: heat transfer which occurs when microwaves, infrared radiation, visible light, or other electromagnetic radiation is emitted or absorbed

Answers

Exercise 3.4.5.1

Exercise 3.4.5.1

Conduction: Heat transfers into your hands as you hold a hot cup of coffee.
Convection: Heat transfers as the barista “steams” cold milk to make hot *cocoa*.
Radiation: Reheating a cold cup of coffee in a microwave oven.

This page titled 3.4.5: Heat Transfer Methods is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.5: Heat Transfer Methods by Andrew Park is licensed CC BY 4.0.

3.4.6: Temperature Change and Heat Capacity

Learning Objectives

- Observe heat transfer and change in temperature and mass.
- Calculate final temperature after heat transfer between two objects.

One of the major effects of heat transfer is temperature change: heating increases the temperature while cooling decreases it. We assume that there is no phase change and that no work is done on or by the system. Experiments show that the transferred heat depends on three factors—the change in temperature, the mass of the system, and the substance and phase of the substance.

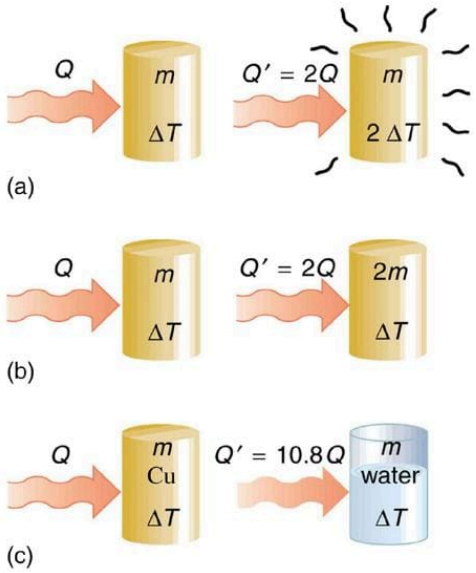


Figure 3.4.6.1: The heat Q transferred to cause a temperature change depends on the magnitude of the temperature change, the mass of the system, and the substance and phase involved. (a) The amount of heat transferred is directly proportional to the temperature change. To double the temperature change of a mass m, you need to add twice the heat. (b) The amount of heat transferred is also directly proportional to the mass. To cause an equivalent temperature change in a doubled mass, you need to add twice the heat. (c) The amount of heat transferred depends on the substance and its phase. If it takes an amount Q of heat to cause a temperature change ΔT in a given mass of copper, it will take 10.8 times that amount of heat to cause the equivalent temperature change in the same mass of water assuming no phase change in either substance.

The dependence on temperature change and mass are easily understood. Owing to the fact that the (average) kinetic energy of an atom or molecule is proportional to the absolute temperature, the internal energy of a system is proportional to the absolute temperature and the number of atoms or molecules. Owing to the fact that the transferred heat is equal to the change in the internal energy, the heat is proportional to the mass of the substance and the temperature change. The transferred heat also depends on the substance so that, for example, the heat necessary to raise the temperature is less for alcohol than for water. For the same substance, the transferred heat also depends on the phase (gas, liquid, or solid).

HEAT TRANSFER AND TEMPERATURE CHANGE

The quantitative relationship between heat transfer and temperature change contains all three factors:

$$Q = mc\Delta T,$$

where Q is the symbol for heat transfer, m is the mass of the substance, and ΔT is the change in temperature. The symbol c stands for **specific heat** and depends on the material and phase. The specific heat is the amount of heat necessary to change the temperature of 1.00 kg of mass by 1.00°C. The specific heat c is a property of the substance; its SI unit is $\mathrm{J/(kg \cdot K)}$ or $\mathrm{J/\,(kg \cdot {}^{\circ}C)}$. Recall that the temperature change (ΔT) is the same in units of kelvin and degrees Celsius. If heat transfer is measured in kilocalories, then *the unit of specific heat* is $\mathrm{kcal/\,(kg \cdot {}^{\circ}C)}$.

Values of specific heat must generally be looked up in tables, because there is no simple way to calculate them. In general, the specific heat also depends on the temperature. Table 3.4.6.1 lists representative values of specific heat for various substances. Except for gases, the temperature and volume dependence of the specific heat of most substances is weak. We see from this table that the specific heat of water is five times that of glass and ten times that of iron, which means that it takes five times as much heat to raise the temperature of water the same amount as for glass and ten times as much heat to raise the temperature of water as for iron. In fact, water has one of the largest specific heats of any material, which is important for sustaining life on Earth.

Example 3.4.6.1: Calculating the Required Heat: Heating Water in an Aluminum Pan

A 0.500 kg aluminum pan on a stove is used to heat 0.250 liters of water from $20.0^{\circ}\mathrm{C}$ to $80.0^{\circ}\mathrm{C}$. (a) How much heat is required? What percentage of the heat is used to raise the temperature of (b) the pan and (c) the water?

Strategy

The pan and the water are always at the same temperature. When you put the pan on the stove, the temperature of the water and the pan is increased by the same amount. We use the equation for the heat transfer for the given temperature change and mass of water and aluminum. The specific heat values for water and aluminum are given in Table 3.4.6.1.

Solution

Because water is in thermal contact with the aluminum, the pan and the water are at the same temperature.

1. Calculate the temperature difference:

$$\Delta T = T_{\mathrm{f}} - T_{\mathrm{i}} = 60.0^{\circ}\mathrm{C}.$$

2. Calculate the mass of water. Because the density of water is $1000\ \mathrm{kg/m^3}$, one liter of water has a mass of 1 kg, and the mass of 0.250 liters of water is $m_{\mathrm{w}} = 0.250\ \mathrm{kg}$.
3. Calculate the heat transferred to the water. Use the specific heat of water in Table 3.4.6.1:

$$Q_{\mathrm{w}} = m_{\mathrm{w}} c_{\mathrm{w}} \Delta T = (0.250\ \mathrm{kg})\,(4186\ \mathrm{J/kg^{\circ}C})\,(60.0^{\circ}\mathrm{C}) = 62.8\ \mathrm{kJ}.$$

4. Calculate the heat transferred to the aluminum. Use the specific heat for aluminum in Table 3.4.6.1:

$$Q_{\mathrm{Al}} = m_{\mathrm{Al}} c_{\mathrm{Al}} \Delta T = (0.500\ \mathrm{kg})\,(900\ \mathrm{J/kg^{\circ}C})\,(60.0^{\circ}\mathrm{C}) = 27.0 \times 10^4\ \mathrm{J} = 27.0\ \mathrm{kJ}.$$

5. Compare the percentage of heat going into the pan versus that going into the water. First, find the total transferred heat:

$$Q_{\mathrm{Total}} = Q_{\mathrm{W}} + Q_{\mathrm{Al}} = 62.8\ \mathrm{kJ} + 27.0\ \mathrm{kJ} = 89.8\ \mathrm{kJ}.$$

Thus, the amount of heat going into heating the pan is

$$\frac{27.0\ \mathrm{kJ}}{89.8\ \mathrm{kJ}} \times 100\% = 30.1\%,$$

and the amount going into heating the water is

$$\frac{62.8\ \mathrm{kJ}}{89.8\ \mathrm{kJ}} \times 100\% = 69.9\%.$$

Discussion

In this example, the heat transferred to the container is a significant fraction of the total transferred heat. Although the mass of the pan is twice that of the water, the specific heat of water is over four times greater than that of aluminum. Therefore, it takes a bit more than twice the heat to achieve the given temperature change for the water as compared to the aluminum pan.



Figure 3.4.6.2: The smoking brakes on this truck are a visible evidence of the mechanical equivalent of heat.

Example 3.4.6.2: Calculating the Temperature Increase from the Work Done on a Substance: Truck Brakes Overheat on Downhill Runs

Truck brakes used to control speed on a downhill run do work, converting gravitational potential energy into increased internal energy (higher temperature) of the brake material. This conversion prevents the gravitational potential energy from being converted into kinetic energy of the truck. The problem is that the mass of the truck is large compared with that of the brake material absorbing the energy, and the temperature increase may occur too fast for sufficient heat to transfer from the brakes to the environment.

Calculate the temperature increase of 100 kg of brake material with an average specific heat of $800\ \mathrm{J/kg \cdot {}^\circ C}$ if the material retains 10% of the energy from a 10,000-kg truck descending 75.0 m (in vertical displacement) at a constant speed.

Strategy

If the brakes are not applied, gravitational potential energy is converted into kinetic energy. When brakes are applied, gravitational potential energy is converted into internal energy of the brake material. We first calculate the gravitational potential energy (Mgh) that the entire truck loses in its descent and then find the temperature increase produced in the brake material alone.

Solution

1. Calculate the change in gravitational potential energy as the truck goes downhill

$$Mgh = (10,000\ \mathrm{kg})\left(9.80\ \mathrm{m/s^2}\right)(75.0\ \mathrm{m}) = 7.35 \times 10^6\ \mathrm{J}.$$

2. Calculate the temperature from the heat transferred using $Q = (0.10)Mgh$ (0.10 for 10% of energy being retained by brake material) and

$$\Delta T = \frac{Q}{mc},$$

where m is the mass of the brake material. Insert the values $m = 100\ \mathrm{kg}$ and $c = 800\ \mathrm{J/kg \cdot {}^\circ C}$ to find

$$\Delta T = \frac{\left(7.35 \times 10^5\ \mathrm{J}\right)}{(100\ \mathrm{kg})\,(800\ \mathrm{J/kg^\circ C})} = 9.2^\circ\mathrm{C}.$$

Discussion

The same idea underlies the recent hybrid technology of cars, where mechanical energy (gravitational potential energy) is converted by the brakes into electrical energy (battery).

Table 3.4.6.1: *Specific Heats*[1] *of Various Substances*

Substances	**Specific heat (*c*)**	
Solids	J/kg · °C	kcal/kg · °C[2]
Aluminum	900	0.215
Asbestos	800	0.19
Concrete, granite (average)	840	0.20
Copper	387	0.0924
Glass	840	0.20
Gold	129	0.0308
Human body (average at 37 °C)	3500	0.83
Ice (average, -50°C to 0°C)	2090	0.50
Iron, steel	452	0.108
Lead	128	0.0305
Silver	235	0.0562
Wood	1700	0.4
Liquids		
Benzene	1740	0.415
Ethanol	2450	0.586
Glycerin	2410	0.576
Mercury	139	0.0333
Water (15.0 °C)	4186	1.000
Gases [3]		
Air (dry)	721 (1015)	0.172 (0.242)
Ammonia	1670 (2190)	0.399 (0.523)
Carbon dioxide	638 (833)	0.152 (0.199)
Nitrogen	739 (1040)	0.177 (0.248)
Oxygen	651 (913)	0.156 (0.218)
Steam (100°C)	1520 (2020)	0.363 (0.482)

Note that Example 3.4.6.2 is an illustration of the mechanical equivalent of heat. Alternatively, the temperature increase could be produced by a blow torch instead of mechanically.

Example 3.4.6.3: Calculating the Final Temperature When Heat Is Transferred Between Two Bodies: Pouring Cold Water in a Hot Pan

Suppose you pour 0.250 kg of $20.0^{\circ}\mathrm{C}$ water (about a cup) into a 0.500-kg aluminum pan off the stove with a temperature of $150^{\circ}\mathrm{C}$. Assume that the pan is placed on an insulated pad and that a negligible amount of water boils off. What is the temperature when the water and pan reach thermal equilibrium a short time later?

Strategy

The pan is placed on an insulated pad so that little heat transfer occurs with the surroundings. Originally the pan and water are not in thermal equilibrium: the pan is at a higher temperature than the water. Heat transfer then restores thermal equilibrium once the water and pan are in contact. Because heat transfer between the pan and water takes place rapidly, the mass of evaporated water is negligible and the magnitude of the heat lost by the pan is equal to the heat gained by the water. The exchange of heat stops once a thermal equilibrium between the pan and the water is achieved. The heat exchange can be written as $|Q_{\mathrm{hot}}| = Q_{\mathrm{cold}}$.

Solution

1. Use the equation for heat transfer $Q = mc\Delta T$ to express the heat lost by the aluminum pan in terms of the mass of the pan, the specific heat of aluminum, the initial temperature of the pan, and the final temperature:

$$Q_{\mathrm{hot}} = m_{\mathrm{Al}} c_{\mathrm{Al}} \left(T_{\mathrm{f}} - 150^{\circ}\mathrm{C}\right).$$

2. Express the heat gained by the water in terms of the mass of the water, the specific heat of water, the initial temperature of the water and the final temperature:

$$Q_{\mathrm{cold}} = m_{\mathrm{W}} c_{\mathrm{W}} \left(T_{\mathrm{f}} - 20.0^{\circ}\mathrm{C}\right).$$

3. Note that $Q_{\mathrm{hot}} < 0$ and $Q_{\mathrm{cold}} > 0$ and that they must sum to zero because the heat lost by the hot pan must be the same as the heat gained by the cold water:

$$\begin{aligned} Q_{\mathrm{cold}} + Q_{\mathrm{hot}} &= 0, \\ Q_{\mathrm{cold}} &= -\mathrm{Q}_{\mathrm{hot}}, \\ m_{\mathrm{W}} c_{\mathrm{W}} \left(T_{\mathrm{f}} - 20.0^{\circ}\mathrm{C}\right) &= -\mathrm{m}_{\mathrm{Al}} c_{\mathrm{Al}} \left(T_{\mathrm{f}} - 150^{\circ}\mathrm{C}.\right) \end{aligned}$$

4. This an equation for the unknown final temperature, T_{f}
5. Bring all terms involving T_{f} on the left hand side and all other terms on the right hand side. Solve for T_{f},

$$T_{\mathrm{f}} = \frac{m_{\mathrm{Al}} c_{\mathrm{Al}} \left(150^{\circ}\mathrm{C}\right) + m_{\mathrm{W}} c_{\mathrm{W}} \left(20.0^{\circ}\mathrm{C}\right)}{m_{\mathrm{Al}} c_{\mathrm{Al}} + m_{\mathrm{W}} c_{\mathrm{W}}},$$

and insert the numerical values:

$$\begin{aligned} T_{\mathrm{f}} &= \frac{(0.500\ \mathrm{kg})\,(900\ \mathrm{J/kg^{\circ}C})\,(150^{\circ}\mathrm{C}) + (0.250\ \mathrm{kg})\,(4186\ \mathrm{J/kg^{\circ}C})\,(20.0^{\circ}\mathrm{C})}{(0.500\ \mathrm{kg})\,(900\ \mathrm{J/kg^{\circ}C}) + (0.250\ \mathrm{kg})\,(4186\ \mathrm{J/kg^{\circ}C})} \\ &= \frac{88430\ \mathrm{J}}{1496.5\ \mathrm{J/^{\circ}C}} \\ &= 59.1^{\circ}\mathrm{C}. \end{aligned}$$

Discussion

This is a typical *calorimetry* problem—two bodies at different temperatures are brought in contact with each other and exchange heat until a common temperature is reached. Why is the final temperature so much closer to 20.0°C than 150°C? The reason is that water has a greater specific heat than most common substances and thus undergoes a small temperature change for a given heat transfer. A large body of water, such as a lake, requires a large amount of heat to increase its temperature appreciably. This explains why the temperature of a lake stays relatively constant during a day even when the temperature change of the air is large. However, the water temperature does change over longer times (e.g., summer to winter).

TAKE-HOME EXPERIMENT: TEMPERATURE CHANGE OF LAND AND WATER

What heats faster, land or water?

To study differences in heat capacity:

- Place equal masses of dry sand (or soil) and water at the same temperature into two small jars. (The average density of soil or sand is about 1.6 times that of water, so you can achieve approximately equal masses by using 50% more water by volume.)
- Heat both (using an oven or a heat lamp) for the same amount of time.
- Record the final temperature of the two masses.
- Now bring both jars to the same temperature by heating for a longer period of time.
- Remove the jars from the heat source and measure their temperature every 5 minutes for about 30 minutes.

Which sample cools off the fastest? This activity replicates the phenomena responsible for land breezes and sea breezes.

Exercise 3.4.6.1

If 25 kJ is necessary to raise the temperature of a block from 25°C to 30°C, how much heat is necessary to heat the block from 45°C to 50°C?

Answer available at the end of this section.

Answer available at the end of this section.

Summary

- The transfer of heat Q that leads to a change ΔT in the temperature of a body with mass m is $Q = mc\Delta T$, where c is the specific heat of the material. This relationship can also be considered as the definition of specific heat.

Footnotes

- 1 The values for solids and liquids are at constant volume and at 25°C, except as noted.
- 2 These values are identical in units of cal/g · °C.
- 3 c_{v} at constant volume and at 20.0°C, except as noted, and at 1.00 atm average pressure. Values in parentheses are c_{p} at a constant pressure of 1.00 atm.

Glossary

specific heat
: the amount of heat necessary to change the temperature of 1.00 kg of a substance by 1.00 °C

Answers

Exercise 3.4.6.1

Exercise 3.4.6.1

The heat transfer depends only on the temperature difference. Since the temperature differences are the same in both cases, the same 25 kJ is necessary in the second case.

This page titled 3.4.6: Temperature Change and Heat Capacity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.6: Temperature Change and Heat Capacity by Andrew Park is licensed CC BY 4.0.

3.4.7: Phase Change and Latent Heat

Learning Objectives

- Describe phase changes.
- Explain the relationship between phase changes and heat transfer.

So far we have discussed temperature change due to heat transfer. No temperature change occurs from heat transfer if ice melts and becomes liquid water (i.e., during a phase change). For example, consider water dripping from icicles melting on a roof warmed by the Sun. Conversely, water freezes in an ice tray cooled by lower-temperature surroundings.



Figure 3.4.7.1: Heat from the air transfers to the ice causing it to melt. (credit: Mike Brand)

Energy is required to melt a solid because the cohesive bonds between the molecules in the solid must be broken apart such that, in the liquid, the molecules can move around at comparable kinetic energies; thus, there is no rise in temperature. Similarly, energy is needed to vaporize a liquid, because molecules in a liquid interact with each other via attractive forces. There is no temperature change until a phase change is complete. The temperature of a cup of soda initially at $0^{\circ}\mathrm{C}$ stays at $0^{\circ}\mathrm{C}$ until all the ice has melted. Conversely, energy is released during freezing and condensation, usually in the form of thermal energy. Work is done by cohesive forces when molecules are brought together. The corresponding energy must be given off (dissipated) to allow them to stay together Figure 3.4.7.2.

The energy involved in a phase change depends on two major factors: the number and strength of bonds or force pairs. The number of bonds is proportional to the number of molecules and thus to the mass of the sample. The strength of forces depends on the type of molecules. The heat Q required to change the phase of a sample of mass m is given by

$$Q = mL_{\mathrm{f}} \text{ (melting/freezing)},$$

$$Q = mL_{\mathrm{v}} \text{ (vaporization/condensation)},$$

where the latent heat of fusion, L_{f}, and latent heat of vaporization, L_{v}, are material constants that are determined experimentally. See (Table 3.4.7.1).

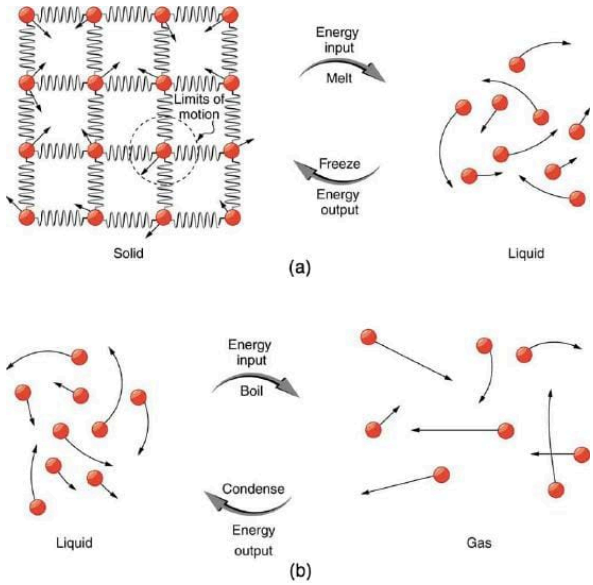


Figure 3.4.7.2: (a) Energy is required to partially overcome the attractive forces between molecules in a solid to form a liquid. That same energy must be removed for freezing to take place. (b) Molecules are separated by large distances when going from liquid to vapor, requiring significant energy to overcome molecular attraction. The same energy must be removed for condensation to take place. There is no temperature change until a phase change is complete.

Latent heat is measured in units of J/kg. Both L_{f} and L_{v} depend on the substance, particularly on the strength of its molecular forces as noted earlier. L_{f} and L_{v} are collectively called **latent heat coefficients**. They are *latent*, or hidden, because in phase changes, energy enters or leaves a system without causing a temperature change in the system; so, in effect, the energy is hidden. Table 3.4.7.1 lists representative values of L_{f} and L_{v}, together with melting and boiling points.

The table shows that significant amounts of energy are involved in phase changes. Let us look, for example, at how much energy is needed to melt a kilogram of ice at $0^{\circ}\mathrm{C}$ to produce a kilogram of water at $0^{\circ}\mathrm{C}$. Using the equation for a change in temperature and the value for water from Table 3.4.7.1, we find that $Q = mL_{\mathrm{f}} = (1.0\ \mathrm{kg})(334\ \mathrm{kJ/kg}) = 334\ \mathrm{kJ}$ is the energy to melt a kilogram of ice. This is a lot of energy as it represents the same amount of energy needed to raise the temperature of 1 kg of liquid water from $0^{\circ}\mathrm{C}$ to 79.8°C. Even more energy is required to vaporize water; it would take 2256 kJ to change 1 kg of liquid water at the normal boiling point ($100^{\circ}\mathrm{C}$ at atmospheric pressure) to steam (water vapor). This example shows that the energy for a phase change is enormous compared to energy associated with temperature changes without a phase change.

Table 3.4.7.1: *Heats of Fusion and Vaporization* [1]

Substance	Melting point (°C)	L_f		Boiling point (°C)	L_v	
		kJ/kg	kcal/kg		kJ/kg	kcal/kg
Helium	−269.7	5.23	1.25	−268.9	20.9	4.99
Hydrogen	−259.3	58.6	14.0	−252.9	452	108
Nitrogen	−210.0	25.5	6.09	−195.8	201	48.0
Oxygen	−218.8	13.8	3.30	−183.0	213	50.9
Ethanol	−114	104	24.9	78.3	854	204
Ammonia	−75		108	−33.4	1370	327
Mercury	−38.9	11.8	2.82	357	272	65.0
Water	0.00	334	79.8	100.0	2256[2]	539[3]
Sulfur	119	38.1	9.10	444.6	326	77.9
Lead	327	24.5	5.85	1750	871	208
Antimony	631	165	39.4	1440	561	134

Aluminum	660	380	90	2450	11400	2720
Silver	961	88.3	21.1	2193	2336	558
Gold	1063	64.5	15.4	2660	1578	377
Copper	1083	134	32.0	2595	5069	1211
Uranium	1133	84	20	3900	1900	454
Tungsten	3410	184	44	5900	4810	1150

Phase changes can have a tremendous stabilizing effect even on temperatures that are not near the melting and boiling points, because evaporation and condensation (conversion of a gas into a liquid state) occur even at temperatures below the boiling point. Take, for example, the fact that air temperatures in humid climates rarely go above $35.0^{\circ}\mathrm{C}$, which is because most heat transfer goes into evaporating water into the air. Similarly, temperatures in humid weather rarely fall below the dew point because enormous heat is released when water vapor condenses.

We examine the effects of phase change more precisely by considering adding heat into a sample of ice at $-20^{\circ}\mathrm{C}$ (Figure 3.4.7.3). The temperature of the ice rises linearly, absorbing heat at a constant rate of $0.50\ \mathrm{cal/g}\cdot{}^{\circ}\mathrm{C}$ until it reaches $0^{\circ}\mathrm{C}$. Once at this temperature, the ice begins to melt until all the ice has melted, absorbing 79.8 cal/g of heat. The temperature remains constant at $0^{\circ}\mathrm{C}$ during this phase change. Once all the ice has melted, the temperature of the liquid water rises, absorbing heat at a new constant rate of $1.00\ \mathrm{cal/g}\cdot{}^{\circ}\mathrm{C}$. At $100^{\circ}\mathrm{C}$, the water begins to boil and the temperature again remains constant while the water absorbs 539 cal/g of heat during this phase change. When all the liquid has become steam vapor, the temperature rises again, absorbing heat at a rate of $0.482\ \mathrm{cal/g}\cdot{}^{\circ}\mathrm{C}$.

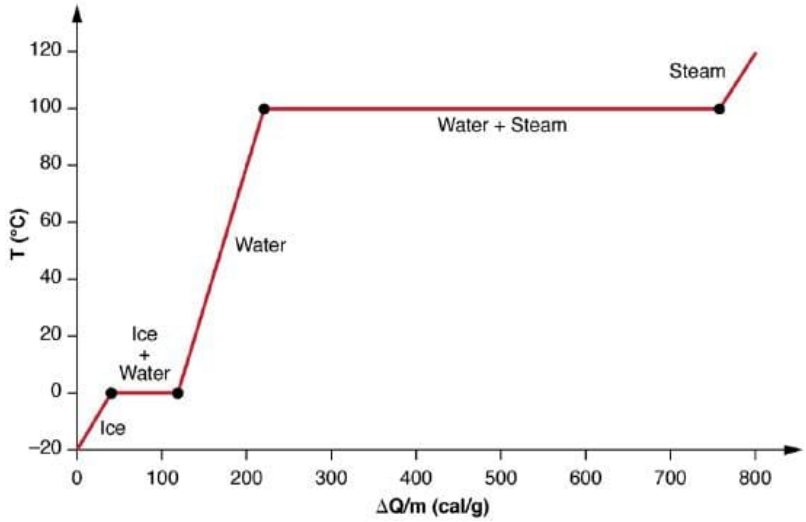


Figure 3.4.7.3: A graph of temperature versus energy added. The system is constructed so that no vapor evaporates while ice warms to become liquid water, and so that, when vaporization occurs, the vapor remains in of the system. The long stretches of constant temperature values at $0^{\circ}\mathrm{C}$ and $100^{\circ}\mathrm{C}$ reflect the large latent heat of melting and vaporization, respectively.

Water can evaporate at temperatures below the boiling point. More energy is required than at the boiling point, because the kinetic energy of water molecules at temperatures below $100^{\circ}\mathrm{C}$ is less than that at $100^{\circ}\mathrm{C}$, hence less energy is available from random thermal motions. Take, for example, the fact that, at body temperature, perspiration from the skin requires a heat input of 2428 kJ/kg, which is about 10 percent higher than the latent heat of vaporization at $100^{\circ}\mathrm{C}$. This heat comes from the skin, and thus provides an effective cooling mechanism in hot weather. High humidity inhibits evaporation, so that body temperature might rise, leaving unevaporated sweat on your brow.

Example 3.4.7.1: Calculate Final Temperature from Phase Change: Cooling Soda with Ice Cubes

Three ice cubes are used to chill a soda at $20^{\circ}\mathrm{C}$ with $m_{\mathrm{soda}} = 0.25\ \mathrm{kg}$. The ice is at $0^{\circ}\mathrm{C}$ and each ice cube has a mass of 6.0 g. Assume that the soda is kept in a foam container so that heat loss can be ignored. Assume the soda has the same heat capacity as water. Find the final temperature when all ice has melted.

Strategy

The ice cubes are at the melting temperature of $0^{\circ}\mathrm{C}$. Heat is transferred from the soda to the ice for melting. Melting of ice occurs in two steps: first the phase change occurs and solid (ice) transforms into liquid water at the melting temperature, then the temperature of this water rises. Melting yields water at $0^{\circ}\mathrm{C}$, so more heat is transferred from the soda to this water until the water plus soda system reaches thermal equilibrium,

$$Q_{\text{ice}} = -Q_{\text{soda}}.$$

The heat transferred to the ice is $Q_{\text{ice}} = m_{\text{ice}} L_{\text{f}} + m_{\text{ice}} c_{\text{W}} (T_{\text{f}} - 0^{\circ}\mathrm{C})$. The heat given off by the soda is $Q_{\text{soda}} = m_{\text{soda}} c_{\text{W}} (T_{\text{f}} - 20^{\circ}\mathrm{C})$. Since no heat is lost, $Q_{\text{ice}} = -Q_{\text{soda}}$, so that

$$m_{\text{ice}} L_{\text{f}} + m_{\text{ice}} c_{\text{W}} (T_{\text{f}} - 0^{\circ}\mathrm{C}) = -m_{\text{soda}} c_{\text{W}} (T_{\text{f}} - 20^{\circ}\mathrm{C}).$$

Bring all terms involving T_{f} on the left-hand-side and all other terms on the right-hand-side. Solve for the unknown quantity T_{f}:

$$T_{\text{f}} = \frac{m_{\text{soda}} c_{\text{W}} (20^{\circ}\mathrm{C}) - m_{\text{ice}} L_{\text{f}}}{(m_{\text{soda}} + m_{\text{ice}}) c_{\text{W}}}.$$

Solution

1. Identify the known quantities. The mass of ice is $m_{\text{ice}} = 3 \times 6.0\ \text{g} = 0.018\ \text{kg}$ and the mass of soda is $m_{\text{soda}} = 0.25\ \text{kg}$.
2. Calculate the terms in the numerator:

   $$m_{\text{soda}} c_{\text{W}} (20^{\circ}\mathrm{C}) = (0.25\ \text{kg})(4186\ \text{J/kg} \cdot {}^{\circ}\mathrm{C})(20^{\circ}\mathrm{C}) = 20,930\ \text{J}$$

   and

   $$m_{\text{ice}} L_{\text{f}} = (0.018\ \text{kg})(334,000\ \text{J/kg}) = 6012\ \text{J}.$$

3. Calculate the denominator:

   $$(m_{\text{soda}} + m_{\text{ice}}) c_{\text{W}} = (0.25\ \text{kg} + 0.018\ \text{kg})(4186\ \text{K}/(\text{kg} \cdot {}^{\circ}\mathrm{C}) = 1122\ \text{J}/{}^{\circ}\mathrm{C}.$$

4. Calculate the final temperature:

   $$T_{\text{f}} = \frac{20,930\ \text{J} - 6012\ \text{J}}{1122\ \text{J}/{}^{\circ}\mathrm{C}} = 13^{\circ}\mathrm{C}.$$

Discussion

This example illustrates the enormous energies involved during a phase change. The mass of ice is about 7 percent the mass of water but leads to a noticeable change in the temperature of soda. Although we assumed that the ice was at the freezing temperature, this is incorrect: the typical temperature is $-6^{\circ}\mathrm{C}$. However, this correction gives a final temperature that is essentially identical to the result we found. Can you explain why?

We have seen that vaporization requires heat transfer to a liquid from the surroundings, so that energy is released by the surroundings. Condensation is the reverse process, increasing the temperature of the surroundings. This increase may seem surprising, since we associate condensation with cold objects—the glass in the figure, for example. However, energy must be removed from the condensing molecules to make a vapor condense. The energy is exactly the same as that required to make the phase change in the other direction, from liquid to vapor, and so it can be calculated from $Q = mL_{\text{v}}$.

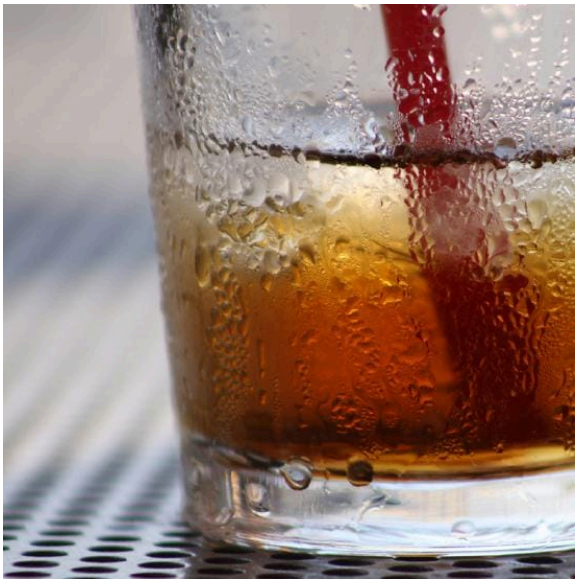


Figure 3.4.7.4: Condensation forms on this glass of iced tea because the temperature of the nearby air is reduced to below the dew point. The air cannot hold as much water as it did at room temperature, and so water condenses. Energy is released when the water condenses, speeding the melting of the ice in the glass. (credit: Jenny Downing)

REAL-WORLD APPLICATION

Energy is also released when a liquid freezes. This phenomenon is used by fruit growers in Florida to protect oranges when the temperature is close to the freezing point ($0^{\circ}\mathrm{C}$). Growers spray water on the plants in orchards so that the water freezes and heat is released to the growing oranges on the trees. This prevents the temperature inside the orange from dropping below freezing, which would damage the fruit.



Figure 3.4.7.5: The ice on these trees released large amounts of energy when it froze, helping to prevent the temperature of the trees from dropping below $0^{\circ}\mathrm{C}$. Water is intentionally sprayed on orchards to help prevent hard frosts. (credit: Hermann Hammer)

Sublimation is the transition from solid to vapor phase. You may have noticed that snow can disappear into thin air without a trace of liquid water, or the disappearance of ice cubes in a freezer. The reverse is also true: Frost can form on very cold windows without going through the liquid stage. A popular effect is the making of “smoke” from dry ice, which is solid carbon dioxide. Sublimation occurs because the equilibrium vapor pressure of solids is not zero. Certain air fresheners use the sublimation of a solid to inject a perfume into the room. Moth balls are a slightly toxic example of a phenol (an organic compound) that sublimates, while some solids, such as osmium tetroxide, are so toxic that they must be kept in sealed containers to prevent human exposure to their sublimation-produced vapors.

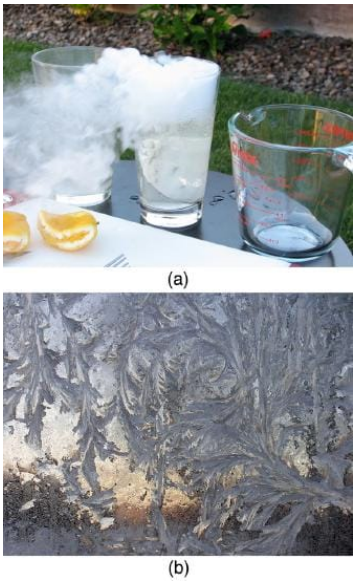


Figure 3.4.7.6: Direct transitions between solid and vapor are common, sometimes useful, and even beautiful. (a) Dry ice sublimates directly to carbon dioxide gas. The visible vapor is made of water droplets. (credit: Windell Oskay) (b) Frost forms patterns on a very cold window, an example of a solid formed directly from a vapor. (credit: Liz West)

All phase transitions involve heat. In the case of direct solid-vapor transitions, the energy required is given by the equation $Q = mL_{\mathrm{s}}$, where L_{s} is the **heat of sublimation**, which is the energy required to change 1.00 kg of a substance from the solid phase to the vapor phase. L_{s} is analogous to L_{f} and L_{v}, and its value depends on the substance. Sublimation requires energy input, so that dry ice is an effective coolant, whereas the reverse process (i.e., frosting) releases energy. The amount of energy required for sublimation is of the same order of magnitude as that for other phase transitions.

The material presented in this section and the preceding section allows us to calculate any number of effects related to temperature and phase change. In each case, it is necessary to identify which temperature and phase changes are taking place and then to apply the appropriate equation. Keep in mind that heat transfer and work can cause both temperature and phase changes.

Exercise 3.4.7.1

Why does snow remain on mountain slopes even when daytime temperatures are higher than the freezing temperature?

Answer available at the end of this section.

Answer available at the end of this section.

Summary

- Most substances can exist either in solid, liquid, and gas forms, which are referred to as “phases.”
- Phase changes occur at fixed temperatures for a given substance at a given pressure, and these temperatures are called boiling and freezing (or melting) points.
- During phase changes, heat absorbed or released is given by:

$$Q = mL,$$

where L is the latent heat coefficient.

Footnotes

- 1 Values quoted at the normal melting and boiling temperatures at standard atmospheric pressure (1 atm).

- 2 At $37.0^{\circ}\mathrm{C}$ (body temperature), the heat of vaporization L_{v} for water is 2430 kJ/kg or 580 kcal/kg
- 3 At $37.0^{\circ}\mathrm{C}$ (body temperature), the heat of vaporization L_{v} for water is 2430 kJ/kg or 580 kcal/kg

Glossary

heat of sublimation
: the energy required to change a substance from the solid phase to the vapor phase

latent heat coefficient
: a physical constant equal to the amount of heat transferred for every 1 kg of a substance during the change in phase of the substance

sublimation
: the transition from the solid phase to the vapor phase

Answers

Exercise 3.4.7.1

Exercise 3.4.7.1

Snow is formed from ice crystals and thus is the solid phase of water. Because enormous heat is necessary for phase changes, it takes a certain amount of time for this heat to be accumulated from the air, even if the air is above $0^{\circ}\mathrm{C}$. The warmer the air is, the faster this heat exchange occurs and the faster the snow melts.

Snow is formed from ice crystals and thus is the solid phase of water. Because enormous heat is necessary for phase changes, it takes a certain amount of time for this heat to be accumulated from the air, even if the air is above $0^{\circ}\mathrm{C}$. The warmer the air is, the faster this heat exchange occurs and the faster the snow melts.

This page titled 3.4.7: Phase Change and Latent Heat is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.7: Phase Change and Latent Heat by Andrew Park is licensed CC BY 4.0.

3.4.8: The First Law of Thermodynamics

Learning Objectives

- Define the first law of thermodynamics.
- Describe how conservation of energy relates to the first law of thermodynamics.
- Identify instances of the first law of thermodynamics working in everyday situations.



Figure 3.4.8.1: This boiling tea kettle represents energy in motion. The water in the kettle is turning to water vapor because heat is being transferred from the stove to the kettle. As the entire system gets hotter, work is done—from the evaporation of the water to the whistling of the kettle. (credit: Gina Hamilton)

If we are interested in how heat transfer is converted into doing work, then the conservation of energy principle is important. The first law of thermodynamics applies the conservation of energy principle to systems where heat transfer and doing work are the methods of transferring energy into and out of the system. The **first law of thermodynamics** states that the change in internal energy of a system equals the net heat transfer *into* the system minus the net work done *by* the system. In equation form, the first law of thermodynamics is

$$\Delta U = Q - W.$$

Here ΔU is the *change in internal energy* U of the system. Q is the *net heat transferred into the system*—that is, Q is the sum of all heat transfer into and out of the system. W is the *net work done by the system*—that is, W is the sum of all work done on or by the system. We use the following sign conventions: if Q is positive, then there is a net heat transfer into the system; if W is positive, then there is net work done by the system. So positive Q adds energy to the system and positive W takes energy from the system. Thus $\Delta U = Q - W$. Note also that if more heat transfer into the system occurs than work done, the difference is stored as internal energy. Heat engines are a good example of this—heat transfer into them takes place so that they can do work. (See Figure 3.4.8.2.) We will now examine Q, W, and ΔU further.

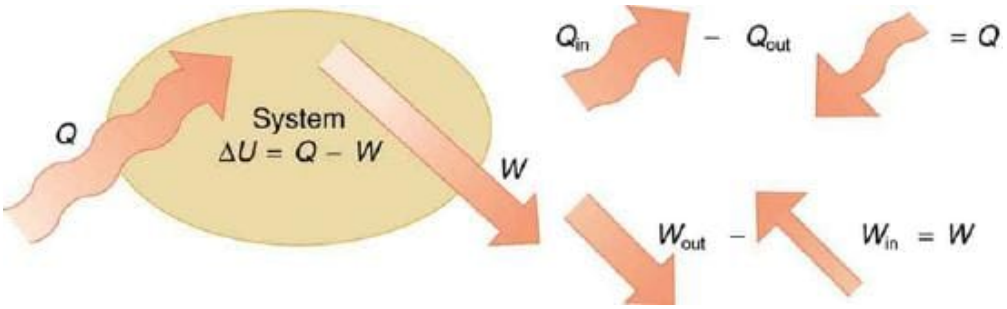


Figure 3.4.8.2: The first law of thermodynamics is the conservation-of-energy principle stated for a system where heat and work are the methods of transferring energy for a system in thermal equilibrium. Q represents the net heat transfer—it is the sum of all heat transfers into and out of the system. Q is positive for net heat transfer *into* the system. W is the total work done on and by the system. W is positive when more work is done *by* the system than on it. The change in the internal energy of the system, ΔU, is related to heat and work by the first law of thermodynamics, $\Delta U = Q - W$.

MAKING CONNECTIONS: LAW OF THERMODYNAMICS AND LAW OF CONSERVATION OF ENERGY

The first law of thermodynamics is actually the law of conservation of energy stated in a form most useful in thermodynamics. The first law gives the relationship between heat transfer, work done, and the change in internal energy of a system.

Heat Q and Work *W*

Heat transfer (Q) and doing work (W) are the two everyday means of bringing energy into or taking energy out of a system. The processes are quite different. Heat transfer, a less organized process, is driven by temperature differences. Work, a quite organized process, involves a macroscopic force exerted through a distance. Nevertheless, heat and work can produce identical results.For example, both can cause a temperature increase. Heat transfer into a system, such as when the Sun warms the air in a bicycle tire, can increase its temperature, and so can work done on the system, as when the bicyclist pumps air into the tire. Once the temperature increase has occurred, it is impossible to tell whether it was caused by heat transfer or by doing work. This uncertainty is an important point. Heat transfer and work are both energy in transit—neither is stored as such in a system. However, both can change the internal energy U of a system. Internal energy is a form of energy completely different from either heat or work.

Internal Energy *U*

We can think about the internal energy of a system in two different but consistent ways. The first is the atomic and molecular view, which examines the system on the atomic and molecular scale. The **internal energy** U of a system is the sum of the kinetic and potential energies of its atoms and molecules. Recall that kinetic plus potential energy is called mechanical energy. Thus internal energy is the sum of atomic and molecular mechanical energy. Because it is impossible to keep track of all individual atoms and molecules, we must deal with averages and distributions. A second way to view the internal energy of a system is in terms of its macroscopic characteristics, which are very similar to atomic and molecular average values.

Macroscopically, we define the change in internal energy ΔU to be that given by the first law of thermodynamics:

$$\Delta U = Q - W.$$

Many detailed experiments have verified that $\Delta U = Q - W$, where ΔU is the change in total kinetic and potential energy of all atoms and molecules in a system. It has also been determined experimentally that the internal energy U of a system depends only on the state of the system and *not how it reached that state*. More specifically, U is found to be a function of a few macroscopic quantities (pressure, volume, and temperature, for example), independent of past history such as whether there has been heat transfer or work done. This independence means that if we know the state of a system, we can calculate changes in its internal energy U from a few macroscopic variables.

MAKING CONNECTIONS: MACROSCOPIC AND MICROSCOPIC

In thermodynamics, we often use the macroscopic picture when making calculations of how a system behaves, while the atomic and molecular picture gives underlying explanations in terms of averages and distributions. We shall see this again in later sections of this chapter. For example, in the topic of entropy, calculations will be made using the atomic and molecular view.

Section Summary

- The first law of thermodynamics is given as $\Delta U = Q - W$, where ΔU is the change in internal energy of a system, Q is the net heat transfer (the sum of all heat transfer into and out of the system), and W is the net work done (the sum of all work done on or by the system).
- Both Q and W are energy in transit; only ΔU represents an independent quantity capable of being stored.
- The internal energy U of a system depends only on the state of the system and not how it reached that state.

Glossary

first law of thermodynamics
: states that the change in internal energy of a system equals the net heat transfer *into* the system minus the net work done *by* the system

internal energy
: the sum of the kinetic and potential energies of a system's atoms and molecules

human metabolism
: conversion of food into heat transfer, work, and stored fat

This page titled 3.4.8: The First Law of Thermodynamics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.8: The First Law of Thermodynamics by Andrew Park is licensed CC BY 4.0.

3.4.9: The First Law of Thermodynamics and Heat Engine Processes

Learning Objectives

- Describe the processes of a simple heat engine.
- Explain the differences among the simple thermodynamic processes—isobaric, isochoric, isothermal, and adiabatic.
- Explain the relationship between work done by a gas and change in its volume.

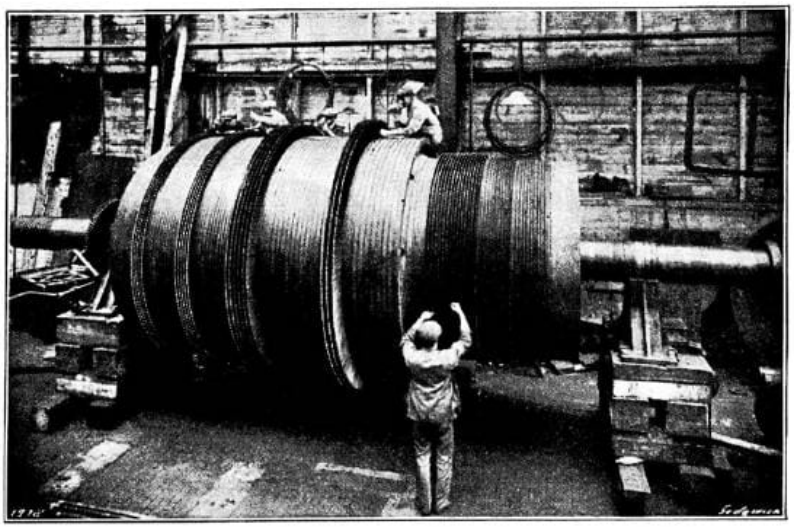


Figure 3.4.9.1: Beginning with the Industrial Revolution, humans have harnessed power through the use of the first law of thermodynamics, before we even understood it completely. This photo, of a steam engine at the Turbinia Works, dates from 1911, a mere 61 years after the first explicit statement of the first law of thermodynamics by Rudolph Clausius. (credit: public domain; author unknown)

One of the most important things we can do with heat transfer is to use it to do work for us. Such a device is called a **heat engine**. Car engines and steam turbines that generate electricity are examples of heat engines. Figure 3.4.9.2 shows schematically how the first law of thermodynamics applies to the typical heat engine.

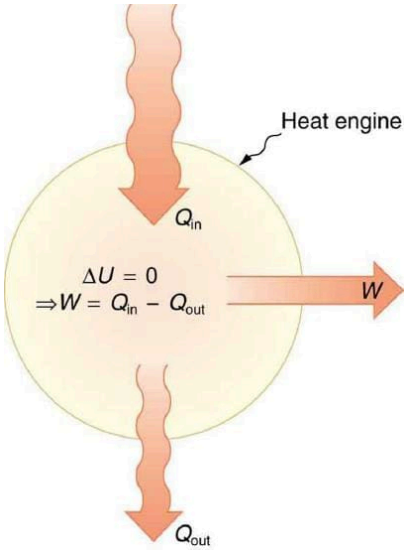


Figure 3.4.9.2: Schematic representation of a heat engine, governed, of course, by the first law of thermodynamics (and other laws of thermodynamics we will discuss later).

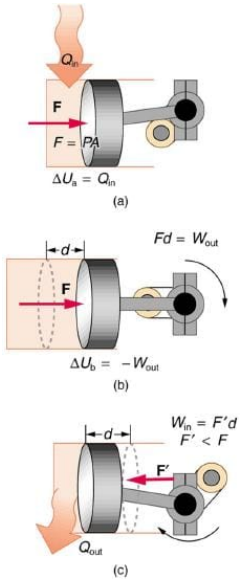


Figure 3.4.9.3: (a) Heat transfer to the gas in a cylinder increases the internal energy of the gas, creating higher pressure and temperature. (b) The force exerted on the movable cylinder does work as the gas expands. Gas pressure and temperature decrease when it expands, indicating that the gas's internal energy has been decreased by doing work. (c) Heat transfer to the environment further reduces pressure in the gas so that the piston can be more easily returned to its starting position.

The illustrations above show one of the ways in which heat transfer does work. Fuel combustion produces heat transfer to a gas in a cylinder, increasing the pressure of the gas and thereby the force it exerts on a movable piston. The gas does work on the outside world, as this force moves the piston through some distance. Heat transfer to the gas cylinder results in work being done. To repeat this process, the piston needs to be returned to its starting point. Heat transfer now occurs from the gas to the surroundings so that its pressure decreases, and a force is exerted by the surroundings to push the piston back through some distance. Variations of this process are employed daily in hundreds of millions of heat engines. Here, we consider some of the thermodynamic processes on which heat engines are based.

Work Done by a Gas

A process by which a gas does work on a piston at constant pressure is called an **isobaric process**. Since the pressure is constant, the force exerted is constant and the work done is given as

$$P\Delta V.$$

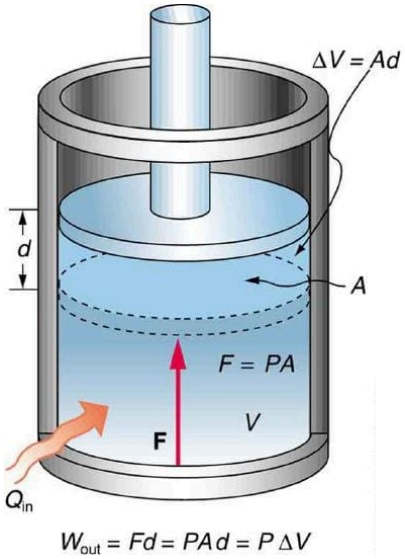


Figure 3.4.9.4: An isobaric expansion of a gas requires heat transfer to keep the pressure constant. Since pressure is constant, the work done is $P\Delta V$.

Recall from mechanics that work done by a force F on an object undergoing displacement d is

$$W = Fd.$$

See the symbols as shown in Figure 3.4.9.4. Now force is pressure times area ($F = PA$), and so

$$W = \text{ PAd } .$$

Because the volume of a cylinder is its cross-sectional area A times its length d, we see that $Ad = \Delta V$, the change in volume; thus,

$$W = P\Delta V \text{ (isobaric process) } .$$

Note that if ΔV is positive, then W is positive, meaning that positive work is done by the gas on the outside world.

(Note that the pressure involved in this work that we've called P is the pressure of the gas *inside* the tank. If we call the pressure outside the tank P_{ext} , an expanding gas would be working *against* the external pressure; the work done would therefore be $W = -P_{\text{ext}}\ \Delta V$ (isobaric process). There are some—especially chemists—who use this definition of work, and not the definition based on internal pressure, as the basis of the First Law of Thermodynamics. This definition reverses the sign conventions for work, and results in a statement of the first law that becomes $\Delta U = Q + W$. In this textbook, we will use the *physics* convention of using work done by the system on the surrounding, not the other way around.)

This is the key lesson from the above derivation: *a gas expanding under pressure does work on its surrounding, and unless additional energy is added through heat transfer, the internal energy of the gas decreases*. We will examine the experimental results that come about as a consequence of this fact later.

Thermodynamic Processes

We introduced the **isobaric process** above in discussing work done by a gas. Isobaric process is an example of a **thermodynamic process**. A thermodynamic process describes a change that happens to a gas, which results in change in its pressure (P), volume (V), and/or temperature (T). An isobaric process is a thermodynamic process that takes place under constant pressure (so the volume and temperature of the gas may change in an isobaric process).

There are three more named thermodynamic processes. These processes are given special names because, like the isobaric process, they occur under some restrictions, which gives them their special properties, as described briefly below. These three additional named thermodynamic processes are: isochoric, isothermal, and adiabatic processes.

An **isochoric process** is a thermodynamic process in which no change in volume takes place. Because the work done by a gas is

proportional to the change in volume, in an isochoric process, no work is done by (or on) the gas. Instead, in an isochoric process, a heat transfer takes place, and the energy from the heat transfer goes into increasing (or decreasing) the internal energy of the gas, increasing (or decreasing) its temperature.

An **isothermal process** is a thermodynamic process in which no change in temperature takes place. A gas expanding isothermally, for example, does work on the surrounding, but its internal energy (as represented by the temperature) does not change, because enough heat flows in to balance out the energy expended in doing work. This is consistent with the first law of thermodynamics ($0 = \Delta U = Q - W$, because $Q = W$). An isothermal process occurs if a thermodynamic process in a gas occurs slowly enough so that the gas remains in thermal equilibrium with the surrounding at all times.

The adiabatic process is, in some sense, the opposite of an isothermal process. In an **adiabatic process**, no heat transfer takes place (that is, $Q = 0$). This may happen because the gas is well-insulated from the surrounding. It may also happen because the process occurs so quickly that no significant heat transfer can take place. In an adiabatic expansion, for example, the internal energy of the gas decreases, because of the work done by the gas in expansion. This is perhaps the clearest experimental evidence one can observe that it takes work for a gas to expand under pressure.

Figure $3.4.9.5$ illustrates these three processes on a plot of pressure and volume (a PV diagram). In an isothermal process, as the gas expands, the pressure decreases. This can be predicted from the ideal gas law ($PV = NkT$). Since the temperature is constant, if volume increases, the pressure must decrease, to keep PV constant. You can also see that in an adiabatic process, the pressure decreases with expanding volume more steeply than an isothermal process, because in an adiabatic process, the temperature is not constant, but it decreases. So with increasing V, the pressure decrease even more rapidly, so that PV actually decreases (for decreasing temperature).

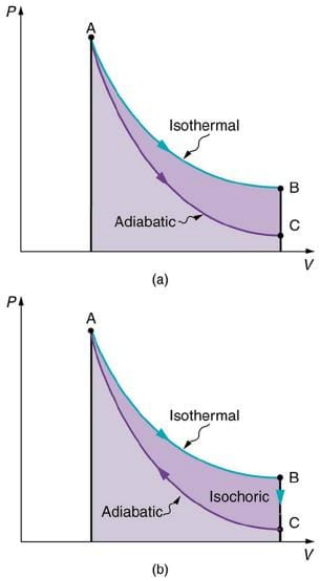


Figure $3.4.9.5$: (a) The upper curve is an isothermal process ($\Delta T = 0$), whereas the lower curve is an adiabatic process ($Q = 0$). Both start from the same point A, but the isothermal process does more work than the adiabatic because heat transfer into the gas takes place to keep its temperature constant. This keeps the pressure higher all along the isothermal path than along the adiabatic path, producing more work. The adiabatic path thus ends up with a lower pressure and temperature at point C, even though the final volume is the same as for the isothermal process. (b) The cycle ABCA produces a net work output.

Reversible Processes

Both isothermal and adiabatic processes such as shown in Figure $3.4.9.5$ are reversible in principle. A **reversible process** is one in which both the system and its environment can return to exactly the states they were in by following the reverse path. The reverse isothermal and adiabatic paths are BA and CA, respectively. Real macroscopic processes are never exactly reversible. In the previous examples, our system is a gas (like that in Figure $3.4.9.4$), and its environment is the piston, cylinder, and the rest of the universe. If there are any energy-dissipating mechanisms, such as friction or turbulence, then heat transfer to the environment occurs for either direction of the piston. So, for example, if the path BA is followed and there is friction, then the gas will be returned to its original state but the environment will not—it will have been heated in both directions. Reversibility requires the direction of heat transfer to reverse for the reverse path. Since dissipative mechanisms cannot be completely eliminated, real

processes cannot be reversible.

There must be reasons that real macroscopic processes cannot be reversible. We can imagine them going in reverse. For example, heat transfer occurs spontaneously from hot to cold and never spontaneously the reverse. Yet it would not violate the first law of thermodynamics for this to happen. In fact, all spontaneous processes, such as bubbles bursting, never go in reverse. There is a second thermodynamic law that forbids them from going in reverse. When we study this law, we will learn something about nature and also find that such a law limits the efficiency of heat engines. We will find that heat engines with the greatest possible theoretical efficiency would have to use reversible processes, and even they cannot convert all heat transfer into doing work. Table 3.4.9.1 summarizes the simpler thermodynamic processes and their definitions.

Table 3.4.9.1*: Summary of Simple Thermodynamic Processes*

Isobaric	Constant pressure $$W = P\Delta V$$
Isochoric	Constant volume $$W = 0$$
Isothermal	Constant temperature $$Q = W$$
Adiabatic	No heat transfer $$Q = 0$$

Section Summary

- One of the important implications of the first law of thermodynamics is that machines can be harnessed to do work that humans previously did by hand or by external energy supplies such as running water or the heat of the Sun. A machine that uses heat transfer to do work is known as a heat engine.
- There are several simple processes, used by heat engines, that flow from the first law of thermodynamics. Among them are the isobaric, isochoric, isothermal and adiabatic processes.
- These processes differ from one another based on how they affect pressure, volume, temperature, and heat transfer.
- If the work done is performed on the outside environment, work (W) will be a positive value. If the work done is done to the heat engine system, work (W) will be a negative value.
- Some thermodynamic processes, including isothermal and adiabatic processes, are reversible in theory; that is, both the thermodynamic system and the environment can be returned to their initial states. However, because of loss of energy owing to the second law of thermodynamics, complete reversibility does not work in practice.

Glossary

first law of thermodynamics
: states that the change in internal energy of a system equals the net heat transfer *into* the system minus the net work done *by* the system

internal energy
: the sum of the kinetic and potential energies of a system's atoms and molecules

human metabolism
: conversion of food into heat transfer, work, and stored fat

This page titled 3.4.9: The First Law of Thermodynamics and Heat Engine Processes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.9: The First Law of Thermodynamics and Heat Engine Processes by Andrew Park is licensed CC BY 4.0.

3.4.10: Introduction to the Second Law of Thermodynamics- Heat Engines and Their Efficiency

Learning Objectives

- State the expressions of the second law of thermodynamics.
- Calculate the efficiency of a coal-fired electricity plant, using second law characteristics.
- Describe and define the Otto cycle.



Figure 3.4.10.1: These ice floes melt during the Arctic summer. Some of them refreeze in the winter, but the second law of thermodynamics predicts that it would be extremely unlikely for the water molecules contained in these particular floes to reform the distinctive alligator-like shape they formed when the picture was taken in the summer of 2009. (credit: Patrick Kelley, U.S. Coast Guard, U.S. Geological Survey)

The second law of thermodynamics deals with the direction taken by spontaneous processes. Many processes occur spontaneously in one direction only—that is, they are irreversible, under a given set of conditions. Here are some examples of irreversible processes seen in day-to-day life: a glass breaking, heat transferring from a hot object to a cold object, conversion of kinetic energy into thermal energy, and a puff of gas expanding from a corner of a vacuum chamber into the entire chamber. If you saw a video of glass un-breaking, or two objects in thermal contact where one gets hotter and the other gets colder, or a hot stationary object spontaneously cooling off and moving in some direction, or a gas in a chamber "regrouping" into a corner, you would say this is a video run backwards. This is because you intuitively recognize these are examples of **irreversible process**, and these processes in nature only take place in the forward direction, not reverse (See Figure 3.4.10.2).

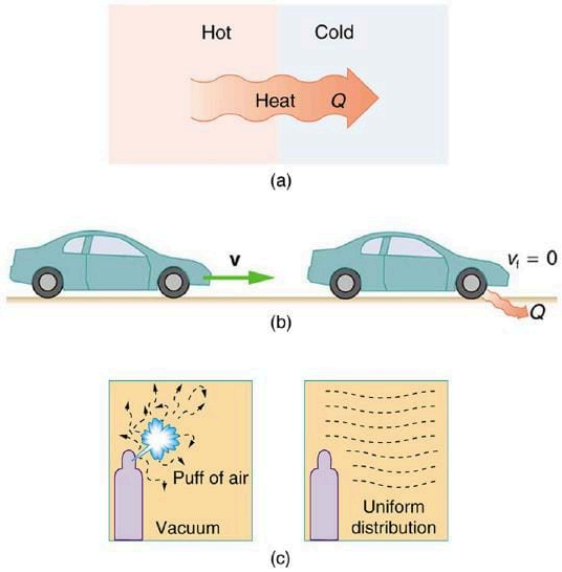


Figure 3.4.10.2: Examples of one-way processes in nature. (a) Heat transfer occurs spontaneously from hot to cold and not from cold to hot. (b) The brakes of this car convert its kinetic energy to heat transfer to the environment. The reverse process is impossible. (c) The burst of gas let into this vacuum chamber quickly expands to uniformly fill every part of the chamber. The random motions of the gas molecules will never return them to the corner.

The fact that certain processes never occur suggests that there is a law forbidding them to occur. The first law of thermodynamics would allow them to occur—none of those processes violate conservation of energy. The law that forbids these processes is called the second law of thermodynamics. We shall see that the second law can be stated in many ways that may seem different, but which in fact are equivalent. Like all natural laws, the second law of thermodynamics gives insights into nature, and its several statements imply that it is broadly applicable, fundamentally affecting many apparently disparate processes.

The already familiar direction of heat transfer from hot to cold is the basis of our first version of the **second law of thermodynamics**.

THE SECOND LAW OF THERMODYNAMICS (FIRST EXPRESSION)

Heat transfer occurs spontaneously from higher- to lower-temperature bodies but never spontaneously in the reverse direction.

Another way of stating this: It is impossible for any process to have as its sole result heat transfer from a cooler to a hotter object.

Heat Engines

Now let us consider a device that uses heat transfer to do work. As noted in the previous section, such a device is called a heat engine, and one is shown schematically in Figure 3.4.10.3(b). Gasoline and diesel engines, jet engines, and steam turbines are all heat engines that do work by using part of the heat transfer from some source. Heat transfer from the hot object (or hot reservoir) is denoted as Q_{h}, while heat transfer into the cold object (or cold reservoir) is Q_{c}, and the work done by the engine is W. The temperatures of the hot and cold reservoirs are T_{h} and T_{c}, respectively.

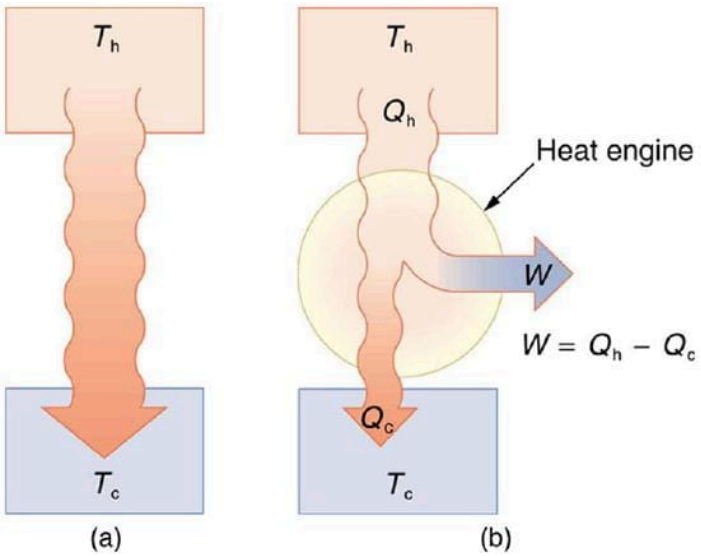


Figure 3.4.10.3: (a) Heat transfer occurs spontaneously from a hot object to a cold one, consistent with the second law of thermodynamics. (b) A heat engine, represented here by a circle, uses part of the heat transfer to do work. The hot and cold objects are called the hot and cold reservoirs. Q_{h} is the heat transfer out of the hot reservoir, W is the work output, and Q_{c} is the heat transfer into the cold reservoir.

Because the hot reservoir is heated externally, which is energy intensive, it is important that the work is done as efficiently as possible. In fact, we would like W to equal Q_{h}, and for there to be no heat transfer to the environment ($Q_{\mathrm{c}} = 0$). Unfortunately, this is impossible. The **second law of thermodynamics** also states, with regard to using heat transfer to do work (the second expression of the second law):

THE SECOND LAW OF THERMODYNAMICS (SECOND EXPRESSION)

It is impossible in any system for heat transfer from a reservoir to completely convert to work in a cyclical process in which the system returns to its initial state.

A **cyclical process** brings a system, such as the gas in a cylinder, back to its original state at the end of every cycle. Most heat engines, such as reciprocating piston engines and rotating turbines, use cyclical processes. The second law, just stated in its second form, clearly states that such engines cannot have perfect conversion of heat transfer into work done. Before going into the underlying reasons for the limits on converting heat transfer into work, we need to explore the relationships among W, Q_{h}, and Q_{c}, and to define the efficiency of a cyclical heat engine. As noted, a cyclical process brings the system back to its original condition at the end of every cycle. Such a system's internal energy U is the same at the beginning and end of every cycle—that is, $\Delta U = 0$. The first law of thermodynamics states that

$$\Delta U = Q - W,$$

where Q is the *net* heat transfer during the cycle ($Q = Q_{\mathrm{h}} - Q_{\mathrm{c}}$) and W is the net work done by the system. Since $\Delta U = 0$ for a complete cycle, we have

$$0 = Q - W,$$

so that

$$W = Q.$$

Thus the net work done by the system equals the net heat transfer into the system, or

$$W = Q_{\mathrm{h}} - Q_{\mathrm{c}} \text{ (cyclical process)},$$

just as shown schematically in Figure 3.4.10.3(b). The problem is that in all processes, there is some heat transfer Q_{c} to the environment—and usually a very significant amount at that.

In the conversion of energy to work, we are always faced with the problem of getting less out than we put in. We define *conversion efficiency* Eff to be the ratio of useful work output to the energy input (or, in other words, the ratio of what we get to what we spend). In that spirit, we define the efficiency of a heat engine to be its net work output W divided by heat transfer to the engine Q_{h}; that is,

$$Eff = \frac{W}{Q_{\text{h}}}.$$

Since $W = Q_{\text{h}} - Q_{\text{c}}$ in a cyclical process, we can also express this as

$$Eff = \frac{Q_{\text{h}} - Q_{\text{c}}}{Q_{\text{h}}} = 1 - \frac{Q_{\text{c}}}{Q_{\text{h}}} \text{ (cyclical process)},$$

making it clear that an efficiency of 1, or 100%, is possible only if there is no heat transfer to the environment ($Q_{\text{c}} = 0$). Note that all Qs are positive. The direction of heat transfer is indicated by a plus or minus sign. For example, Q_{c} is out of the system and so is preceded by a minus sign.

Example 3.4.10.1: Daily Work Done by a Coal-Fired Power Station, Its Efficiency and Carbon Dioxide Emissions

A coal-fired power station is a huge heat engine. It uses heat transfer from burning coal to do work to turn turbines, which are used to generate electricity. In a single day, a large coal power station has 2.50×10^{14} J of heat transfer from coal and 1.48×10^{14} J of heat transfer into the environment. (a) What is the work done by the power station? (b) What is the efficiency of the power station? (c) In the combustion process, the following chemical reaction occurs: $\text{C} + \text{O}_2 \rightarrow \text{CO}_2$. This implies that every 12 kg of coal puts 12 kg + 16 kg + 16 kg = 44 kg of carbon dioxide into the atmosphere. Assuming that 1 kg of coal can provide 2.5×10^6 J of heat transfer upon combustion, how much CO_2 is emitted per day by this power plant?

Strategy for (a)

We can use $W = Q_{\text{h}} - Q_{\text{c}}$ to find the work output W, assuming a cyclical process is used in the power station. In this process, water is boiled under pressure to form high-temperature steam, which is used to run steam turbine-generators, and then condensed back to water to start the cycle again.

Solution for (a)

Work output is given by:

$$W = Q_{\text{h}} - Q_{\text{c}}.$$

Substituting the given values:

$$\begin{aligned} W &= 2.50 \times 10^{14} \text{ J} - 1.48 \times 10^{14} \text{ J} \\ &= 1.02 \times 10^{14} \text{ J}. \end{aligned}$$

Strategy for (b)

The efficiency can be calculated with $Eff = \frac{W}{Q_{\text{h}}}$ since Q_{h} is given and work W was found in the first part of this example.

Solution for (b)

Efficiency is given by: $Eff = \frac{W}{Q_{\text{h}}}$ since Q_{h}. The work W was just found to be 1.02×10^{14} J, and Q_{h} is given, so the efficiency is

$$\begin{aligned} Eff &= \frac{1.02 \times 10^{14} \text{ J}}{2.50 \times 10^{14} \text{ J}} \\ &= 0.408, \text{ or } 40.8\% \end{aligned}$$

Strategy for (c)

The daily consumption of coal is calculated using the information that each day there is 2.50×10^{14} J of heat transfer from coal. In the combustion process, we have $C + O_2 \rightarrow CO_2$. So every 12 kg of coal puts 12 kg + 16 kg + 16 kg = 44 kg of CO_2 into the atmosphere.

Solution for (c)

The daily coal consumption is

$$\frac{2.50 \times 10^{14} \text{ J}}{2.50 \times 10^{6} \text{ J/kg}} = 1.0 \times 10^{8} \text{ kg}.$$

Assuming that the coal is pure and that all the coal goes toward producing carbon dioxide, the carbon dioxide produced per day is

$$1.0 \times 10^{8} \text{ kg coal} \times \frac{44 \text{ kg}CO_2}{12 \text{ kg coal}} = 3.7 \times 10^{8} \text{ kg}CO_2.$$

This is 370,000 metric tons of CO_2 produced every day.

Discussion

If all the work output is converted to electricity in a period of one day, the average power output is 1180 MW (this is left to you as an end-of-chapter problem). This value is about the size of a large-scale conventional power plant. The efficiency found is acceptably close to the value of 42% given for coal power stations. It means that fully 59.2% of the energy is heat transfer to the environment, which usually results in warming lakes, rivers, or the ocean near the power station, and is implicated in a warming planet generally. While the laws of thermodynamics limit the efficiency of such plants—including plants fired by nuclear fuel, oil, and natural gas—the heat transfer to the environment could be, and sometimes is, used for heating homes or for industrial processes. The generally low cost of energy has not made it economical to make better use of the waste heat transfer from most heat engines. Coal-fired power plants produce the greatest amount of CO_2 per unit energy output (compared to natural gas or oil), making coal the least efficient fossil fuel.

With the information given in Example 3.4.10.1, we can find characteristics such as the efficiency of a heat engine without any knowledge of how the heat engine operates, but looking further into the mechanism of the engine will give us greater insight. Figure 3.4.10.4 illustrates the operation of the common four-stroke gasoline engine. The four steps shown complete this heat engine's cycle, bringing the gasoline-air mixture back to its original condition.

The **Otto cycle** shown in Figure 3.4.10.5(a) is used in four-stroke internal combustion engines, although in fact the true Otto cycle paths do not correspond exactly to the strokes of the engine.

The adiabatic process AB corresponds to the nearly adiabatic compression stroke of the gasoline engine. In both cases, work is done on the system (the gas mixture in the cylinder), increasing its temperature and pressure. Along path BC of the Otto cycle, heat transfer Q_h into the gas occurs at constant volume, causing a further increase in pressure and temperature. This process corresponds to burning fuel in an internal combustion engine, and takes place so rapidly that the volume is nearly constant. Path CD in the Otto cycle is an adiabatic expansion that does work on the outside world, just as the power stroke of an internal combustion engine does in its nearly adiabatic expansion. The work done by the system along path CD is greater than the work done on the system along path AB, because the pressure is greater, and so there is a net work output. Along path DA in the Otto cycle, heat transfer Q_c from the gas at constant volume reduces its temperature and pressure, returning it to its original state. In an internal combustion engine, this process corresponds to the exhaust of hot gases and the intake of an air-gasoline mixture at a considerably lower temperature. In both cases, heat transfer into the environment occurs along this final path.

The net work done by a cyclical process is the area inside the closed path on a PV diagram, such as that inside path ABCDA in Figure 3.4.10.5. Note that in every imaginable cyclical process, it is absolutely necessary for heat transfer from the system to occur in order to get a net work output. In the Otto cycle, heat transfer occurs along path DA. If no heat transfer occurs, then the return path is the same, and the net work output is zero. The lower the temperature on the path AB, the less work has to be done to compress the gas. The area inside the closed path is then greater, and so the engine does more work and is thus more efficient. Similarly, the higher the temperature along path CD, the more work output there is. (See Figure 3.4.10.6.) So efficiency is related to the temperatures of the hot and cold reservoirs. In the next section, we shall see what the absolute limit to the efficiency of a heat engine is, and how it is related to temperature.

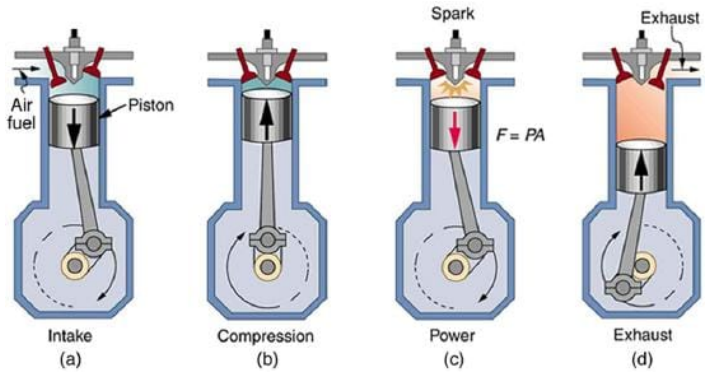


Figure 3.4.10.4: In the four-stroke internal combustion gasoline engine, heat transfer into work takes place in the cyclical process shown here. The piston is connected to a rotating crankshaft, which both takes work out of and does work on the gas in the cylinder. (a) Air is mixed with fuel during the intake stroke. (b) During the compression stroke, the air-fuel mixture is rapidly compressed in a nearly adiabatic process, as the piston rises with the valves closed. Work is done on the gas. (c) The power stroke has two distinct parts. First, the air-fuel mixture is ignited, converting chemical potential energy into thermal energy almost instantaneously, which leads to a great increase in pressure. Then the piston descends, and the gas does work by exerting a force through a distance in a nearly adiabatic process. (d) The exhaust stroke expels the hot gas to prepare the engine for another cycle, starting again with the intake stroke.

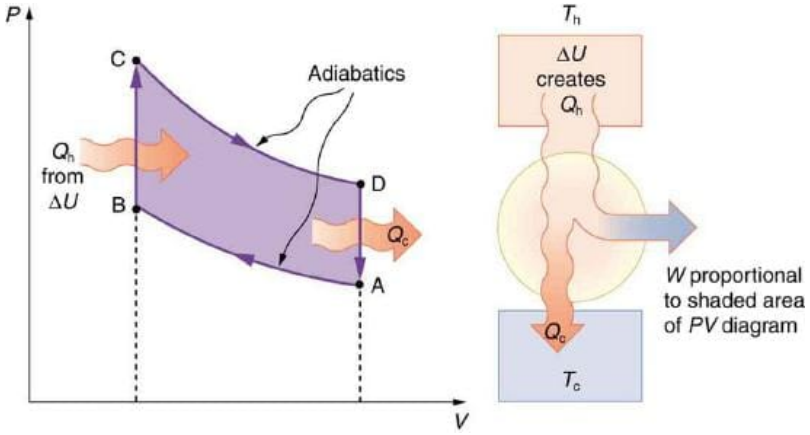


Figure 3.4.10.5: PV diagram for a simplified Otto cycle, analogous to that employed in an internal combustion engine. Point A corresponds to the start of the compression stroke of an internal combustion engine. Paths AB and CD are adiabatic and correspond to the compression and power strokes of an internal combustion engine, respectively. Paths BC and DA are isochoric and accomplish similar results to the ignition and exhaust-intake portions, respectively, of the internal combustion engine's cycle. Work is done on the gas along path AB, but more work is done by the gas along path CD, so that there is a net work output.

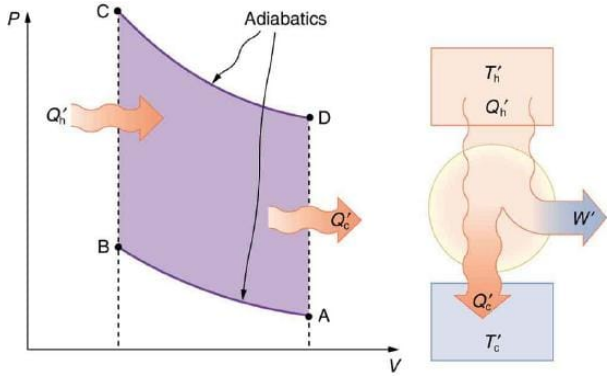


Figure 3.4.10.6: This Otto cycle produces a greater work output than the one in Figure 3.4.10.5, because the starting temperature of path CD is higher and the starting temperature of path AB is lower. The area inside the loop is greater, corresponding to greater net work output.

Section Summary

- The two expressions of the second law of thermodynamics are: (i) Heat transfer occurs spontaneously from higher- to lower-temperature bodies but never spontaneously in the reverse direction; and (ii) It is impossible in any system for heat transfer

from a reservoir to completely convert to work in a cyclical process in which the system returns to its initial state. These expressions describe the direction of irreversible processes that occur in nature.

- Cyclical processes are processes that return to their original state at the end of every cycle.
- In a cyclical process, such as a heat engine, the net work done by the system equals the net heat transfer into the system, or $W = Q_{\mathrm{h}} - Q_{\mathrm{c}}$, where Q_{h} is the heat transfer from the hot object (hot reservoir), and Q_c is the heat transfer into the cold object (cold reservoir).
- Efficiency can be expressed as $Eff = \frac{W}{Q_{\mathrm{h}}}$, the ratio of work output divided by the amount of energy input.
- The four-stroke gasoline engine is often explained in terms of the Otto cycle, which is a repeating sequence of processes that convert heat into work.

Glossary

irreversible process
: a process which occurs in only one direction in nature; a process that cannot be exactly reversed

second law of thermodynamics
: heat transfer flows from a hotter to a cooler object, never the reverse, and some heat energy in any process is lost to available work in a cyclical process

cyclical process
: a process in which the path returns to its original state at the end of every cycle

Otto cycle
: a thermodynamic cycle, consisting of a pair of adiabatic processes and a pair of isochoric processes, that converts heat into work, e.g., the four-stroke engine cycle of intake, compression, ignition, and exhaust

This page titled 3.4.10: Introduction to the Second Law of Thermodynamics- Heat Engines and Their Efficiency is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.10: Introduction to the Second Law of Thermodynamics- Heat Engines and Their Efficiency by Andrew Park is licensed CC BY 4.0.

3.4.11: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated

Learning Objectives

- Identify a Carnot cycle.
- Calculate maximum theoretical efficiency of a nuclear reactor.
- Explain how dissipative processes affect the ideal Carnot engine.



Figure 3.4.11.1: This novelty toy, known as the drinking bird, is an example of Carnot's engine. It contains methylene chloride (mixed with a dye) in the abdomen, which boils at a very low temperature—about 100°F. To operate, one gets the bird's head wet. As the water evaporates, fluid moves up into the head, causing the bird to become top-heavy and dip forward back into the water. This cools down the methylene chloride in the head, and it moves back into the abdomen, causing the bird to become bottom heavy and tip up. Except for a very small input of energy—the original head-wetting—the bird becomes a perpetual motion machine of sorts. (credit: Arabesk.nl, Wikimedia Commons)

We know from the second law of thermodynamics that a heat engine cannot be 100% efficient, since there must always be some heat transfer Q_c to the environment, which is often called waste heat. How efficient, then, can a heat engine be? This question was answered at a theoretical level in 1824 by a young French engineer, Sadi Carnot (1796–1832), in his study of the then-emerging heat engine technology crucial to the Industrial Revolution. He devised a theoretical cycle, now called the **Carnot cycle**, which is the most efficient cyclical process possible. The second law of thermodynamics can be restated in terms of the Carnot cycle, and so what Carnot actually discovered was this fundamental law. Any heat engine employing the Carnot cycle is called a **Carnot engine**.

What is crucial to the Carnot cycle—and, in fact, defines it—is that only reversible processes are used. Irreversible processes involve dissipative factors, such as friction and turbulence. This increases heat transfer Q_c to the environment and reduces the efficiency of the engine. Obviously, then, reversible processes are superior.

CARNOT ENGINE

Stated in terms of reversible processes, the **second law of thermodynamics** has a third form:

A Carnot engine operating between two given temperatures has the greatest possible efficiency of any heat engine operating between these two temperatures. Furthermore, all engines employing only reversible processes have this same maximum efficiency when operating between the same given temperatures.

Figure 3.4.11.2 shows the PV diagram for a Carnot cycle. The cycle comprises two isothermal and two adiabatic processes. Recall that both isothermal and adiabatic processes are, in principle, reversible.

Carnot also determined the efficiency of a perfect heat engine—that is, a Carnot engine. It is always true that the efficiency of a cyclical heat engine is given by:

$$Eff = \frac{Q_{\mathrm{h}} - Q_{\mathrm{c}}}{Q_{\mathrm{h}}} = 1 - \frac{Q_{\mathrm{c}}}{Q_{\mathrm{h}}}.$$

What Carnot found was that for a perfect heat engine, the ratio $Q_{\mathrm{c}}/Q_{\mathrm{h}}$ equals the ratio of the absolute temperatures of the heat reservoirs. That is, $Q_{\mathrm{c}}/Q_{\mathrm{h}} = T_{\mathrm{c}}/T_{\mathrm{h}}$ for a Carnot engine, so that the maximum or **Carnot efficiency** Eff_{C} is given by

$$Eff_{\mathrm{C}} = 1 - \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}},$$

where T_{h} and T_{c} are in kelvins (or any other absolute temperature scale). No real heat engine can do as well as the Carnot efficiency—an actual efficiency of about 0.7 of this maximum is usually the best that can be accomplished. But the ideal Carnot engine, like the drinking bird above, while a fascinating novelty, has zero power. This makes it unrealistic for any applications.

Carnot's interesting result implies that 100% efficiency would be possible only if $T_{\mathrm{c}} = 0\ \mathrm{K}$ —that is, only if the cold reservoir were at absolute zero, a practical and theoretical impossibility. But the physical implication is this—the only way to have all heat transfer go into doing work is to remove *all* thermal energy, and this requires a cold reservoir at absolute zero.

It is also apparent that the greatest efficiencies are obtained when the ratio $T_{\mathrm{c}}/T_{\mathrm{h}}$ is as small as possible. Just as discussed for the Otto cycle in the previous section, this means that efficiency is greatest for the highest possible temperature of the hot reservoir and lowest possible temperature of the cold reservoir. (This setup increases the area inside the closed loop on the PV diagram; also, it seems reasonable that the greater the temperature difference, the easier it is to divert the heat transfer to work.) The actual reservoir temperatures of a heat engine are usually related to the type of heat source and the temperature of the environment into which heat transfer occurs. Consider the following example.

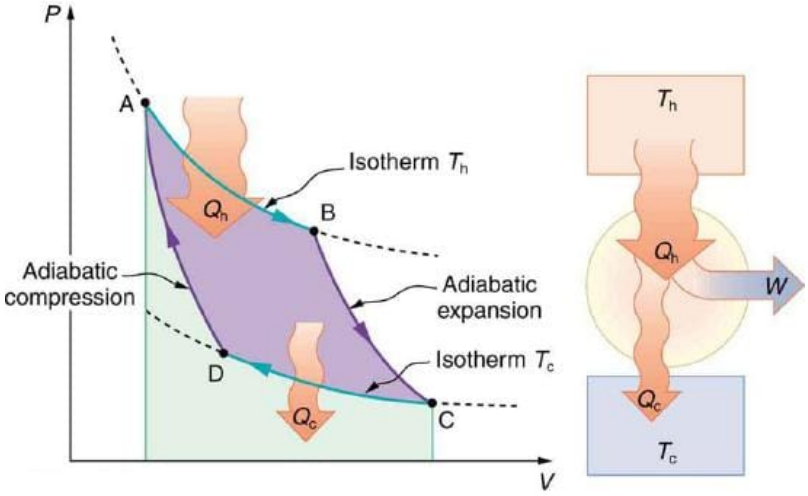


Figure 3.4.11.2: PV diagram for a Carnot cycle, employing only reversible isothermal and adiabatic processes. Heat transfer Q_{h} occurs into the working substance during the isothermal path AB, which takes place at constant temperature T_{h}. Heat transfer Q_{c} occurs out of the working substance during the isothermal path CD, which takes place at constant temperature T_{c}. The net work output W equals the area inside the path ABCDA. Also shown is a schematic of a Carnot engine operating between hot and cold reservoirs at temperatures T_{h} and T_{c}. Any heat engine using reversible processes and operating between these two temperatures will have the same maximum efficiency as the Carnot engine.

Example 3.4.11.1: Maximum Theoretical Efficiency for a Nuclear Reactor

A nuclear power reactor has pressurized water at $300^{\circ}\mathrm{C}$. (Higher temperatures are theoretically possible but practically not, due to limitations with materials used in the reactor.) Heat transfer from this water is a complex process (see Figure). Steam, produced in the steam generator, is used to drive the turbine-generators. Eventually the steam is condensed to water at $27^{\circ}\mathrm{C}$ and then heated again to start the cycle over. Calculate the maximum theoretical efficiency for a heat engine operating between these two temperatures.

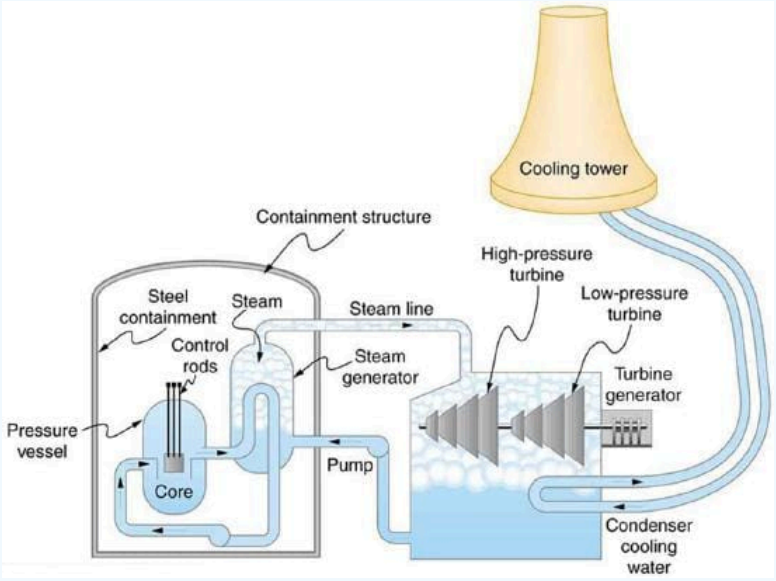


Figure 3.4.11.3: Schematic diagram of a pressurized water nuclear reactor and the steam turbines that convert work into electrical energy. Heat exchange is used to generate steam, in part to avoid contamination of the generators with radioactivity. Two turbines are used because this is less expensive than operating a single generator that produces the same amount of electrical energy. The steam is condensed to liquid before being returned to the heat exchanger, to keep exit steam pressure low and aid the flow of steam through the turbines (equivalent to using a lower-temperature cold reservoir). The considerable energy associated with condensation must be dissipated into the local environment; in this example, a cooling tower is used so there is no direct heat transfer to an aquatic environment. (Note that the water going to the cooling tower does not come into contact with the steam flowing over the turbines.)

Strategy

Since temperatures are given for the hot and cold reservoirs of this heat engine, $Eff_{\mathrm{C}} = 1 - \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}$ can be used to calculate the Carnot (maximum theoretical) efficiency. Those temperatures must first be converted to kelvins.

Solution

The hot and cold reservoir temperatures are given as $300^{\circ}\mathrm{C}$ and $27.0^{\circ}\mathrm{C}$, respectively. In kelvins, then, $T_{\mathrm{h}} = 573\ \mathrm{K}$ and $T_{\mathrm{c}} = 300\ \mathrm{K}$, so that the maximum efficiency is

$$Eff_{\mathrm{C}} = 1 - \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}.$$

Thus,

$$\begin{aligned} Eff_{\mathrm{C}} &= 1 - \frac{300\ \mathrm{K}}{573\ \mathrm{K}} \\ &= 0.476,\ \text{or}\ 47.6\%. \end{aligned}$$

Discussion

A typical nuclear power station's actual efficiency is about 35%, a little better than 0.7 times the maximum possible value, a tribute to superior engineering. Electrical power stations fired by coal, oil, and natural gas have greater actual efficiencies (about 42%), because their boilers can reach higher temperatures and pressures. The cold reservoir temperature in any of

these power stations is limited by the local environment. Figure 3.4.11.4 shows (a) the exterior of a nuclear power station and (b) the exterior of a coal-fired power station. Both have cooling towers into which water from the condenser enters the tower near the top and is sprayed downward, cooled by evaporation.

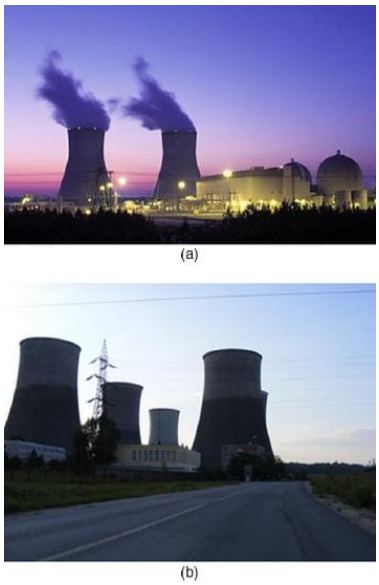


Figure 3.4.11.4: (a) A nuclear power station (credit: BlatantWorld.com) and (b) a coal-fired power station. Both have cooling towers in which water evaporates into the environment, representing Q_c. The nuclear reactor, which supplies Q_h, is housed inside the dome-shaped containment buildings. (credit: Robert & Mihaela Vicol, publicphoto.org)

Since all real processes are irreversible, the actual efficiency of a heat engine can never be as great as that of a Carnot engine, as illustrated in Figure 3.4.11.5(a). Even with the best heat engine possible, there are always dissipative processes in peripheral equipment, such as electrical transformers or car transmissions. These further reduce the overall efficiency by converting some of the engine's work output back into heat transfer, as shown in Figure 3.4.11.5(b).

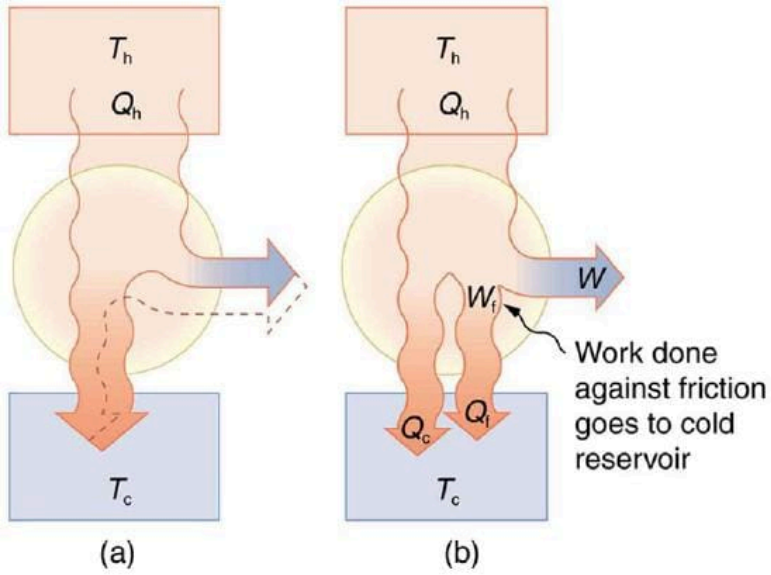


Figure 3.4.11.5: Real heat engines are less efficient than Carnot engines. (a) Real engines use irreversible processes, reducing the heat transfer to work. Solid lines represent the actual process; the dashed lines are what a Carnot engine would do between the same two reservoirs. (b) Friction and other dissipative processes in the output mechanisms of a heat engine convert some of its work output into heat transfer to the environment.

Section Summary

- The Carnot cycle is a theoretical cycle that is the most efficient cyclical process possible. Any engine using the Carnot cycle, which uses only reversible processes (adiabatic and isothermal), is known as a Carnot engine.
- Any engine that uses the Carnot cycle enjoys the maximum theoretical efficiency.
- While Carnot engines are ideal engines, in reality, no engine achieves Carnot's theoretical maximum efficiency, since dissipative processes, such as friction, play a role. Carnot cycles without heat loss may be possible at absolute zero, but this has never been seen in nature.

Glossary

Carnot cycle
: a cyclical process that uses only reversible processes, the adiabatic and isothermal processes

Carnot engine
: a heat engine that uses a Carnot cycle

Carnot efficiency
: the maximum theoretical efficiency for a heat engine

This page titled 3.4.11: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.11: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated by Andrew Park is licensed CC BY 4.0.

3.4.12: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy

Learning Objectives

- Define entropy and calculate the increase of entropy in a system with reversible and irreversible processes.
- Calculate the increasing disorder of a system.



Figure 3.4.12.1: The ice in this drink is slowly melting. Eventually the liquid will reach thermal equilibrium, as predicted by the second law of thermodynamics. (credit: Jon Sullivan, PDPhoto.org)

There is yet another way of expressing the second law of thermodynamics. This version relates to a concept called **entropy**. By examining it, we shall see that the directions associated with the second law—heat transfer from hot to cold, for example—are related to the tendency in nature for systems to become disordered and for less energy to be available for use as work. The entropy of a system can in fact be shown to be a measure of its disorder and of the unavailability of energy to do work.

MAKING CONNECTIONS: ENTROPY, ENERGY, AND WORK

Recall that the simple definition of energy is the ability to do work. Entropy is a measure of how much energy is not available to do work. Although all forms of energy are interconvertible, and all can be used to do work, it is not always possible, even in principle, to convert the entire available energy into work. That unavailable energy is of interest in thermodynamics, because the field of thermodynamics arose from efforts to convert heat to work.

We can see how entropy is defined by recalling our discussion of the Carnot engine. We noted that for a Carnot cycle, and hence for any reversible processes, $Q_{\mathrm{c}}/Q_{\mathrm{h}} = T_{\mathrm{c}}/T_{\mathrm{h}}$. Rearranging terms yields

$$\frac{Q_{\mathrm{c}}}{T_{\mathrm{c}}} = \frac{Q_{\mathrm{h}}}{T_{\mathrm{h}}}$$

for any reversible process. Q_{c} and Q_{h} are absolute values of the heat transfer at temperatures T_{c} and T_{h}, respectively. This ratio of Q/T is defined to be the **change in entropy** ΔS for a reversible process,

$$\Delta S = \left(\frac{Q}{T}\right)_{\mathrm{rev}},$$

where Q is the heat transfer, which is positive for heat transfer into and negative for heat transfer out of, and T is the absolute

temperature at which the reversible process takes place. The SI unit for entropy is joules per kelvin (J/K). If temperature changes during the process, then it is usually a good approximation (for small changes in temperature) to take T to be the average temperature, avoiding the need to use integral calculus to find ΔS.

The definition of ΔS is strictly valid only for reversible processes, such as used in a Carnot engine. However, we can find ΔS precisely even for real, irreversible processes. The reason is that the entropy S of a system, like internal energy U, depends only on the state of the system and not how it reached that condition. Entropy is a property of state. Thus the change in entropy ΔS of a system between state 1 and state 2 is the same no matter how the change occurs. We just need to find or imagine a reversible process that takes us from state 1 to state 2 and calculate ΔS for that process. That will be the change in entropy for any process going from state 1 to state 2. (See Figure 3.4.12.2.)

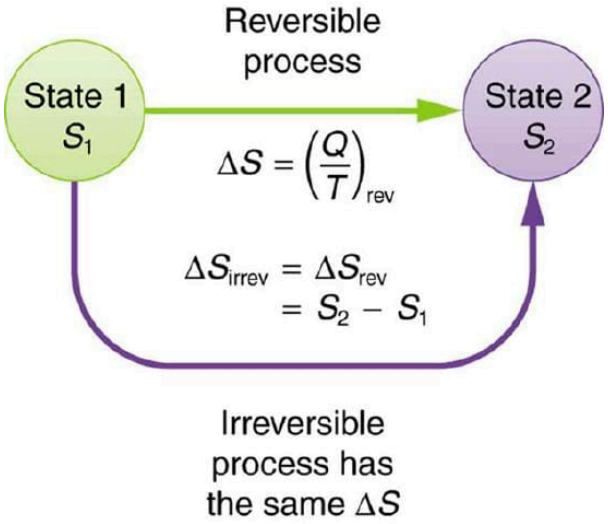


Figure 3.4.12.2: When a system goes from state 1 to state 2, its entropy changes by the same amount ΔS, whether a hypothetical reversible path is followed or a real irreversible path is taken.

Now let us take a look at the change in entropy of a Carnot engine and its heat reservoirs for one full cycle. The hot reservoir has a loss of entropy $\Delta S_{\mathrm{h}} = -Q_{\mathrm{h}}/T_{\mathrm{h}}$, because heat transfer occurs out of it (remember that when heat transfers out, then Q has a negative sign). The cold reservoir has a gain of entropy $\Delta S_{\mathrm{c}} = Q_{\mathrm{c}}/T_{\mathrm{c}}$, because heat transfer occurs into it. (We assume the reservoirs are sufficiently large that their temperatures are constant.) So the total change in entropy is

$$\Delta S_{\mathrm{tot}} = \Delta S_{\mathrm{h}} + \Delta S_{\mathrm{c}}$$

Thus, since we know that $Q_{\mathrm{h}}/T_{\mathrm{h}} = Q_{\mathrm{c}}/T_{\mathrm{c}}$ for a Carnot engine,

$$\Delta S_{\mathrm{tot}} = -\frac{Q_{\mathrm{h}}}{T_{\mathrm{h}}} + \frac{Q_{\mathrm{c}}}{T_{\mathrm{c}}} = 0.$$

This result, which has general validity, means that *the total change in entropy for a system in any reversible process is zero*.

The entropy of various parts of the system may change, but the total change is zero. Furthermore, the system does not affect the entropy of its surroundings, since heat transfer between them does not occur. Thus the reversible process changes neither the total entropy of the system nor the entropy of its surroundings. Sometimes this is stated as follows: *Reversible processes do not affect the total entropy of the universe*. Real processes are not reversible, though, and they do change total entropy. We can, however, use hypothetical reversible processes to determine the value of entropy in real, irreversible processes. The following example illustrates this point.

Example 3.4.12.1: Entropy Increases in an Irreversible (Real) Process

Spontaneous heat transfer from hot to cold is an irreversible process. Calculate the total change in entropy if 4000 J of heat transfer occurs from a hot reservoir at $T_{\mathrm{h}} = 600\ \mathrm{K}\ (327^{\circ}\mathrm{C})$ to a cold reservoir at $T_{\mathrm{c}} = 250\ \mathrm{K}\ (-23^{\circ}\mathrm{C})$, assuming there is no temperature change in either reservoir. (See Figure 3.4.12.3.)

Strategy

How can we calculate the change in entropy for an irreversible process when $\Delta S_{\text{tot}} = \Delta S_{\text{h}} + \Delta S_{\text{c}}$ is valid only for reversible processes? Remember that the total change in entropy of the hot and cold reservoirs will be the same whether a reversible or irreversible process is involved in heat transfer from hot to cold. So we can calculate the change in entropy of the hot reservoir for a hypothetical reversible process in which 4000 J of heat transfer occurs from it; then we do the same for a hypothetical reversible process in which 4000 J of heat transfer occurs to the cold reservoir. This produces the same changes in the hot and cold reservoirs that would occur if the heat transfer were allowed to occur irreversibly between them, and so it also produces the same changes in entropy.

Solution

We now calculate the two changes in entropy using $\Delta S_{\text{tot}} = \Delta S_{\text{h}} + \Delta S_{\text{c}}$. First, for the heat transfer from the hot reservoir,

$$\Delta S_{\text{h}} = \frac{-Q_{\text{h}}}{T_{\text{h}}} = \frac{-4000\ \text{J}}{600\ \text{K}} = -6.67\ \text{J/K}.$$

And for the cold reservoir,

$$\Delta S_{\text{c}} = \frac{Q_{\text{c}}}{T_{\text{c}}} = \frac{4000\ \text{J}}{250\ \text{K}} = 16.0\ \text{J/K}.$$

Thus the total is

$$\begin{aligned} \Delta S_{\text{tot}} &= \Delta S_{\text{h}} + \Delta S_{\text{c}} \\ &= (-6.67 + 16.0)\text{J/K} \\ &= 9.33\ \text{J/K}. \end{aligned}$$

Discussion

There is an *increase* in entropy for the system of two heat reservoirs undergoing this irreversible heat transfer. We will see that this means there is a loss of ability to do work with this transferred energy. Entropy has increased, and energy has become unavailable to do work.

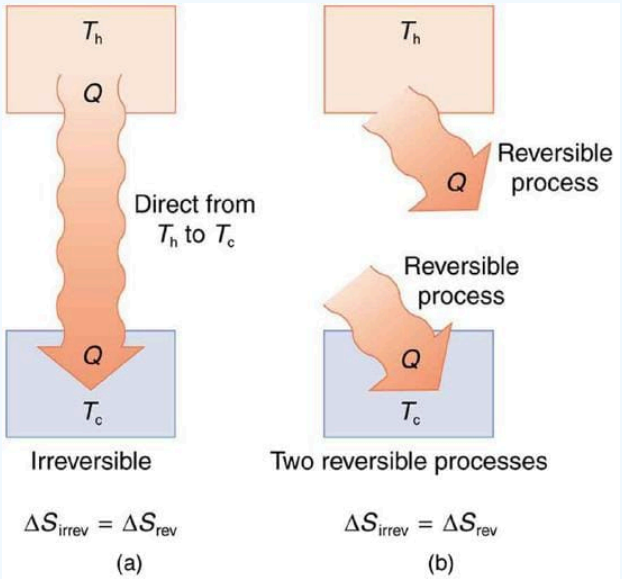


Figure 3.4.12.3: (a) Heat transfer from a hot object to a cold one is an irreversible process that produces an overall increase in entropy. (b) The same final state and, thus, the same change in entropy is achieved for the objects if reversible heat transfer processes occur between the two objects whose temperatures are the same as the temperatures of the corresponding objects in the irreversible process.

It is reasonable that entropy increases for heat transfer from hot to cold. Since the change in entropy is Q/T, there is a larger change at lower temperatures. The decrease in entropy of the hot object is therefore less than the increase in entropy of the cold object, producing an overall increase, just as in the previous example. This result is very general:

There is an increase in entropy for any system undergoing an irreversible process.

With respect to entropy, there are only two possibilities: entropy is constant for a reversible process, and it increases for an irreversible process. There is a fourth version of **the second law of thermodynamics stated in terms of entropy**:

The total entropy of a system either increases or remains constant in any process; it never decreases.

For example, heat transfer cannot occur spontaneously from cold to hot, because entropy would decrease.

Entropy is very different from energy. Entropy is *not* conserved but increases in all real processes. Reversible processes (such as in Carnot engines) are the processes in which the most heat transfer to work takes place and are also the ones that keep entropy constant. Thus we are led to make a connection between entropy and the availability of energy to do work.

Order to Disorder

Entropy is related not only to the unavailability of energy to do work—it is also a measure of disorder. This notion was initially postulated by Ludwig Boltzmann in the 1800s. For example, melting a block of ice means taking a highly structured and orderly system of water molecules and converting it into a disorderly liquid in which molecules have no fixed positions. (See Figure 3.4.12.4.) There is a large increase in entropy in the process, as seen in the following example.

Example 3.4.12.2: Entropy Associated with Disorder

Find the increase in entropy of 1.00 kg of ice originally at 0°C that is melted to form water at 0°C.

Strategy

As before, the change in entropy can be calculated from the definition of ΔS once we find the energy Q needed to melt the ice.

Solution

The change in entropy is defined as:

$$\Delta S = \frac{Q}{T}.$$

Here Q is the heat transfer necessary to melt 1.00 kg of ice and is given by

$$Q = mL_\text{f},$$

where m is the mass and L_f is the latent heat of fusion. $L_\text{f} = 334\ \text{kJ/kg}$ for water, so that

$$Q = (1.00\ \text{kg})(334\ \text{kJ/kg}) = 3.34 \times 10^5\ \text{J}.$$

Now the change in entropy is positive, since heat transfer occurs into the ice to cause the phase change; thus,

$$\Delta S = \frac{Q}{T} = \frac{3.34 \times 10^5\ \text{J}}{T}.$$

T is the melting temperature of ice. That is, $T = 0^\circ\text{C} = 273\ \text{K}$. So the change in entropy is

$$\begin{aligned}\Delta S &= \frac{3.34 \times 10^5\ \text{J}}{273\ \text{K}} \\ &= 1.22 \times 10^3\ \text{J/K}.\end{aligned}$$

Discussion

This is a significant increase in entropy accompanying an increase in disorder.

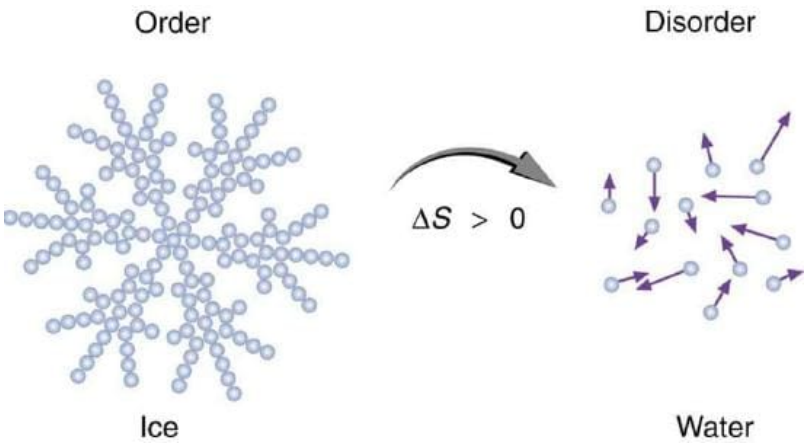


Figure 3.4.12.4: When ice melts, it becomes more disordered and less structured. The systematic arrangement of molecules in a crystal structure is replaced by a more random and less orderly movement of molecules without fixed locations or orientations. Its entropy increases because heat transfer occurs into it. Entropy is a measure of disorder.

In another easily imagined example, suppose we mix equal masses of water originally at two different temperatures, say $20.0^{\circ}\mathrm{C}$ and $40.0^{\circ}\mathrm{C}$. The result is water at an intermediate temperature of $30.0^{\circ}\mathrm{C}$. Three outcomes have resulted: entropy has increased, some energy has become unavailable to do work, and the system has become less orderly. Let us think about each of these results.

First, entropy has increased for the same reason that it did in the example above. Mixing the two bodies of water has the same effect as heat transfer from the hot one and the same heat transfer into the cold one. The mixing decreases the entropy of the hot water but increases the entropy of the cold water by a greater amount, producing an overall increase in entropy.

Second, once the two masses of water are mixed, there is only one temperature—you cannot run a heat engine with them. The energy that could have been used to run a heat engine is now unavailable to do work.

Third, the mixture is less orderly, or to use another term, less structured. Rather than having two masses at different temperatures and with different distributions of molecular speeds, we now have a single mass with a uniform temperature.

These three results—entropy, unavailability of energy, and disorder—are not only related but are in fact essentially equivalent.

Section Summary

- Entropy is the loss of energy available to do work.
- Another form of the second law of thermodynamics states that the total entropy of a system either increases or remains constant; it never decreases.
- Change of entropy is zero in a reversible process; it increases in an irreversible process.
- Entropy is also associated with the tendency toward disorder in an *isolated* system.

Glossary

entropy
: a measurement of a system's disorder and its inability to do work in a system

change in entropy
: the ratio of heat transfer to temperature Q/T

second law of thermodynamics stated in terms of entropy
: the total entropy of a system either increases or remains constant; it never decreases

This page titled 3.4.12: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.13: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy by Andrew Park is licensed CC BY 4.0.

3.4.13: Statistical Interpretation of Entropy and the Second Law of Thermodynamics- The Underlying Explanation

Learning Objectives

- Identify probabilities in entropy.
- Analyze statistical probabilities in entropic systems.



Figure 3.4.13.1: When you toss a coin a large number of times, heads and tails tend to come up in roughly equal numbers. Why doesn't heads come up 100, 90, or even 80% of the time? (credit: Jon Sullivan, PDPhoto.org)

The various ways of formulating the second law of thermodynamics tell what happens rather than why it happens. Why should heat transfer occur only from hot to cold? Why should energy become ever less available to do work? Why should the universe become increasingly disorderly? The answer is that it is a matter of overwhelming probability. Disorder is simply vastly more likely than order.

When you watch an emerging rain storm begin to wet the ground, you will notice that the drops fall in a disorganized manner both in time and in space. Some fall close together, some far apart, but they never fall in straight, orderly rows. It is not impossible for rain to fall in an orderly pattern, just highly unlikely, because there are many more disorderly ways than orderly ones. To illustrate this fact, we will examine some random processes, starting with coin tosses.

Coin Tosses

What are the possible outcomes of tossing 5 coins? Each coin can land either heads or tails. On the large scale, we are concerned only with the total heads and tails and not with the order in which heads and tails appear. The following possibilities exist:

Table 3.4.13.1*: Possible Outcomes of 5 Coin Tosses*

Combination of Heads and Tails	
5 heads	0 tails
4 heads	1 tails
3 heads	2 tails
2 heads	3 tails

1 heads	4 tails
0 heads	5 tails

These are what we call macrostates. A **macrostate** is an overall property of a system. It does not specify the details of the system, such as the order in which heads and tails occur or which coins are heads or tails.

Using this nomenclature, a system of 5 coins has the 6 possible macrostates just listed. Some macrostates are more likely to occur than others. For instance, there is only one way to get 5 heads, but there are several ways to get 3 heads and 2 tails, making the latter macrostate more probable. Table 3.4.13.2 lists of all the ways in which 5 coins can be tossed, taking into account the order in which heads and tails occur. Each sequence is called a **microstate**—a detailed description of every element of a system.

Table 3.4.13.2*: 5-Coin Toss*

Combinations	Individual microstates	Number of microstates
5 heads, 0 tails	HHHHH	1
4 heads, 1 tail	HHHHT, HHHTH, HHTHH, HTHHH, THHHH	5
3 heads, 2 tails	HTHTH, THTHH, HTHHT, THHTH, THHHT HTHTH, THTHH, HTHHT, THHTH, THHHT	10
2 heads, 3 tails	TTTHH, TTHHT, THHTT, HHTTT, TTHTH, THTHT, HTHTT, THTTH, HTTHT, HTTTH	10
1 head, 4 tails	TTTTH, TTTHT, TTHTT, THTTT, HTTTT	5
0 heads, 5 tails	TTTTT	1
		Total: 32

The macrostate of 3 heads and 2 tails can be achieved in 10 ways and is thus 10 times more probable than the one having 5 heads. Not surprisingly, it is equally probable to have the reverse, 2 heads and 3 tails. Similarly, it is equally probable to get 5 tails as it is to get 5 heads. Note that all of these conclusions are based on the crucial assumption that each microstate is equally probable. With coin tosses, this requires that the coins not be asymmetric in a way that favors one side over the other, as with loaded dice. With any system, the assumption that all microstates are equally probable must be valid, or the analysis will be erroneous.

The two most orderly possibilities are 5 heads or 5 tails. (They are more structured than the others.) They are also the least likely, only 2 out of 32 possibilities. The most disorderly possibilities are 3 heads and 2 tails and its reverse. (They are the least structured.) The most disorderly possibilities are also the most likely, with 20 out of 32 possibilities for the 3 heads and 2 tails and its reverse. If we start with an orderly array like 5 heads and toss the coins, it is very likely that we will get a less orderly array as a result, since 30 out of the 32 possibilities are less orderly. So even if you start with an orderly state, there is a strong tendency to go from order to disorder, from low entropy to high entropy. The reverse can happen, but it is unlikely.

Table 3.4.13.3*: 100-Coin Toss*

Macrostate		Number of microstates
Heads	Tails	(W)
100	0	1
99	1	1.0×10^{2}
95	5	7.5×10^{7}
90	10	1.7×10^{13}

Macrostate		Number of microstates
Heads	Tails	(W)
75	25	2.4×10^{23}
60	40	1.4×10^{28}
55	45	6.1×10^{28}
51	49	9.9×10^{28}
50	50	1.0×10^{29}
49	51	9.9×10^{28}
45	55	6.1×10^{28}
40	60	1.4×10^{28}
25	75	2.4×10^{23}
10	90	1.7×10^{13}
5	95	7.5×10^{7}
1	99	1.0×10^{2}
0	100	1
		Total: 1.27×10^{30}

This result becomes dramatic for larger systems. Consider what happens if you have 100 coins instead of just 5. The most orderly arrangements (most structured) are 100 heads or 100 tails. The least orderly (least structured) is that of 50 heads and 50 tails. There is only 1 way (1 microstate) to get the most orderly arrangement of 100 heads. There are 100 ways (100 microstates) to get the next most orderly arrangement of 99 heads and 1 tail (also 100 to get its reverse). And there are 1.0×10^{29} ways to get 50 heads and 50 tails, the least orderly arrangement. Table 3.4.13.3 is an abbreviated list of the various macrostates and the number of microstates for each macrostate. The total number of microstates—the total number of different ways 100 coins can be tossed—is an impressively large 1.27×10^{30}. Now, if we start with an orderly macrostate like 100 heads and toss the coins, there

is a virtual certainty that we will get a less orderly macrostate. If we keep tossing the coins, it is possible, but exceedingly unlikely, that we will ever get back to the most orderly macrostate. If you tossed the coins once each second, you could expect to get either 100 heads or 100 tails once in 2×10^{22} years! This period is 1 trillion (10^{12}) times longer than the age of the universe, and so the chances are essentially zero. In contrast, there is an 8% chance of getting 50 heads, a 73% chance of getting from 45 to 55 heads, and a 96% chance of getting from 40 to 60 heads. Disorder is highly likely.

Disorder in a Gas

The fantastic growth in the odds favoring disorder that we see in going from 5 to 100 coins continues as the number of entities in the system increases. Let us now imagine applying this approach to perhaps a small sample of gas. Because counting microstates and macrostates involves statistics, this is called **statistical analysis**. The macrostates of a gas correspond to its macroscopic properties, such as volume, temperature, and pressure; and its microstates correspond to the detailed description of the positions and velocities of its atoms. Even a small amount of gas has a huge number of atoms: 1.0 cm^3 of an ideal gas at 1.0 atm and 0° C has 2.7×10^{19} atoms. So each macrostate has an immense number of microstates. In plain language, this means that there are an immense number of ways in which the atoms in a gas can be arranged, while still having the same pressure, temperature, and so on.

The most likely conditions (or macrostates) for a gas are those we see all the time—a random distribution of atoms in space with a Maxwell-Boltzmann distribution of speeds in random directions, as predicted by kinetic theory. This is the most disorderly and least structured condition we can imagine. In contrast, one type of very orderly and structured macrostate has all of the atoms in one corner of a container with identical velocities. There are very few ways to accomplish this (very few microstates corresponding to it), and so it is exceedingly unlikely ever to occur. (See Figure 3.4.13.2(b).) Indeed, it is so unlikely that we have a law saying that it is impossible, which has never been observed to be violated—the second law of thermodynamics.

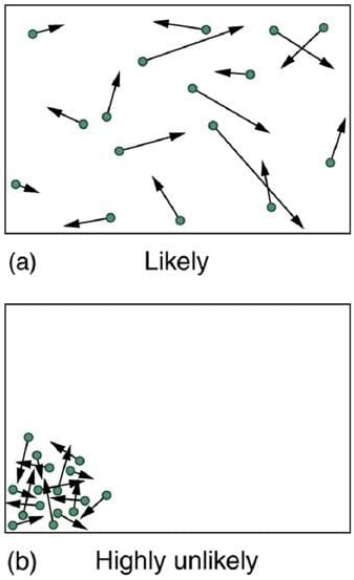


Figure 3.4.13.2: (a) The ordinary state of gas in a container is a disorderly, random distribution of atoms or molecules with a Maxwell-Boltzmann distribution of speeds. It is so unlikely that these atoms or molecules would ever end up in one corner of the container that it might as well be impossible. (b) With energy transfer, the gas can be forced into one corner and its entropy greatly reduced. But left alone, it will spontaneously increase its entropy and return to the normal conditions, because they are immensely more likely.

The disordered condition is one of high entropy, and the ordered one has low entropy. With a transfer of energy from another system, we could force all of the atoms into one corner and have a local decrease in entropy, but at the cost of an overall increase in entropy of the universe. If the atoms start out in one corner, they will quickly disperse and become uniformly distributed and will never return to the orderly original state (Figure 3.4.13.2(b)). Entropy will increase. With such a large sample of atoms, it is possible—but unimaginably unlikely—for entropy to decrease. Disorder is vastly more likely than order.

The arguments that disorder and high entropy are the most probable states are quite convincing. The great Austrian physicist Ludwig Boltzmann (1844–1906)—who, along with Maxwell, made so many contributions to kinetic theory—proved that the entropy of a system in a given state (a macrostate) can be written as

$$S = k \ln W,$$

where $k = 1.38 \times 10^{-23}\ \text{J/K}$ is Boltzmann's constant, and $\ln W$ is the natural logarithm of the number of microstates W corresponding to the given macrostate. W is proportional to the probability that the macrostate will occur. Thus entropy is directly related to the probability of a state—the more likely the state, the greater its entropy. Boltzmann proved that this expression for S is equivalent to the definition $\Delta S = Q/T$, which we have used extensively.

Thus the second law of thermodynamics is explained on a very basic level: entropy either remains the same or increases in every process. This phenomenon is due to the extraordinarily small probability of a decrease, based on the extraordinarily larger number of microstates in systems with greater entropy. Entropy *can* decrease, but for any macroscopic system, this outcome is so unlikely that it will never be observed.

Example 3.4.13.1: Entropy Increases in a Coin Toss

Suppose you toss 100 coins starting with 60 heads and 40 tails, and you get the most likely result, 50 heads and 50 tails. What is the change in entropy?

Strategy

Noting that the number of microstates is labeled W in Table 3.4.13.3 for the 100-coin toss, we can use $\Delta S = S_\text{f} - S_\text{i} = k \ln W_\text{f} - k \ln W_\text{i}$ to calculate the change in entropy.

Solution

The change in entropy is

$$\Delta S = S_\text{f} - S_\text{i} = k \ln W_\text{f} - k \ln W_\text{i},$$

where the subscript i stands for the initial 60 heads and 40 tails state, and the subscript f for the final 50 heads and 50 tails state. Substituting the values for W from Table 3.4.13.3 gives

$$\begin{aligned} \Delta S &= \left(1.38 \times 10^{-23}\ \text{J/K}\right)\left[\ln\left(1.0 \times 10^{29}\right) - \ln\left(1.4 \times 10^{28}\right)\right] \\ &= 2.7 \times 10^{-23}\ \text{J/K} \end{aligned}$$

Discussion

This increase in entropy means we have moved to a less orderly situation. It is not impossible for further tosses to produce the initial state of 60 heads and 40 tails, but it is less likely. There is about a 1 in 90 chance for that decrease in entropy (-2.7×10^{-23} J/K) to occur. If we calculate the decrease in entropy to move to the most orderly state, we get $\Delta S = -92 \times 10^{-23}\ \text{J/K}$. There is about a **1 in 10^{30}** chance of this change occurring. So while very small decreases in entropy are unlikely, slightly greater decreases are impossibly unlikely. These probabilities imply, again, that for a macroscopic system, a decrease in entropy is impossible. For example, for heat transfer to occur spontaneously from 1.00 kg of 0°C ice to its 0°C environment, there would be a decrease in entropy of 1.22×10^3 J/K. Given that a ΔS of $10^{-21}\ \text{J/K}$ corresponds to about a 1 in 1030 chance, a decrease of this size (10^3 J/K) is an *utter* impossibility. Even for a milligram of melted ice to spontaneously refreeze is impossible.

PROBLEM-SOLVING STRATEGIES FOR ENTROPY

1. *Examine the situation to determine if entropy is involved.*
2. *Identify the system of interest and draw a labeled diagram of the system showing energy flow.*
3. *Identify exactly what needs to be determined in the problem (identify the unknowns).* A written list is useful.
4. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).* You must carefully identify the heat transfer, if any, and the temperature at which the process takes place. It is also important to identify the initial and final states.
5. *Solve the appropriate equation for the quantity to be determined (the unknown).* Note that the change in entropy can be determined between any states by calculating it for a reversible process.

6. *Substitute the known value along with their units into the appropriate equation, and obtain numerical solutions complete with units.*
7. *To see if it is reasonable: Does it make sense?* For example, total entropy should increase for any real process or be constant for a reversible process. Disordered states should be more probable and have greater entropy than ordered states.

Section Summary

- Disorder is far more likely than order, which can be seen statistically.
- The entropy of a system in a given state (a macrostate) can be written as

$$S = k \ln W,$$

where k=1.38×10^{-23} J/K is Boltzmann's constant, and lnW is the natural logarithm of the number of microstates W corresponding to the given macrostate.

Glossary

macrostate
an overall property of a system

microstate
each sequence within a larger macrostate

statistical analysis
using statistics to examine data, such as counting microstates and macrostates

This page titled 3.4.13: Statistical Interpretation of Entropy and the Second Law of Thermodynamics- The Underlying Explanation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.14: Statistical Interpretation of Entropy and the Second Law of Thermodynamics- The Underlying Explanation by Andrew Park is licensed CC BY 4.0.

3.4.14: Applications of Thermodynamics- Heat Pumps and Refrigerators

Learning Objectives

- Describe the use of heat engines in heat pumps and refrigerators.
- Demonstrate how a heat pump works to warm an interior space.
- Explain the differences between heat pumps and refrigerators.
- Calculate a heat pump's coefficient of performance.



Figure 3.4.14.1: Almost every home contains a refrigerator. Most people don't realize they are also sharing their homes with a heat pump. (credit: Id1337x, Wikimedia Commons)

Heat pumps, air conditioners, and refrigerators utilize heat transfer from cold to hot. They are heat engines run backward. We say backward, rather than reverse, because except for Carnot engines, all heat engines, though they can be run backward, cannot truly be reversed. Heat transfer occurs from a cold reservoir Q_{c} and into a hot one. This requires work input W, which is also converted to heat transfer. Thus the heat transfer to the hot reservoir is $Q_{\mathrm{h}} = Q_{\mathrm{c}} + W$. (Note that Q_{h}, Q_{c}, and W are positive, with their directions indicated on schematics rather than by sign.) A heat pump's mission is for heat transfer Q_{h} to occur into a warm environment, such as a home in the winter. The mission of air conditioners and refrigerators is for heat transfer Q_{c} to occur from a cool environment, such as chilling a room or keeping food at lower temperatures than the environment. (Actually, a heat pump can be used both to heat and cool a space. It is essentially an air conditioner and a heating unit all in one. In this section we will concentrate on its heating mode.)

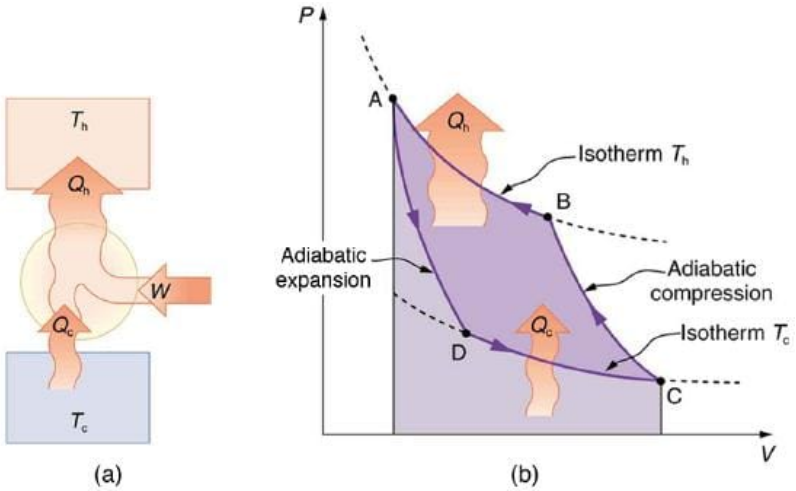


Figure 3.4.14.2: Heat pumps, air conditioners, and refrigerators are heat engines operated backward. The one shown here is based on a Carnot (reversible) engine. (a) Schematic diagram showing heat transfer from a cold reservoir to a warm reservoir with a heat pump. The directions of W, Q_{h}, and Q_{c} are opposite what they would be in a heat engine. (b) PV diagram for a Carnot cycle similar to that in Figure 3.4.14.3 but reversed, following path ADCBA. The area inside the loop is negative, meaning there is a net work input. There is heat transfer Q_{c} into the system from a cold reservoir along path DC, and heat transfer Q_{h} out of the system into a hot reservoir along path BA.

Heat Pumps

The great advantage of using a heat pump to keep your home warm, rather than just burning fuel, is that a heat pump supplies $Q_{\mathrm{h}} = Q_{\mathrm{c}} + W$. Heat transfer is from the outside air, even at a temperature below freezing, to the indoor space. You only pay for W, and you get an additional heat transfer of Q_{c} from the outside at no cost; in many cases, at least twice as much energy is transferred to the heated space as is used to run the heat pump. When you burn fuel to keep warm, you pay for all of it. The disadvantage is that the work input (required by the second law of thermodynamics) is sometimes more expensive than simply burning fuel, especially if the work is done by electrical energy.

The basic components of a heat pump in its heating mode are shown in Figure 3.4.14.3. A working fluid such as a non-CFC refrigerant is used. In the outdoor coils (the evaporator), heat transfer Q_{c} occurs to the working fluid from the cold outdoor air, turning it into a gas.

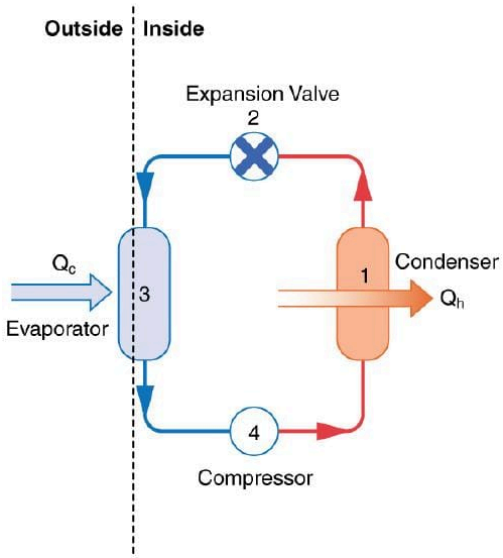


Figure 3.4.14.3: A simple heat pump has four basic components: (1) condenser, (2) expansion valve, (3) evaporator, and (4) compressor. In the heating mode, heat transfer Q_c occurs to the working fluid in the evaporator (3) from the colder outdoor air, turning it into a gas. The electrically driven compressor (4) increases the temperature and pressure of the gas and forces it into the condenser coils (1) inside the heated space. Because the temperature of the gas is higher than the temperature in the room, heat transfer from the gas to the room occurs as the gas condenses to a liquid. The working fluid is then cooled as it flows back through an expansion valve (2) to the outdoor evaporator coils.

The electrically driven compressor (work input W) raises the temperature and pressure of the gas and forces it into the condenser coils that are inside the heated space. Because the temperature of the gas is higher than the temperature inside the room, heat transfer to the room occurs and the gas condenses to a liquid. The liquid then flows back through a pressure-reducing valve to the outdoor evaporator coils, being cooled through expansion. (In a cooling cycle, the evaporator and condenser coils exchange roles and the flow direction of the fluid is reversed.)

The quality of a heat pump is judged by how much heat transfer Q_{h} occurs into the warm space compared with how much work input W is required. In the spirit of taking the ratio of what you get to what you spend, we define a **heat pump's coefficient of performance** (COP_{hp}) to be

$$COP_{\text{hp}} = \frac{Q_{\text{h}}}{W}.$$

Since the efficiency of a heat engine is $Eff = W/Q_{\text{h}}$, we see that $COP_{\text{hp}} = 1/Eff$, an important and interesting fact. First, since the efficiency of any heat engine is less than 1, it means that COP_{hp} is always greater than 1—that is, a heat pump always has more heat transfer Q_{h} than work put into it. Second, it means that heat pumps work best when temperature differences are small. The efficiency of a perfect, or Carnot, engine is $Eff_{\text{C}} = 1 - (T_{\text{c}}/T_{\text{h}})$; thus, the smaller the temperature difference, the smaller the efficiency and the greater the COP_{hp} (because $COP_{\text{hp}} = 1/Eff$). In other words, heat pumps do not work as well in very cold climates as they do in more moderate climates.

Friction and other irreversible processes reduce heat engine efficiency, but they do *not* benefit the operation of a heat pump—instead, they reduce the work input by converting part of it to heat transfer back into the cold reservoir before it gets into the heat pump.

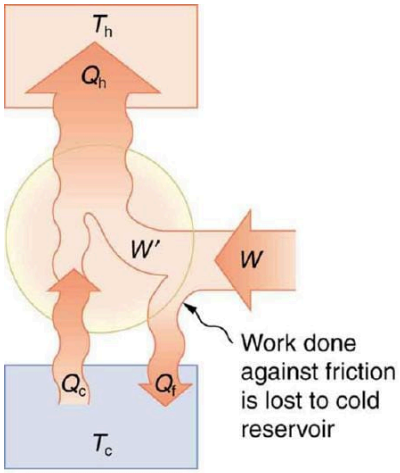


Figure 3.4.14.4: When a real heat engine is run backward, some of the intended work input (W) goes into heat transfer before it gets into the heat engine, thereby reducing its coefficient of performance COP_{hp}. In this figure, W' represents the portion of W that goes into the heat pump, while the remainder of W is lost in the form of frictional heat (Q_f) to the cold reservoir. If all of W had gone into the heat pump, then Q_{h} would have been greater. The best heat pump uses adiabatic and isothermal processes, since, in theory, there would be no dissipative processes to reduce the heat transfer to the hot reservoir.

Example 3.4.14.1: The Best COP hp of a Heat Pump for Home Use

A heat pump used to warm a home must employ a cycle that produces a working fluid at temperatures greater than typical indoor temperature so that heat transfer to the inside can take place. Similarly, it must produce a working fluid at temperatures that are colder than the outdoor temperature so that heat transfer occurs from outside. Its hot and cold reservoir temperatures therefore cannot be too close, placing a limit on its COP_{hp}. (See Figure 3.4.14.5.) What is the best coefficient of performance possible for such a heat pump, if it has a hot reservoir temperature of 45.0°C and a cold reservoir temperature of -15.0°C?

Strategy

A Carnot engine reversed will give the best possible performance as a heat pump. As noted above, $COP_{\text{hp}} = 1/Eff$, so that we need to first calculate the Carnot efficiency to solve this problem.

Solution

Carnot efficiency in terms of absolute temperature is given by**:**

$$Eff_{\text{C}} = 1 - \frac{T_{\text{c}}}{T_{\text{h}}}.$$

The temperatures in kelvins are $T_{\text{h}} = 318\ \text{K}$ and $T_{\text{c}} = 258\ \text{K}$, so that

$$Eff_{\text{C}} = 1 - \frac{258\ \text{K}}{318\ \text{K}} = 0.1887.$$

Thus, from the discussion above,

$$COP_{\text{hp}} = \frac{1}{Eff} = \frac{1}{0.1887} = 5.30,$$

or

$$COP_{\text{hp}} = \frac{Q_{\text{h}}}{W} = 5.30,$$

so that

$$Q_{\text{h}} = 5.30 \text{ W}.$$

Discussion

This result means that the heat transfer by the heat pump is 5.30 times as much as the work put into it. It would cost 5.30 times as much for the same heat transfer by an electric room heater as it does for that produced by this heat pump. This is not a violation of conservation of energy. Cold ambient air provides 4.3 J per 1 J of work from the electrical outlet.

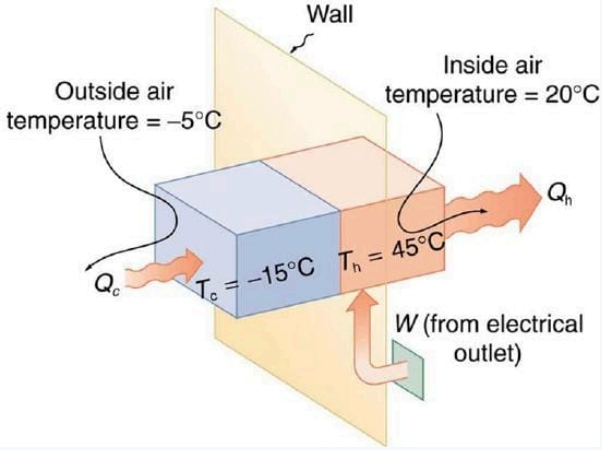


Figure 3.4.14.5: Heat transfer from the outside to the inside, along with work done to run the pump, takes place in the heat pump of the example above**.** Note that the cold temperature produced by the heat pump is lower than the outside temperature, so that heat transfer into the working fluid occurs. The pump's compressor produces a temperature greater than the indoor temperature in order for heat transfer into the house to occur.

Real heat pumps do not perform quite as well as the ideal one in the previous example; their values of COP_{hp} range from about 2 to 4. This range means that the heat transfer Q_{h} from the heat pumps is 2 to 4 times as great as the work W put into them. Their economical feasibility is still limited, however, since W is usually supplied by electrical energy that costs more per joule than heat transfer by burning fuels like natural gas. Furthermore, the initial cost of a heat pump is greater than that of many furnaces, so that a heat pump must last longer for its cost to be recovered. Heat pumps are most likely to be economically superior where winter temperatures are mild, electricity is relatively cheap, and other fuels are relatively expensive. Also, since they can cool as well as heat a space, they have advantages where cooling in summer months is also desired. Thus some of the best locations for heat pumps are in warm summer climates with cool winters. Figure 3.4.14.6 shows a heat pump, called a *"reverse cycle"* or *"split-system cooler"* in some countries.



Figure 3.4.14.6: In hot weather, heat transfer occurs from air inside the room to air outside, cooling the room. In cool weather, heat transfer occurs from air outside to air inside, warming the room. This switching is achieved by reversing the direction of flow of the working fluid.

Air Conditioners and Refrigerators

Air conditioners and refrigerators are designed to cool something down in a warm environment. As with heat pumps, work input is required for heat transfer from cold to hot, and this is expensive. The quality of air conditioners and refrigerators is judged by how much heat transfer Q_{c} occurs from a cold environment compared with how much work input W is required. What is considered

the benefit in a heat pump is considered waste heat in a refrigerator. We thus define the **coefficient of performance** (COP_{ref}) of an air conditioner or refrigerator to be

$$COP_{\text{ref}} = \frac{Q_{\text{c}}}{W}.$$

Noting again that $Q_{\text{h}} = Q_{\text{c}} + W$, we can see that an air conditioner will have a lower coefficient of performance than a heat pump, because $COP_{\text{hp}} = Q_{\text{h}}/W$ and Q_{h} is greater than Q_{c}. In this module's Problems and Exercises, you will show that

$$COP_{\text{ref}} = COP_{\text{hp}} - 1$$

for a heat engine used as either an air conditioner or a heat pump operating between the same two temperatures. Real air conditioners and refrigerators typically do remarkably well, having values of COP_{ref} ranging from 2 to 6. These numbers are better than the COP_{hp} values for the heat pumps mentioned above, because the temperature differences are smaller, but they are less than those for Carnot engines operating between the same two temperatures.

A type of COP rating system called the "energy efficiency rating" (EER) has been developed. This rating is an example where non-SI units are still used and relevant to consumers. To make it easier for the consumer, Australia, Canada, New Zealand, and the U.S. use an Energy Star Rating out of 5 stars—the more stars, the more energy efficient the appliance. EERs are expressed in mixed units of British thermal units (Btu) per hour of heating or cooling divided by the power input in watts. Room air conditioners are readily available with EERs ranging from 6 to 12. Although not the same as the COPs just described, these EERs are good for comparison purposes—the greater the EER , the cheaper an air conditioner is to operate (but the higher its purchase price is likely to be).

The EER of an air conditioner or refrigerator can be expressed as

$$EER = \frac{Q_{\text{c}}/t_1}{W/t_2},$$

where Q_{c} is the amount of heat transfer from a cold environment in British thermal units, t_1 is time in hours, W is the work input in joules, and t_2 is time in seconds.

PROBLEM-SOLVING STRATEGIES FOR THERMODYNAMICS

1. *Examine the situation to determine whether heat, work, or internal energy are involved.* Look for any system where the primary methods of transferring energy are heat and work. Heat engines, heat pumps, refrigerators, and air conditioners are examples of such systems.
2. *Identify the system of interest and draw a labeled diagram of the system showing energy flow.*
3. *Identify exactly what needs to be determined in the problem (identify the unknowns).* A written list is useful. Maximum efficiency means a Carnot engine is involved. Efficiency is not the same as the coefficient of performance.
4. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).* Be sure to distinguish heat transfer into a system from heat transfer out of the system, as well as work input from work output. In many situations, it is useful to determine the type of process, such as isothermal or adiabatic.
5. *Solve the appropriate equation for the quantity to be determined (the unknown).*
6. *Substitute the known quantities along with their units into the appropriate equation and obtain numerical solutions complete with units.*
7. *Check the answer to see if it is reasonable: Does it make sense?* For example, efficiency is always less than 1, whereas coefficients of performance are greater than 1.

Section Summary

- An artifact of the second law of thermodynamics is the ability to heat an interior space using a heat pump. Heat pumps compress cold ambient air and, in so doing, heat it to room temperature without violation of conservation principles.
- To calculate the heat pump's coefficient of performance, use the equation $COP_{\text{hp}} = \frac{Q_{\text{h}}}{W}$.
- A refrigerator is a heat pump; it takes warm ambient air and expands it to chill it.

Glossary

heat pump
: a machine that generates heat transfer from cold to hot

coefficient of performance
: for a heat pump, it is the ratio of heat transfer at the output (the hot reservoir) to the work supplied; for a refrigerator or air conditioner, it is the ratio of heat transfer from the cold reservoir to the work supplied

This page titled 3.4.14: Applications of Thermodynamics- Heat Pumps and Refrigerators is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.12: Applications of Thermodynamics- Heat Pumps and Refrigerators by Andrew Park is licensed CC BY 4.0.

3.4.15: Thermal Physics (Exercises)

Conceptual Questions

8.1: Temperature

1. What does it mean to say that two systems are in thermal equilibrium?

2. Give an example of a physical property that varies with temperature and describe how it is used to measure temperature.

3. When a cold alcohol thermometer is placed in a hot liquid, the column of alcohol goes *down* slightly before going up. Explain why.

4. If you add boiling water to a cup at room temperature, what would you expect the final equilibrium temperature of the unit to be? You will need to include the surroundings as part of the system. Consider the zeroth law of thermodynamics.

8.2: The Ideal Gas Law

5. Under what circumstances would you expect a gas to behave significantly differently than predicted by the ideal gas law?

6. A constant-volume gas thermometer contains a fixed amount of gas. What property of the gas is measured to indicate its temperature?

8.3: Heat

7. How is heat transfer related to temperature?

8. Describe a situation in which heat transfer occurs. What are the resulting forms of energy?

9. When heat transfers into a system, is the energy stored as heat? Explain briefly.

8.4: Heat Transfer Methods

10. What are the main methods of heat transfer from the hot core of Earth to its surface? From Earth's surface to outer space?

11. When our bodies get too warm, they respond by sweating and increasing blood circulation to the surface to transfer thermal energy away from the core. What effect will this have on a person in a 40.0°C hot tub?

12. Figure 3.4.15.1 shows a cut-away drawing of a thermos bottle (also known as a Dewar flask), which is a device designed specifically to slow down all forms of heat transfer. Explain the functions of the various parts, such as the vacuum, the silvering of the walls, the thin-walled long glass neck, the rubber support, the air layer, and the stopper.

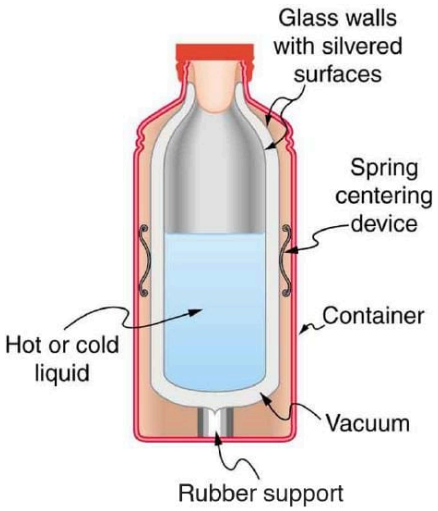


Figure 3.4.15.1: The construction of a thermos bottle is designed to inhibit all methods of heat transfer.

8.5: Temperature Change and Heat Capacity

13. What three factors affect the heat transfer that is necessary to change an object's temperature?

14. The brakes in a car increase in temperature by ΔT when bringing the car to rest from a speed v. How much greater would ΔT be if the car initially had twice the speed? You may assume the car to stop sufficiently fast so that no heat transfers out of the brakes.

8.6: Phase Change and Latent Heat

15. Heat transfer can cause temperature and phase changes. What else can cause these changes?

16. How does the latent heat of fusion of water help slow the decrease of air temperatures, perhaps preventing temperatures from falling significantly below $0^{\circ}\mathrm{C}$, in the vicinity of large bodies of water?

17. What is the temperature of ice right after it is formed by freezing water?

18. If you place $0^{\circ}\mathrm{C}$ ice into $0^{\circ}\mathrm{C}$ water in an insulated container, what will happen? Will some ice melt, will more water freeze, or will neither take place?

19. What effect does condensation on a glass of ice water have on the rate at which the ice melts? Will the condensation speed up the melting process or slow it down?

20. In very humid climates where there are numerous bodies of water, such as in Florida, it is unusual for temperatures to rise above about 35°C(95°F). In deserts, however, temperatures can rise far above this. Explain how the evaporation of water helps limit high temperatures in humid climates.

21. In winters, it is often warmer in San Francisco than in nearby Sacramento, 150 km inland. In summers, it is nearly always hotter in Sacramento. Explain how the bodies of water surrounding San Francisco moderate its extreme temperatures.

22. Putting a lid on a boiling pot greatly reduces the heat transfer necessary to keep it boiling. Explain why.

23. Freeze-dried foods have been dehydrated in a vacuum. During the process, the food freezes and must be heated to facilitate dehydration. Explain both how the vacuum speeds up dehydration and why the food freezes as a result.

24. When still air cools by radiating at night, it is unusual for temperatures to fall below the dew point. Explain why.

25. In a physics classroom demonstration, an instructor inflates a balloon by mouth and then cools it in liquid nitrogen. When cold, the shrunken balloon has a small amount of light blue liquid in it, as well as some snow-like crystals. As it warms

up, the liquid boils, and part of the crystals sublimate, with some crystals lingering for awhile and then producing a liquid. Identify the blue liquid and the two solids in the cold balloon. Justify your identifications using data from Table 9.6.1.

8.7: The First Law of Thermodynamics

26. Describe the photo of the tea kettle at the beginning of this section in terms of heat transfer, work done, and internal energy. How is heat being transferred? What is the work done and what is doing it? How does the kettle maintain its internal energy?

27. The first law of thermodynamics and the conservation of energy are clearly related. How do they differ in the types of energy considered?

28. Heat transfer Q and work done W are always energy in transit, whereas internal energy U is energy stored in a system. Give an example of each type of energy, and state specifically how it is either in transit or resides in a system.

29. How do heat transfer and internal energy differ? In particular, which can be stored as such in a system and which cannot?

30. If you run down some stairs and stop, what happens to your kinetic energy and your initial gravitational potential energy?

31. Give an explanation of how food energy (calories) can be viewed as molecular potential energy (consistent with the atomic and molecular definition of internal energy).

32. Identify the type of energy transferred to your body in each of the following as either internal energy, heat transfer, or doing work:

(a) basking in sunlight;

(b) eating food;

(c) riding an elevator to a higher floor.

8.8: The First Law of Thermodynamics and Heat Engine Processes

33. A great deal of effort, time, and money has been spent in the quest for the so-called perpetual-motion machine, which is defined as a hypothetical machine that operates or produces useful work indefinitely and/or a hypothetical machine that produces more work or energy than it consumes. Explain, in terms of heat engines and the first law of thermodynamics, why or why not such a machine is likely to be constructed.

34. One method of converting heat transfer into doing work is for heat transfer into a gas to take place, which expands, doing work on a piston, as shown in the figure below.

(a) Is the heat transfer converted directly to work in an isobaric process, or does it go through another form first? Explain your answer.

(b) What about in an isothermal process?

(c) What about in an adiabatic process (where heat transfer occurred prior to the adiabatic process)?

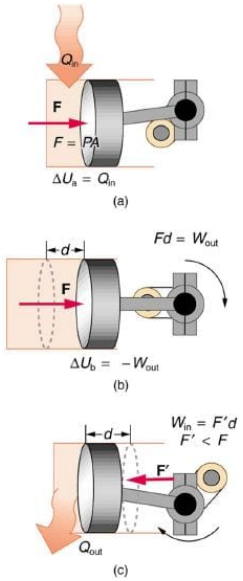


Figure 3.4.15.2

35. Would the previous question make any sense for an isochoric process? Explain your answer.

36. We ordinarily say that $\Delta U = 0$ for an isothermal process. Does this assume no phase change takes place? Explain your answer.

37. The temperature of a rapidly expanding gas decreases. Explain why in terms of the first law of thermodynamics. (Hint: Consider whether the gas does work and whether heat transfer occurs rapidly into the gas through conduction.)

38. A real process may be nearly adiabatic if it occurs over a very short time. How does the short time span help the process to be adiabatic?

39. It is unlikely that a process can be isothermal unless it is a very slow process. Explain why. Is the same true for isobaric and isochoric processes? Explain your answer.

8.9: Introduction to the Second Law of Thermodynamics- Heat Engines and Their Efficiency

40. Imagine you are driving a car up Pike's Peak in Colorado. To raise a car weighing 1000 kilograms a distance of 100 meters would require about a million joules. You could raise a car 12.5 kilometers with the energy in a gallon of gas. Driving up Pike's Peak (a mere 3000-meter climb) should consume a little less than a quart of gas. But other considerations have to be taken into account. Explain, in terms of efficiency, what factors may keep you from realizing your ideal energy use on this trip.

41. Is a temperature difference necessary to operate a heat engine? State why or why not.

42. Definitions of efficiency vary depending on how energy is being converted. Compare the definitions of efficiency for the human body and heat engines. How does the definition of efficiency in each relate to the type of energy being converted into doing work?

43. Why—other than the fact that the second law of thermodynamics says reversible engines are the most efficient—should heat engines employing reversible processes be more efficient than those employing irreversible processes? Consider that dissipative mechanisms are one cause of irreversibility.

8.10: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated

44. Think about the drinking bird at the beginning of this section (Figure 9.10.1). Although the bird enjoys the theoretical maximum efficiency possible, if left to its own devices over time, the bird will cease "drinking." What are some of the dissipative processes that might cause the bird's motion to cease?

45. Can improved engineering and materials be employed in heat engines to reduce heat transfer into the environment? Can they eliminate heat transfer into the environment entirely?

46. Does the second law of thermodynamics alter the conservation of energy principle?

8.11: Applications of Thermodynamics- Heat Pumps and Refrigerators

47. Explain why heat pumps do not work as well in very cold climates as they do in milder ones. Is the same true of refrigerators?

48. In some Northern European nations, homes are being built without heating systems of any type. They are very well insulated and are kept warm by the body heat of the residents. However, when the residents are not at home, it is still warm in these houses. What is a possible explanation?

49. Why do refrigerators, air conditioners, and heat pumps operate most cost-effectively for cycles with a small difference between T_{h} and T_{c}? (Note that the temperatures of the cycle employed are crucial to its COP.)

50. Grocery store managers contend that there is *less* total energy consumption in the summer if the store is kept at a *low* temperature. Make arguments to support or refute this claim, taking into account that there are numerous refrigerators and freezers in the store.

51. Can you cool a kitchen by leaving the refrigerator door open?

8.12: Entropy and the Second Law of Thermodynamics- Disorder and the Unavailability of Energy

52. Does a gas become more orderly when it liquefies? Does its entropy change? If so, does the entropy increase or decrease? Explain your answer.

53. Explain how water's entropy can decrease when it freezes without violating the second law of thermodynamics. Specifically, explain what happens to the entropy of its surroundings.

54. Is a uniform-temperature gas more or less orderly than one with several different temperatures? Which is more structured? In which can heat transfer result in work done without heat transfer from another system?

55. Give an example of a spontaneous process in which a system becomes less ordered and energy becomes less available to do work. What happens to the system's entropy in this process?

56. What is the change in entropy in an adiabatic process? Does this imply that adiabatic processes are reversible? Can a process be precisely adiabatic for a macroscopic system?

57. Does the entropy of a star increase or decrease as it radiates? Does the entropy of the space into which it radiates (which has a temperature of about 3 K) increase or decrease? What does this do to the entropy of the universe?

58. Explain why a building made of bricks has smaller entropy than the same bricks in a disorganized pile. Do this by considering the number of ways that each could be formed (the number of microstates in each macrostate).

8.13: Statistical Interpretation of Entropy and the Second Law of Thermodynamics- The Underlying Explanation

59. Explain why a building made of bricks has smaller entropy than the same bricks in a disorganized pile. Do this by

considering the number of ways that each could be formed (the number of microstates in each macrostate).

Problems & Exercises

8.1: Temperature

1. What is the Fahrenheit temperature of a person with a 39.0°C fever?

Answer

102°F

2. Frost damage to most plants occurs at temperatures of 28.0°F or lower. What is this temperature on the Kelvin scale?

3. To conserve energy, room temperatures are kept at 68.0°F in the winter and 78.0°F in the summer. What are these temperatures on the Celsius scale?

Answer

20.0°C and 25.6°C

4. A tungsten light bulb filament may operate at 2900 K. What is its Fahrenheit temperature? What is this on the Celsius scale?

5. The surface temperature of the Sun is about 5750 K. What is this temperature on the Fahrenheit scale?

Answer

9890°F

6. One of the hottest temperatures ever recorded on the surface of Earth was 134°F in Death Valley, CA. What is this temperature in Celsius degrees? What is this temperature in Kelvin?

7. (a) Suppose a cold front blows into your locale and drops the temperature by 40.0 Fahrenheit degrees. How many degrees Celsius does the temperature decrease when there is a 40.0°F decrease in temperature?

(b) Show that any change in temperature in Fahrenheit degrees is nine-fifths the change in Celsius degrees.

Answer

(a) 22.2°C

(b)
$$\begin{aligned}\Delta T\,(^\circ\mathrm{F}) &= T_2\,(^\circ\mathrm{F}) - T_1\,(^\circ\mathrm{F}) \\ &= \frac{9}{5}T_2\,(^\circ\mathrm{C}) + 32.0^\circ - \left(\frac{9}{5}T_1\,(^\circ\mathrm{C}) + 32.0^\circ\right) \\ &= \frac{9}{5}\left(T_2\,(^\circ\mathrm{C}) - T_1\,(^\circ\mathrm{C})\right) = \frac{9}{5}\Delta T\,(^\circ\mathrm{C})\end{aligned}$$

8. (a) At what temperature do the Fahrenheit and Celsius scales have the same numerical value?

(b) At what temperature do the Fahrenheit and Kelvin scales have the same numerical value?

8.2: The Ideal Gas Law

9. The gauge pressure in your car tires is $2.50 \times 10^5\ \mathrm{N/m^2}$ at a temperature of 35.0°C when you drive it onto a ferry boat to Alaska. What is their gauge pressure later, when their temperature has dropped to −40.0°C?

Answer

1.62 atm

10. Convert an absolute pressure of $7.00 \times 10^5\ \mathrm{N/m^2}$ to gauge pressure in $\mathrm{lb/in^2}$. (This value was stated to be just less than $90.0\ \mathrm{lb/in^2}$ in Example 9.2.1. Is it?)

11. Suppose a gas-filled incandescent light bulb is manufactured so that the gas inside the bulb is at atmospheric pressure when the bulb has a temperature of 20.0°C.

(a) Find the gauge pressure inside such a bulb when it is hot, assuming its average temperature is 60.0°C (an approximation) and neglecting any change in volume due to thermal expansion or gas leaks.

(b) The actual final pressure for the light bulb will be less than calculated in part (a) because the glass bulb will expand. What will the actual final pressure be, taking this into account? Is this a negligible difference?

Answer

(a) 0.136 atm

(b) 0.135 atm. The difference between this value and the value from part (a) is negligible.

12. Large helium-filled balloons are used to lift scientific equipment to high altitudes. (a) What is the pressure inside such a balloon if it starts out at sea level with a temperature of 10.0°C and rises to an altitude where its volume is twenty times the original volume and its temperature is –50.0°C? (b) What is the gauge pressure? (Assume atmospheric pressure is constant.)

13. In the text, it was shown that $N/V = 2.68 \times 10^{25}\ \mathrm{m^{-3}}$ for gas at STP.

(a) Show that this quantity is equivalent to $N/V = 2.68 \times 10^{19}\ \mathrm{cm^{-3}}$, as stated.

(b) About how many atoms are there in one $\mu\mathrm{m}^3$ (a cubic micrometer) at STP?

(c) What does your answer to part (b) imply about the separation of atoms and molecules?

14. An airplane passenger has $100\ \mathrm{cm^3}$ of air in his stomach just before the plane takes off from a sea-level airport. What volume will the air have at cruising altitude if cabin pressure drops to $7.50 \times 10^4\ \mathrm{N/m^2}$?

15. An expensive vacuum system can achieve a pressure as low as $1.00 \times 10^{-7}\ \mathrm{N/m^2}$ at 20°C. How many atoms are there in a cubic centimeter at this pressure and temperature?

16. The number density of gas atoms at a certain location in the space above our planet is about $1.00 \times 10^{11}\ \mathrm{m^{-3}}$, and the pressure is $2.75 \times 10^{-10}\ \mathrm{N/m^2}$ in this space. What is the temperature there?

Answer

–73.9°C

17. A bicycle tire has a pressure of $7.00 \times 10^5\ \mathrm{N/m^2}$ at a temperature of 18.0°C and contains 2.00 L of gas. What will its pressure be if you let out an amount of air that has a volume of 100 cm^3 at atmospheric pressure? Assume tire temperature and volume remain constant.

18. A high-pressure gas cylinder contains 50.0 L of toxic gas at a pressure of $1.40 \times 10^7\ \mathrm{N/m^2}$ and a temperature of 25.0°C. Its valve leaks after the cylinder is dropped. The cylinder is cooled to dry ice temperature (–78.5°C) to reduce the leak rate and pressure so that it can be safely repaired.

(a) What is the final pressure in the tank, assuming a negligible amount of gas leaks while being cooled and that there is no phase change?

(b) What is the final pressure if one-tenth of the gas escapes?

(c) To what temperature must the tank be cooled to reduce the pressure to 1.00 atm (assuming the gas does not

change phase and that there is no leakage during cooling)?

(d) Does cooling the tank appear to be a practical solution?

Answer

(a) 9.14×10^{6} N/m^2

(b) 8.23×10^{6} N/m^2

(c) 2.16 K

(d) No. The final temperature needed is much too low to be easily achieved for a large object.

19. (a) What is the gauge pressure in a 25.0°C car tire containing 3.60 mol of gas in a 30.0 L volume?

(b) What will its gauge pressure be if you add 1.00 L of gas originally at atmospheric pressure and 25.0°C? Assume the temperature returns to 25.0°C and the volume remains constant.

8.10: Carnot's Perfect Heat Engine- The Second Law of Thermodynamics Restated

20. A certain gasoline engine has an efficiency of 30.0%. What would the hot reservoir temperature be for a Carnot engine having that efficiency, if it operates with a cold reservoir temperature of $200^{\circ}\mathrm{C}$?

Solution

403°C

21. A gas-cooled nuclear reactor operates between hot and cold reservoir temperatures of $700^{\circ}\mathrm{C}$ and $27.0^{\circ}\mathrm{C}$.

(a) What is the maximum efficiency of a heat engine operating between these temperatures?

(b) Find the ratio of this efficiency to the Carnot efficiency of a standard nuclear reactor (found in Example 9.10.1).

22. (a) What is the hot reservoir temperature of a Carnot engine that has an efficiency of 42.0% and a cold reservoir temperature of $27.0^{\circ}\mathrm{C}$?

(b) What must the hot reservoir temperature be for a real heat engine that achieves 0.700 of the maximum efficiency, but still has an efficiency of 42.0% (and a cold reservoir at $27.0^{\circ}\mathrm{C}$)?

(c) Does your answer imply practical limits to the efficiency of car gasoline engines?

Solution

(a) 244°C

(b) 477°C

(c) Yes, since automobiles engines cannot get too hot without overheating, their efficiency is limited.

23. Steam locomotives have an efficiency of 17.0% and operate with a hot steam temperature of $425^{\circ}\mathrm{C}$.

(a) What would the cold reservoir temperature be if this were a Carnot engine?

(b) What would the maximum efficiency of this steam engine be if its cold reservoir temperature were $150^{\circ}\mathrm{C}$?

24. Practical steam engines utilize $450^{\circ}\mathrm{C}$ steam, which is later exhausted at $270^{\circ}\mathrm{C}$.

(a) What is the maximum efficiency that such a heat engine can have?

(b) Since $270^{\circ}\mathrm{C}$ steam is still quite hot, a second steam engine is sometimes operated using the exhaust of the first. What is the maximum efficiency of the second engine if its exhaust has a temperature of $150^{\circ}\mathrm{C}$?

(c) What is the overall efficiency of the two engines? (d) Show that this is the same efficiency as a single Carnot engine operating between $450^{\circ}\mathrm{C}$ and $150^{\circ}\mathrm{C}$. Explicitly show how you follow the steps in the Problem-Solving Strategies for Thermodynamics.

Solution

(a) $Eff_1 = 1 - \frac{T_{c,1}}{T_{h,1}} = 1 - \frac{543\ \mathrm{K}}{723\ \mathrm{K}} = 0.249 \text{ or } 24.9\%$

(b) $Eff_2 = 1 - \frac{423\ \mathrm{K}}{543\ \mathrm{K}} = 0.221 \text{ or } 22.1\%$

(c) $Eff_1 = 1 - \dfrac{T_{c,l}}{T_{h,1}} \Rightarrow T_{c,1} = T_{h,1}\,(1, -, eff_1)$ similarly, $T_{c,2} = T_{h,2}\,(1 - Eff_2)$

$$T_{c,2} = T_{h,1}\,(1 - Eff_1)\,(1 - Eff_2) \equiv T_{h,1}\,(1 - Eff_{\text{overall}})$$

using $T_{h,2} = T_{c,1}$ in above equation gives $\therefore (1 - Eff_{\text{overall}}) = (1 - Eff_1)\,(1 - Eff_2)$

$$Eff_{\text{overall}} = 1 - (1 - 0.249)(1 - 0.221) = 41.5\%$$

(d) $Eff_{\text{overall}} = 1 - \frac{423\ \mathrm{K}}{723\ \mathrm{K}} = 0.415 \text{ or } 41.5\%$

25. A coal-fired electrical power station has an efficiency of 38%. The temperature of the steam leaving the boiler is $550^{\circ}\mathrm{C}$. What percentage of the maximum efficiency does this station obtain? (Assume the temperature of the environment is $20^{\circ}\mathrm{C}$.)

26. Would you be willing to financially back an inventor who is marketing a device that she claims has 25 kJ of heat transfer at 600 K, has heat transfer to the environment at 300 K, and does 12 kJ of work? Explain your answer.

Solution

The heat transfer to the cold reservoir is $Q_c = Q_h - W = 25\ \mathrm{kJ} - 12\ \mathrm{kJ} = 13\ \mathrm{kJ}$, so the efficiency is $Eff = 1 - \frac{Q_c}{Q_h} = 1 - \frac{13\ \mathrm{kJ}}{25\ \mathrm{kJ}} = 0.48$. The Carnot efficiency is $Eff_C = 1 - \frac{T_c}{T_h} = 1 - \frac{300\ \mathrm{K}}{600\ \mathrm{K}} = 0.50$. The actual efficiency is 96% of the Carnot efficiency, which is much higher than the best-ever achieved of about 70%, so her scheme is likely to be fraudulent.

Unreasonable Results

27. (a) Suppose you want to design a steam engine that has heat transfer to the environment at 270°C and has a Carnot efficiency of 0.800. What temperature of hot steam must you use?

(b) What is unreasonable about the temperature?

(c) Which premise is unreasonable?

Unreasonable Results

28. Calculate the cold reservoir temperature of a steam engine that uses hot steam at $450^{\circ}\mathrm{C}$ and has a Carnot efficiency of 0.700.

(b) What is unreasonable about the temperature?

(c) Which premise is unreasonable?

Solution

(a) –56.3°C

(b) The temperature is too cold for the output of a steam engine (the local environment). It is below the freezing point of water.

(c) The assumed efficiency is too high.

8.11: Applications of Thermodynamics- Heat Pumps and Refrigerators

29. What is the coefficient of performance of an ideal heat pump that has heat transfer from a cold temperature of –25.0°C to a hot temperature of 40.0°C?

Answer

4.82

30. Suppose you have an ideal refrigerator that cools an environment at –20.0°C and has heat transfer to another environment at 50.0°C. What is its coefficient of performance?

31. What is the best coefficient of performance possible for a hypothetical refrigerator that could make liquid nitrogen at –200°C and has heat transfer to the environment at 35.0°C?

Answer

0.311

32. In a very mild winter climate, a heat pump has heat transfer from an environment at 5.00°C to one at 35.0°C. What is the best possible coefficient of performance for these temperatures? Explicitly show how you follow the steps in the Problem-Solving Strategies for Thermodynamics.

33. (a) What is the best coefficient of performance for a heat pump that has a hot reservoir temperature of 50.0°C and a cold reservoir temperature of –20.0°C?

(b) How much heat transfer occurs into the warm environment if 3.60×10^7 J of work ($10.0\ \mathrm{kW \cdot h}$) is put into it?

(c) If the cost of this work input is $10.0\ \mathrm{cents\ /kW \cdot h}$, how does its cost compare with the direct heat transfer achieved by burning natural gas at a cost of 85.0 cents per therm. (A therm is a common unit of energy for natural gas and equals 1.055×10^8 J.)

Answer

(a) 4.61

(b) 1.66×10^8 J or $3.97 \times 10^4 \mathrm{kcal}$

(c) To transfer 1.66×10^8 J, heat pump costs $1.00, natural gas costs $1.34.

34. (a) What is the best coefficient of performance for a refrigerator that cools an environment at –30.0°C and has heat transfer to another environment at 45.0°C?

(b) How much work in joules must be done for a heat transfer of 4186 kJ from the cold environment?

(c) What is the cost of doing this if the work costs 10.0 cents per 3.60×10^6 J (a kilowatt-hour)?

(d) How many kJ of heat transfer occurs into the warm environment?

(e) Discuss what type of refrigerator might operate between these temperatures.

35. Suppose you want to operate an ideal refrigerator with a cold temperature of –10.0°C, and you would like it to have a coefficient of performance of 7.00. What is the hot reservoir temperature for such a refrigerator?

Answer

27.6°C

36. An ideal heat pump is being considered for use in heating an environment with a temperature of 22.0°C. What is the cold reservoir temperature if the pump is to have a coefficient of performance of 12.0?

37. A 4-ton air conditioner removes 5.06×10^7 J (48,000 British thermal units) from a cold environment in 1.00 h.

(a) What energy input in joules is necessary to do this if the air conditioner has an energy efficiency rating (EER) of 12.0?

(b) What is the cost of doing this if the work costs 10.0 cents per 3.60×10^6 J (one kilowatt-hour)?

(c) Discuss whether this cost seems realistic. Note that the energy efficiency rating (EER) of an air conditioner or refrigerator is defined to be the number of British thermal units of heat transfer from a cold environment per hour divided by the watts of power input.

Answer

(a) 1.44×10^7 J

(b) 40 cents

(c) This cost seems quite realistic; it says that running an air conditioner all day would cost $9.59 (if it ran continuously).

38. Show that the coefficients of performance of refrigerators and heat pumps are related by $COP_{\text{ref}} = COP_{\text{hp}} - 1$.

39. Start with the definitions of the COPs and the conservation of energy relationship between Q_{h}, Q_{c}, and W.

8.13: Statistical Interpretation of Entropy and the Second Law of Thermodynamics- The Underlying Explanation

40. Using Table 9.13.3, verify the contention that if you toss 100 coins each second, you can expect to get 100 heads or 100 tails once in 2×10^{22} years; calculate the time to two-digit accuracy.

Solution

It should happen twice in every 1.27×10^{30} s or once in every 6.35×10^{29} s

$$\left(6.35 \times 10^{29}\ \text{s}\right)\left(\tfrac{1\ \text{h}}{3600\ \text{s}}\right)\left(\tfrac{1\ \text{d}}{24\ \text{h}}\right)\left(\tfrac{1\text{y}}{365.25\ \text{d}}\right)$$
$$= 2.0 \times 10^{22}\text{y}$$

41. What percent of the time will you get something in the range from 60 heads and 40 tails through 40 heads and 60 tails when tossing 100 coins? The total number of microstates in that range is 1.22×10^{30}. (Consult Table 9.13.3.)

42. (a) If tossing 100 coins, how many ways (microstates) are there to get the three most likely macrostates of 49 heads and 51 tails, 50 heads and 50 tails, and 51 heads and 49 tails?

(b) What percent of the total possibilities is this? (Consult Table 9.13.3.)

Solution

(a) 3.0×10^{29}

(b) 24%

43. (a) What is the change in entropy if you start with 100 coins in the 45 heads and 55 tails macrostate, toss them, and get 51 heads and 49 tails?

(b) What if you get 75 heads and 25 tails?

(c) How much more likely is 51 heads and 49 tails than 75 heads and 25 tails?

(d) Does either outcome violate the second law of thermodynamics?

44. (a) What is the change in entropy if you start with 10 coins in the 5 heads and 5 tails macrostate, toss them, and get 2 heads and 8 tails?

(b) How much more likely is 5 heads and 5 tails than 2 heads and 8 tails? (Take the ratio of the number of microstates to find out.)

(c) If you were betting on 2 heads and 8 tails would you accept odds of 252 to 45? Explain why or why not.

Solution

(a) -2.38×10^{-23} J/K

(b) 5.6 times more likely

(c) If you were betting on two heads and 8 tails, the odds of breaking even are 252 to 45, so on average you would break even. So, no, you wouldn't bet on odds of 252 to 45.

Macrostate		Number of Microstates (W)
Heads	Tails	
10	0	1
9	1	10
8	2	45
7	3	120
6	4	210
5	5	252
4	6	210
3	7	120
2	8	45
1	9	10
0	10	1

45. (a) If you toss 10 coins, what percent of the time will you get the three most likely macrostates (6 heads and 4 tails, 5 heads and 5 tails, 4 heads and 6 tails)?

(b) You can realistically toss 10 coins and count the number of heads and tails about twice a minute. At that rate, how long will it take on average to get either 10 heads and 0 tails or 0 heads and 10 tails?

46. (a) Construct a table showing the macrostates and all of the individual microstates for tossing 6 coins. (Use Table 3.4.15.1 as a guide.)

(b) How many macrostates are there?

(c) What is the total number of microstates?

(d) What percent chance is there of tossing 5 heads and 1 tail?

(e) How much more likely are you to toss 3 heads and 3 tails than 5 heads and 1 tail? (Take the ratio of the number of microstates to find out.)

Solution

(b) 7

(c) 64

(d) 9.38%

(e) 3.33 times more likely (20 to 6)

47. In an air conditioner, 12.65 MJ of heat transfer occurs from a cold environment in 1.00 h.

(a) What mass of ice melting would involve the same heat transfer?

(b) How many hours of operation would be equivalent to melting 900 kg of ice?

(c) If ice costs 20 cents per kg, do you think the air conditioner could be operated more cheaply than by simply using ice? Describe in detail how you evaluate the relative costs.

This page titled 3.4.15: Thermal Physics (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

- 9.15: Thermal Physics (Exercises) by Andrew Park is licensed CC BY 4.0.

2: ASTRONOMY

Chapter 2.1: Introduction to Astronomy

2.1.1: The Basics

2.1.1.1: Module Introduction

2.1.2: Basic Objects in the Universe

2.1.1.3: Scale of the Cosmos

2.1.2: Patterns in the Sky

4.2.1: The Constellations

4.2.2: Other Patterns

Chapter 2.2: Motions of the Moon, Sun, and Stars

2.2.1: The Celestial Sphere

2.2.1.1: Motions of Objects

2.2.1.2: Our Night Sky

2.2.1.3: The Sun and the Moon

2.2.2: Equinoxes, Solstices, and Eclipses

2.2.2.1: Seasons and Seasonal Changes

5.2.2: The Moon in our Skies

2.2.2.3: Eclipses

Chapter 2.3: Light and Spectra

2.3.1: Light

2.3.1.1: Looking at a rainbow, what do you see?

2.3.1.2: Thermal Radiation

2.3.2: Spectra

2.3.2.1: The Electromagnetic Spectrum

2.3.2.2: Spectra

2.3.2.3: Moving Objects, Spectra, and the Doppler Effect

Chapter 2.4: Telescope and Observing

2.4.1: The History of the Telescope

2.4.1.1: The Telescope

2.4.1.2: Early Telescopes

2.4.1.3: Telescopic Observations

2.4.2: The Workings of the Telescope

2.4.2.1: Telescope Optical Types

2.4.2.2: Telescope Mounts

2.4.2.3: Images from the Telescope

2.4.2.4: Light Pollution

2.4.3: Space-based Telescopes
2.4.3.1: Space-based Telescope
2.4.3.2: Hubble Space Telescope

Chapter 2.5: Solar System Formation and Other Stellar Systems
2.5.1: Formation
2.5.1.1: Our Solar System
2.5.1.2: The Nebular Theory
2.5.1.3: What is a Planet?
2.5.1.4: The Nebular Theory - Other Important Evidence
2.5.1.5: The Nebular Theory - Proplyds
2.5.1.6: Systems and Extrasolar Planets
2.5.1.7: The Discovery of Extrasolar Planets
2.5.1.8: The Kepler Mission
2.5.1.9: What We Have Learned about Stellar Systems and Extrasolar Planets?
2.5.1.10: The Circumstellar Habitable Zone
2.5.1.11: Some Interesting Exoplanets
2.5.1.12: Extrasolar Planets
2.5.2: The Rocky Planets
2.5.2.1: What do you think?
2.5.2.2: Comparing Planets
2.5.2.3: Rocky Planets
2.5.2.4: Craters
2.5.2.5: Planet Mercury
2.5.2.6: Planet Venus
2.5.2.7: Planet Earth
2.5.2.8: Earth's Moon
2.5.2.9: Planet Mars
2.5.2.10: Snow on Mars
2.5.2.11: A Martian Controversy
2.5.2.12: Martian Moons
2.5.2.13: Characteristics of the Solar System's Rocky Planets
2.5.3: The Gas Giant Planets
2.5.3.1: What do you think?
2.5.3.2: Gas Giant Planets
2.5.3.3: Rings
2.5.3.4: Planet Jupiter
2.5.3.5: Jupiter's Satellites
2.5.3.6: Planet Saturn
2.5.3.7: Saturn's Rings

2.5.3.8: Saturn's Satellites
2.5.3.9: Planet Uranus
2.5.3.10: Planet Neptune
2.5.3.11: Gas Giant Planets and Select Satellites Overviews

2.5.4: Minor Bodies of the Solar System

2.5.4.1: Trans-Neptunian objects, the Kuiper Belt, and the Oort Cloud
2.5.4.2: The Oort Cloud and Kuiper Belt
2.5.4.3: What do you think?
2.5.4.4: Minor Bodies
2.5.4.5: Asteroids
2.5.4.6: Brightest Asteroids Visible from Earth
2.5.4.7: Asteroid Classification
2.5.4.8: Meteoroids, Meteors, and Meteorites
2.5.4.9: Meteorites
2.5.4.10: The Impact of Meteorites
2.5.4.11: Comets
2.5.4.12: Comet Characteristics
2.5.4.13: Images of Comets
2.5.4.14: Comets in History
2.5.4.15: Scientific Advances
2.5.4.16: Halley's Comet
2.5.4.17: A Comet Impacts a Planet
2.5.4.18: The Centaurs- Are they Asteroids or Comets?
2.5.4.19: Pluto
2.5.4.20: Eris

2.5.5: The Sun

2.5.5.1: What do you think?
2.5.5.2: Fission and Fusion
2.5.5.3: Our Star, the Sun
2.5.5.4: How the Sun Works
2.5.5.5: A "Quick Guide" to Solar Fusion—The Proton-Proton Cycle
2.5.5.6: Solar Regions
2.5.5.7: Major Solar Features
2.5.5.8: Studying the Sun
2.5.5.9: The Dark Side of the Sun
2.5.5.10: The National Solar Observatory
2.5.5.11: NASA's Heliophysics Science Division

Chapter 2.6: Stellar Properties

2.6.1: Stellar Objects

2.6.1.1: Module Introduction
2.6.1.2: Variable Stars
2.6.1.3: Cepheids
2.6.1.4: Cataclysmic Variables
2.6.1.5: Star Clusters
2.6.1.6: Consider this...
2.6.1.7: The Unknown in Astronomy
2.6.1.8: Star and Celestial Object Characteristics

2.6.2: Stellar Distances
2.6.2.1: Distances to the Stars and other Celestial Objects
2.6.2.2: Stellar Distances
2.6.2.3: Everything is Moving
2.6.2.4: How Close are the Close-by Stars?
2.6.2.5: Light-Years
2.6.2.6: Parsec Vs. Light-Year Measurement

2.6.3: Stellar Cartography
2.6.3.1: Stellar and Celestial Object Brightness
2.6.3.2: Magnitude System
2.6.3.3: Celestial cartography
2.6.3.4: Planispheres
2.6.3.5: Star Colors
2.6.3.6: Spectral Type
2.6.3.7: Sample Spectral Types

2.6.4: The Hertzsprung-Russell Diagram
2.6.4.1: The Hertzsprung-Russell Diagram
2.6.4.2: The Four Hertzsprung-Russell Diagram Stellar Groups
2.6.4.3: Types of Star Systems and Stars

Chapter 2.7: Stellar Evolution

2.7.1: Birth
2.7.1.1: Star Life
2.7.1.2: Stellar Birth
2.7.1.3: Stellar Mass
2.7.1.4: Stellar Evolution

2.7.2: Life Cycle - Low-Mass Stars
2.7.2.1: Low-Mass Stars
2.7.2.2: White Dwarf
2.7.2.3: Planetary Nebula
2.7.2.4: White Dwarfs and Neighbors
2.7.2.5: Nova, Novae, and Supernova

2.7.3: Life Cycle - High-Mass Stars
2.7.3.1: High-Mass Stars
2.7.3.2: Neutron stars
2.7.3.3: Neutron Star and Companion Star Scenario
2.7.3.4: High-Mass Star Stellar Endings
2.7.3.5: Gamma-Ray Bursts (GRBs)
2.7.3.6: Checking Out the Theory

Chapter 2.8: Galaxies

2.8.1: The Practice of Studying Galaxies
2.8.1.1: What do you think?
2.8.1.2: Galaxies
2.8.1.3: The History of Studying Galaxies
2.8.1.4: Hubble's Major Contributions
2.8.1.5: Hubble's Law

2.8.2: Galactic Classification
2.8.2.1: Classification of Galaxies
2.8.2.2: Elliptical Galaxies
2.8.2.3: Spiral Galaxies
2.8.2.4: Irregular Galaxies
2.8.2.5: Active Galaxies
2.8.2.6: Quasar-stellar Objects
2.8.2.7: Blazars- Quasar-like Object

2.8.3: Galactic Modern Theory
2.8.3.1: Galaxies, Distance, and Age
2.8.3.2: The Ages of Galaxies and What that Reveals
2.8.3.3: Galactic Groups
2.8.3.4: Dark Matter
2.8.3.5: WIMPs and String Theory
2.8.3.6: Gravitational Lensing

2.8.4: The Milky Way Galaxy
2.8.4.1: What do you think?
2.8.4.2: The Milky Way — Our Home Galaxy
2.8.4.3: Milky Way Galaxy Satellites
2.8.4.4: Milky Way Galaxy Research
2.8.4.5: Galileo Galilei, First to See the Milky Way Galaxy
2.8.4.6: The Milky Way Galaxy's Shape
2.8.4.7: Andromeda Galaxy
2.8.4.8: X-Ray and Gamma-Ray of the Milky Way Galaxy
2.8.4.9: Earth's Position within the Milky Way Galaxy

2.8.4.10: A Spinning Spiral

Chapter 2.9: Cosmology

2.9.1: History of Cosmology

2.9.1.1: What do you think?

2.9.1.2: In the beginning...

2.9.1.3: Physical Cosmology

2.9.1.4: The Big Bang in Eras

2.9.1.5: The Planck Era

2.9.2: "Big Bang" Cosmology

2.9.2.1: Big Bang Evidence

2.9.2.2: Recent Data on the Universe

2.9.2.3: Issues with the Big Bang

2.9.2.4: Beyond the Big Bang

2.9.2.5: What is Next for the Big Bang Theory?

2.9.3: Philosophy of Cosmology

2.9.3.1: Questions Regarding the Eventual Fate of the Universe

2.9.3.2: Religious or Philosophical Cosmology

2.9.3.3: Views on Creation

This page titled 2: Astronomy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 2.1: INTRODUCTION TO ASTRONOMY

Learning Objectives

Upon completion of this module, the student will be able to:

- Describe the characteristics of the major classes of objects in the Universe
- Recognize the scale of the Cosmos, using Astronomical Units and Light travel
- Identify specific Constellations
- Describe the characteristics of specific Constellations
- Identify specific Asterisms
- Differentiate between Constellations and Asterisms

This module starts with the question, “What is Astronomy? ” Simply stated, Astronomy is the study of the Universe. Astronomers classify objects into various groups based on characteristics and locations in our Universe. Some of these classifications, such as stars, are easy to understand scientifically. Other objects are more difficult to classify, and in some instances, to entirely understand. This module identifies basic objects in our Solar System and those beyond, the overall scale of the Universe, and patterns in the night sky. So welcome to your Universe: a beautiful and incredible place, as you will soon learn.

2.1.1: The Basics

2.1.1.1: Module Introduction

2.1.2: Basic Objects in the Universe

2.1.1.3: Scale of the Cosmos

2.1.2: Patterns in the Sky

4.2.1: The Constellations

4.2.2: Other Patterns

This page titled Chapter 2.1: Introduction to Astronomy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 1.1: Module Introduction by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.
- 1: Introduction to Astronomy by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.1.1: THE BASICS

"If you wish to make an apple pie from scratch, you must first invent the universe." — Carl Sagan, *Cosmos*

2.1.1.1: Module Introduction

2.1.2: Basic Objects in the Universe

2.1.1.3: Scale of the Cosmos

This page titled 2.1.1: The Basics is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

2.1.1.1: Module Introduction

“I just looked up at a fine twinkling star and thought that a voyager whom I know now many days sail from this coast, might possibly be looking up at that same star with me. ”

–Henry David Thoreau
Personal Log Entry c. 840’s

Astronomy is one of the most-fascinating subjects; the Universe in which we live is simply a beautiful and incredible place!

As one of the oldest sciences, astronomy is also one of the fastest-progressing sciences today. In centuries past, people used the stars to tell stories through patterns in the sky. They used the positions of the Sun and Moon to tell time and the seasons, and the positions of the stars to navigate.

With the advent space exploration starting in October 1957, satellites have added to our knowledge of the Universe, from the surface of Mars to the beginning of the Universe itself. It hardly seems a week passes without a new discovery, either from earth-based observations or the myriad of satellites.

Images of the Universe, from galaxies containing billions of stars to the rings of Saturn, inspire both the casual and professional stargazer alike. Astronomy is a passion for many and today is a mainstream hobby, as in years past.

Introduction to Astronomy will provide you with an opportunity to explore both the significant historical aspects of astronomy, as well as contemporary exploration of the Universe. Common topics, such as the phases of the Moon, tides, seasons and the space program, will also be explored.

This page titled 2.1.1.1: Module Introduction is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 1.1: Module Introduction by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.1.2: Basic Objects in the Universe

This module starts with the question, “What is Astronomy? ”Simply stated, Astronomy is the study of the Universe. Astronomers classify objects into various groups based on characteristics and locations in our Universe. Some of these classifications, such as stars, are easy to understand scientifically. Other objects are more difficult to classify, and in some instances, to entirely understand. This module identifies basic objects in our Solar System and those beyond, the overall scale of the Universe, and patterns in the night sky.

Stars

Stars are glowing balls of gas that undergo nuclear fusion; the Sun is a star.

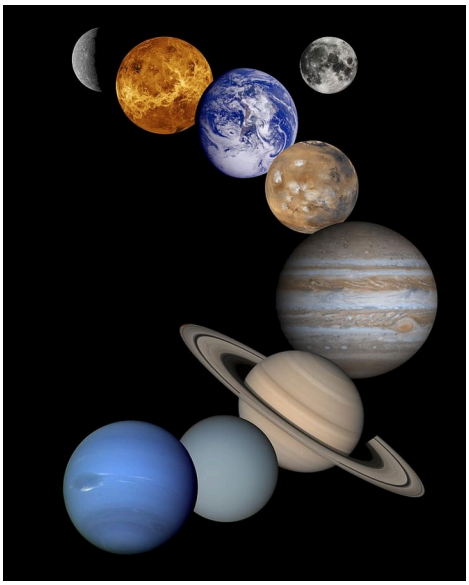


Public Domain | Image courtesy of NASA / ESA.

Planets

Planets are moderately large objects orbiting a star. We see planets because they reflect the light of their central star, or in some cases, stars. Planets are generally rocky or gaseous in nature and spherical-shaped.

A new group of objects has been recently defined: the Dwarf Planets or Plutoids. These are objects that orbit the Sun, but have not cleared their orbits. Pluto is an example of a Dwarf Planet.



Public Domain | Image courtesy of NASA.

Satellite

A **satellite** orbits a planet; these objects are also called moons. For example, the Earth's satellite is the Moon – a proper name.



Public Domain | Image courtesy Pixabay. com.

Asteroid

An **asteroid** is a relatively small, rocky/metallic object *usually* orbiting a star.

Image of the asteroid, Eros.

Public Domain | Image courtesy of NASA.

Comet

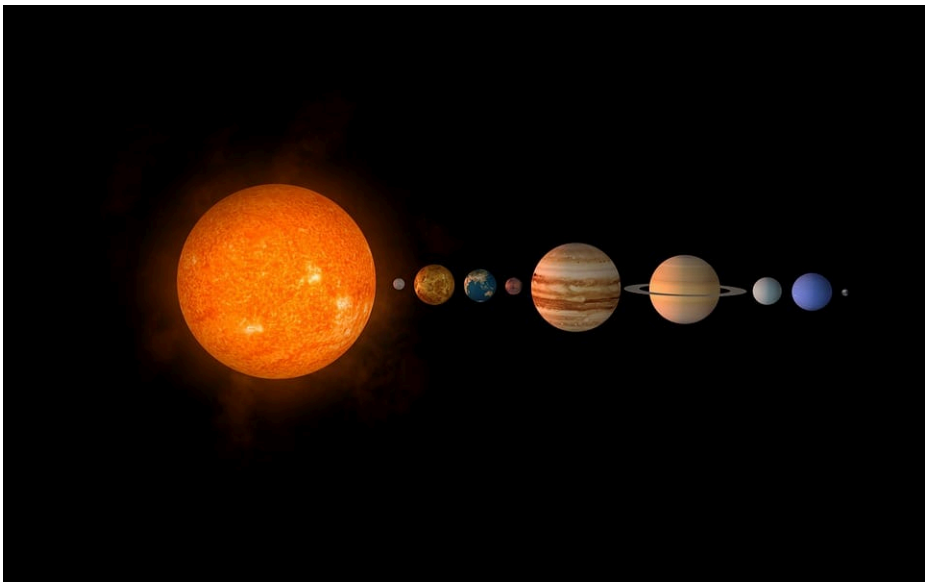
A **comet** is a relatively small, icy object usually orbiting a star. Asteroids, comets, and miscellaneous small/irregular objects and “dust” are often categorized as Minor Bodies.



Public Domain | Image courtesy of NASA.

Solar System

The **Solar System** is the Sun and all the objects that orbit the Sun, including the planets and their moons.



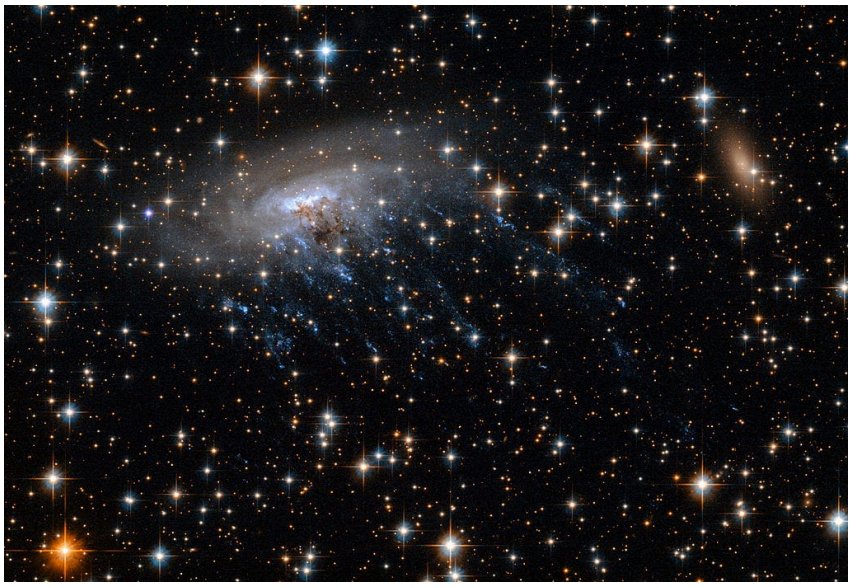
Public Domain | Image courtesy of Pixabay. com.

Stellar System

A **Stellar System** is a star and other objects such as planets and/or other stars and other materials that orbit it.

Galaxy

A **galaxy** is a large island of stars, a few hundred million to over a trillion stars.



CC BY 3. 0 | Image courtesy of ESA/Hubble. NASA, ESA Acknowledgements: Ming Sun (UAH), and Serge Meunier

Galactic Cluster

A **Galactic Cluster** is a collection of galaxies gravitationally bound.

Supercluster

A **Supercluster** is a region where galaxies and galactic clusters are tightly packed.

Universe

The **Universe** is all matter and energy, and is also called the Cosmos.



Public Domain | Image courtesy of Pixabay. com.

This page titled 2.1.2: Basic Objects in the Universe is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 1.2: Basic Objects in the Universe by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.1.1.3: Scale of the Cosmos

Perhaps you have heard the phrase that a large number is **astronomical**. Astronomical comes from the fact that the size of the Universe is so big, it is often hard to comprehend. Even distances between the planets and the Sun in our own Solar System are numbers we are not custom to using. To understand the scale of the Cosmos, we will start with several examples in our Solar System. First, the distance between the Earth and Moon is about 230,000 miles. The closest distance between Earth and Mars is about 32,000,000 miles. The distance between the Earth and Sun is about 93,000,000 miles. The distance between Mercury – the closest planet to the Sun – and the Sun is about 36,000,000 miles. The distance between Neptune – the farthest planet from the Sun – and the Sun is about 2,797,770,000 miles. It is often useful to make a scale model of large systems. Some cities, museums, and parks will create a scale Solar System, based on something like 1 foot equals 1 million miles.

To simplify *distance* in our Solar System, astronomers use the Astronomical Unit (1 AU), which equals 93,000,000 miles or the average distance from Earth to the Sun.

As for other numbers, these are the diameters of three Solar System bodies. First, our Earth's diameter is just less than 8,000 miles. Jupiter's diameter – the largest planet in our Solar System – is around 88,000 miles. And the Sun's diameter is about 850,000 miles.

The idea here is *not* to memorize size and distance; rather, it is to appreciate the grand scale of the Universe within which we live.

Going beyond our Solar System, the closest star is Proxima Centauri, part of the Alpha Centauri triple star system. It is about 24,340,000,000,000 miles (that is 24 trillion, 340 billion miles) away or about 270,000 times more distant than Earth to the Sun. And that is the closest star beyond our Sun. Because of these great distances, astronomers will use another measure: the distance light travels in one year, or the light year. This sounds like a time measure, but it is not.

Light travels 186,000 miles in one second, or 2. 99 x 108 meters per second. That is 5,869,713,600,000 miles in a year. So the Sun is about 8 light minutes from Earth and Proxima Centauri about 4. 24 light years distant.

This page titled 2.1.1.3: Scale of the Cosmos is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 1.3: Scale of the Cosmos by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.1.2: PATTERNS IN THE SKY

4.2.1: The Constellations
4.2.2: Other Patterns

This page titled 2.1.2: Patterns in the Sky is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

4.2.1: The Constellations

Early peoples told stories about the shapes and patterns they saw in the stars, like connect-the-dots. These shapes are called Constellations. Today astronomers recognize 88 “official” Constellations. These are taken from many of the historical constellations. Many of today’s 88 official constellations are Western European in design and history.

Whereas some Constellations are easy to recognize, most are not. Constellations are basically connecting the dots or stars that make up these constellations. To be able to see all 88 Constellations, we would need to be at the equator.



CC BY 3.0 | Image courtesy of Anirban Nandi.

Many peoples throughout civilization identified patterns in the stars. For example, the constellation we now call Orion the Hunter was seen as different shapes by different people around the world, including:

- China: Shen, the supreme warrior
- Egypt: Orion or Osiris in Ancient Egyptian traditions
- India/Hindus: Skanda, a celestial General riding a peacock

This page titled 4.2.1: The Constellations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

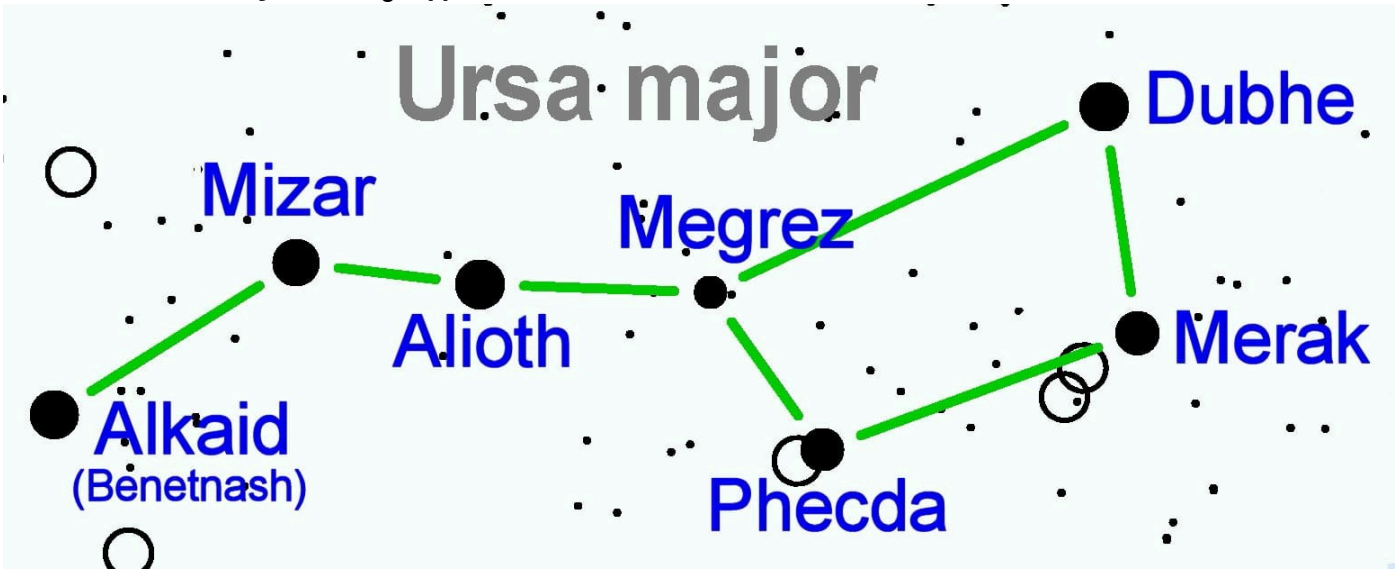
- 1.4: The Constellations by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

4.2.2: Other Patterns

An **Asterism** is a star pattern within a Constellation, but not an official Constellation. These are even more like connecting the dots or stars than the constellations. Some examples of Asterisms within the constellations of Ursa Major and Ursa Minor include the commonly named Big and Little Dippers. In other parts of the world, they have been called the Cart, Plow, and the Drinking Gourd.

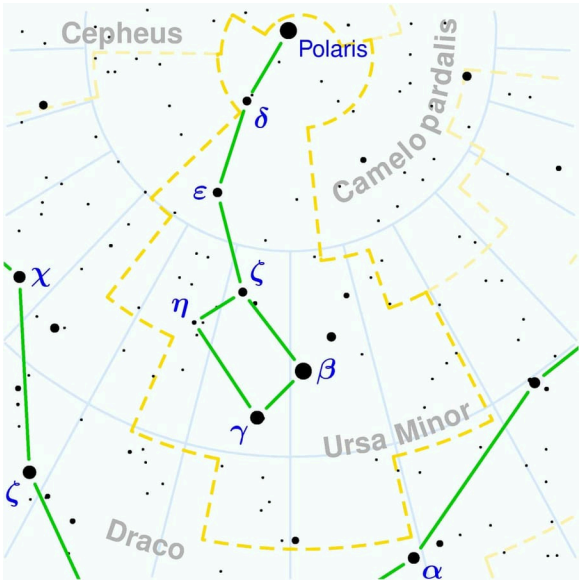
Another example of an Asterism is the Pleiades, within the Constellation Taurus. It is often called the Seven Sisters. In Japan the Pleiades is called Subaru. The Navajo referred to it as Dilyehe, and Hawaiians call the star grouping Makahiki or *many little eyes* .

Constellation Ursa Major—the Big Dipper



GFDL and CC BY-SA 3. 0 | Image courtesy of Luigi Chiesa.

Ursa Minor—the Small Dipper



Public Domain | Image courtesy of Torsten Bronger.

Consider this...

“Ralph Waldo Emerson once asked what we would do if the stars only came out once every thousand years. No one would sleep that night, of course. We would be ecstatic, delirious, made rapturous by the glory of God. Instead the stars come out every night and we watch television. ”

Pawl Hawken (1946 –) Environmentalist, Entrepreneur, and Author

This page titled 4.2.2: Other Patterns is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 1.5: Other Patterns by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

CHAPTER 2.2: MOTIONS OF THE MOON, SUN, AND STARS

The Earth,
Though, in comparison of Heaven, so small
Nor glistering, may of solid good contain
More plenty than the Sun that barren shines.

John Milton,
Paradise Lost, 1665

Learning Objectives

Upon completion of this module, the student will be able to:

- Define motions of objects, including rotation, revolution, and precession
- Identify parts of the points on the Celestial Sphere, such as the celestial poles, meridian, zenith, celestial equator, and ecliptic
- Identify measurement systems used by astronomers
- Explain why we experience seasons, and the descriptions used to discuss seasons
- Explain why the Moon goes through its phases, and the terminology used to explain the phases of the Moon
- Identify types of eclipses
- Explain why eclipses occur

This module will overview movements of Earth, the Moon, and other astronomical bodies and the effects these movements have on what we observe from here on Earth.

2.2.1: The Celestial Sphere

2.2.1.1: Motions of Objects

2.2.1.2: Our Night Sky

2.2.1.3: The Sun and the Moon

2.2.2: Equinoxes, Solstices, and Eclipses

2.2.2.1: Seasons and Seasonal Changes

5.2.2: The Moon in our Skies

2.2.2.3: Eclipses

This page titled Chapter 2.2: Motions of the Moon, Sun, and Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 3: Motions of the Moon, Sun, and Stars by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.2.1: THE CELESTIAL SPHERE

2.2.1.1: Motions of Objects
2.2.1.2: Our Night Sky
2.2.1.3: The Sun and the Moon

This page titled 2.2.1: The Celestial Sphere is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.2.1.1: Motions of Objects



Public Domain | Image courtesy of Pixabay. com.

There are several basic motions of bodies. Rotation is the spinning of an object around its axis. Earth rotates once approximately every 24 hours, whereas Jupiter rotates once approximately every 10 hours.

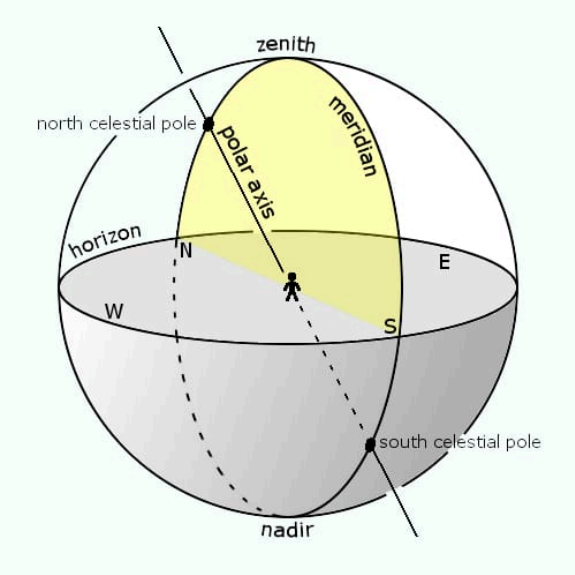
At any moment, half of the Earth is illuminated by the Sun, because of the fact that most of our light comes from the Sun. As the Earth rotates from west to east, your location moves from the dark (night) hemisphere into the illuminated (day) hemisphere and back again. The diurnal (daily) motion of the stars, the Sun, and the Moon is a consequence of Earth's rotation.

Another motion is Revolution; that is, one object orbiting a second. Earth revolves around the Sun approximately every 365. 24 Earth Days (1 Earth year), and Jupiter revolves around the Sun approximately every 11. 86 Earth Years (1 Jovian year). As the Earth revolves or orbits around the Sun, the nighttime side of the Earth (or any other body orbiting the Sun) gradually turns toward different parts of the night sky.

This page titled 2.2.1.1: Motions of Objects is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 3.1: Motions of Objects by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.2.1.2: Our Night Sky



CC BY-SA 3.0 | Image courtesy of Wikimedia Author, Tfr000.

The Celestial Sphere is an imaginary sphere on which all the objects in the sky appear when observed from a specific place, which also appears to move. This sphere appears to surround the Earth. The Sun and Moon, as well as the bright planets, also appear on the Celestial Sphere, moving independently from the background of stars.

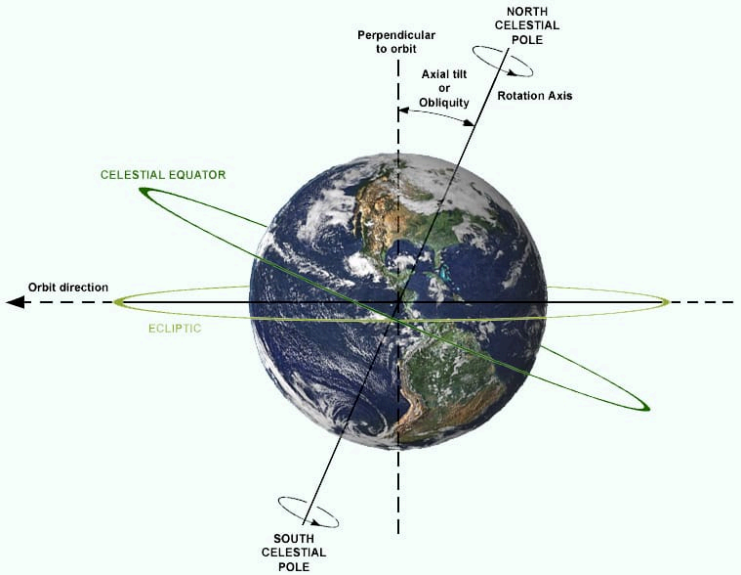
Directions on Earth, in the Universe

On the Earth we use several directional indicators. We use North, East, South, and West. North is 0o, East is 90o, South is 180o, and West is 270o. We use latitude and longitude. Latitude is an imaginary line that runs north or south from the Earth's equator, and longitude runs east or west from the Earth's meridian. We use similar directional and positional indicators for other Solar System bodies.

There are also Celestial Sphere directions. The North and South Celestial Poles are the points at which the Celestial Sphere appears to turn or rotate. This is an extension of the Earth's axis; the stars rotate about these points. Astronomers us Celestial Pole shorthand for these locations

- NCP is North Celestial Pole
- SCP is South Celestial Pole

Points on the Celestial Sphere



CC BY 3.0 | Image courtesy of Author: I, Dennis Nilsson.

There are other points on the Celestial Sphere we see when looking up to the sky. The Local Sky is the sky we see from where we stand. It is called local sky because what one would see in Jacksonville, Florida is different than what one would see in Seattle, Washington and what one would see in Brisbane, Australia.

The Celestial Equator is a projection of Earth's equator onto the Celestial Sphere. The Ecliptic is the Sun's annual path across the Celestial Sphere. Note that the Ecliptic is not the same location as the Celestial Equator due to Earth's tilt of approximately 23.4°. The Zenith is the point directly overhead. And the Meridian is an imaginary line from due north to the zenith to due south. The Meridian divides the sky into halves and can be considered a division between rising objects – those to the east of the meridian, and setting objects – those to the west of the meridian.

Let's go back to the North, East, South, and West system. Generally speaking, objects rise in the East and set in the West. This depends on the object's location on the Celestial Sphere and the observer's location on Earth. For example, if you live at the North or South Poles, the stars will simply circle, neither rising nor setting.

This page titled 2.2.1.2: Our Night Sky is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 3.2: Our Night Sky by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.2.1.3: The Sun and the Moon

The Sun moves along an apparent path on the Celestial Sphere called the Ecliptic. The constellations along Ecliptic – historically 12 but officially 13 – are called the Zodiac. The Moon also travels along the Ecliptic; sometimes in the evening sky, sometimes in the morning, sometimes not visible at all. And the bright planets also travel along the ecliptic.

Measurements along the Celestial Sphere

Astronomers use degrees, minutes, and seconds to measure distances across the sky or sizes. These are referred to as Angular Sizes: the angle the object appears to span or distance between objects, for example:

- The Sun and Moon appear to be about ½ degree as we see them from Earth. This is not their true size; it is just how they appear to us.

  ← → ½ degree = 30 arcminutes

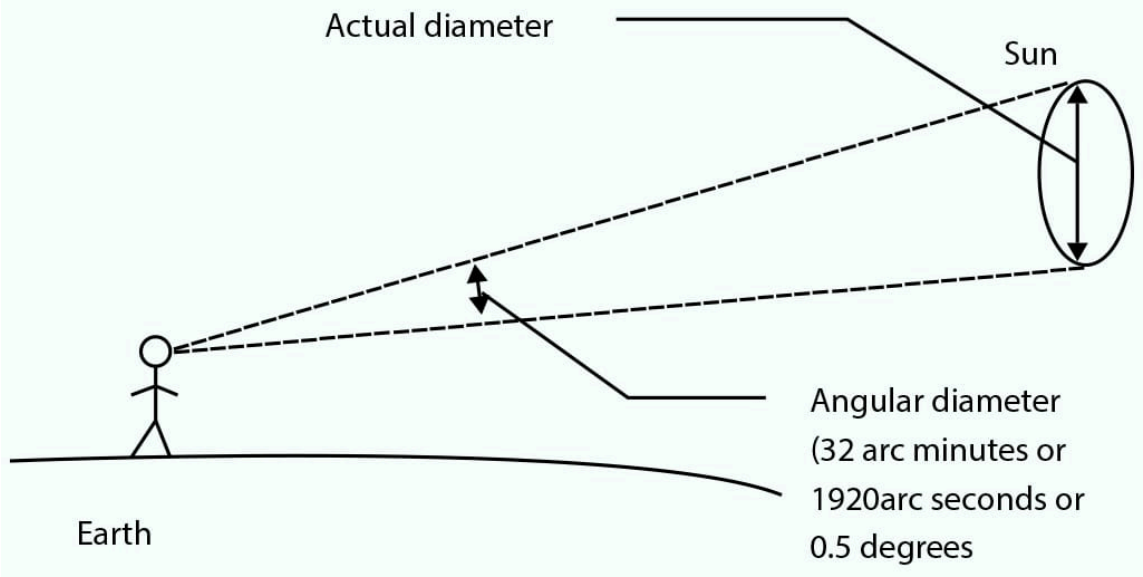


Image courtesy of Florida State College at Jacksonville.

This page titled 2.2.1.3: The Sun and the Moon is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 3.3: The Sun and the Moon by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.2.2: EQUINOXES, SOLSTICES, AND ECLIPSES

For a long as we have been watching the heavens, we have been fascinated by Solar and Lunar Eclipses. From the times of Ancient Man, and even into the modern era, we have seen these events as being ominous and signs of great happenings. No matter how we may feel about them emotionally or spiritually, we cannot deny the inherent beauty contained in the occurrences ranging from the Diamond Ring of a Solar Eclipse going into totality to the Blood Moon of a total Lunar Eclipse.



When a total Solar eclipse nears totality, the one bright spot of the sun still exposed makes the image look like a shiny diamond ring.

Image courtesy Wikipedia, author Tuanna2010. Liscense.



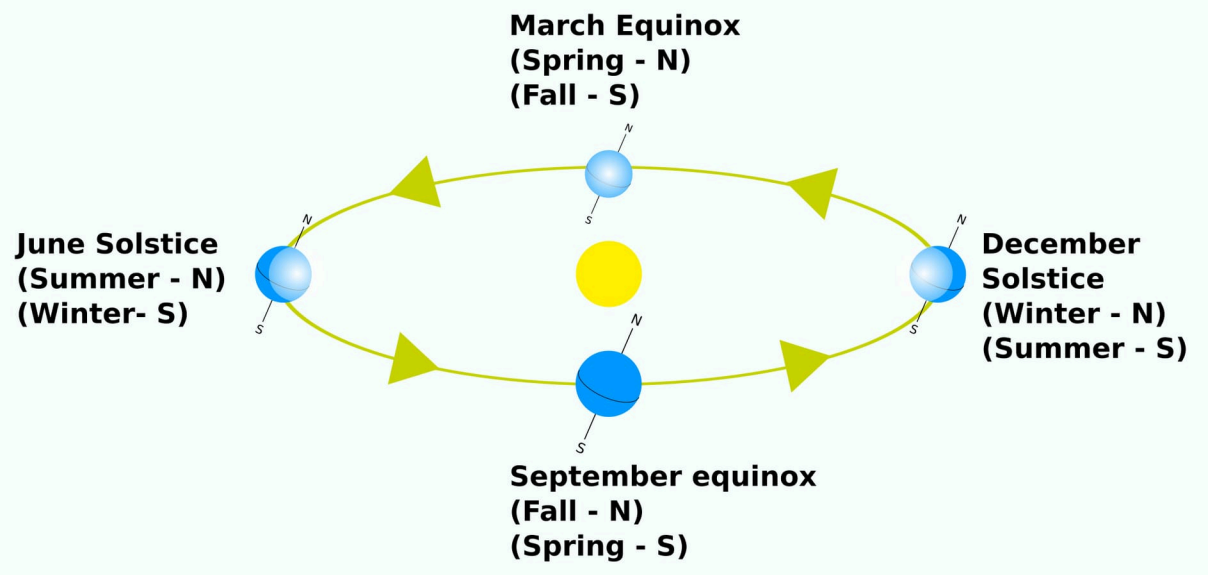
The moon's color resembles that of blood when in its deepest shadow of a total Lunar Eclipse.

Image courtesy of Wikipedia, author Jacek Rużyczka. Liscense.

2.2.2.1: Seasons and Seasonal Changes
5.2.2: The Moon in our Skies
2.2.2.3: Eclipses

This page titled 2.2.2: Equinoxes, Solstices, and Eclipses is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.2.2.1: Seasons and Seasonal Changes



Season of the Year IllustratedPublic Domain | Image courtesy of NASA.

The Seasons are the divisions of a year, which are marked by changes in weather and hours of daylight, and hours of darkness. Seasons occur on a planet or moon because the axis of rotation remains tilted in one direction during its orbit about its star; Earth about the Sun. The tilt is relative to the planet's or satellite's orbit. This tilt results in a planet or moon having Seasons. Earth's axis is tilted about 23 ½o, whereas Uranus' axis is tilted about 98o. Seasons have nothing to do with the distance to the star; for example Earth to the Sun. The distance to the Sun is often incorrectly given as the reason Earth experiences seasons. In fact, Earth is closest to the Sun in January, not June.

Seasonal Terminology include:

- **Summer Solstice** – Sun's rays are most direct
- **Winter's Solstice** – Sun's rays are least direct
- **Spring or Vernal Equinox** – Equal day and night, going from shorter to longer days
- **Fall or Autumnal Equinox** – Equal day and night, going from longer to shorter days

At the Equinoxes, the Sun rises due east and sets due west – only at the Equinoxes. An egg will not stand on end because it's the 1st day of spring (or autumn). At the time of summer solstice the day is longer and the night is shorter. This is extreme at far latitudes and creates what is called the midnight Sun. At the time of winter solstice the day is shorter and the night is longer; E=extreme at far latitudes and results in no Sun during part of the winter season!

Will the Seasons always remain the same?

No, because the orientation of Earth's axis changes over time. This is called Precession, which is the circular motion of a planet's tilted axis and similar to a top's wobble as it slows down. For astronomical bodies, it is a slow process. Earth takes 26,000 years to complete one precession. What are the effects of precession? The effects are the timing of the Seasons and changes in the Celestial poles. Precession is not a perfect path; a wobble in the precessional motion called Nutation causes a small irregularity in the precession.



Image courtesy Mike Reynolds, Ph. D. of Florida State College of Jacksonville.

This page titled 2.2.2.1: Seasons and Seasonal Changes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 3.4: Seasons and Seasonal Changes by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

5.2.2: The Moon in our Skies

It is the very error of the Moon; She comes more near the Earth than she wont, And makes men mad

Shakespeare – Othello

On one side lay the ocean, and on one Lay a great water, and the Moon was full.

Alfred, Lord Tennyson – Morte d'Arthur (1842)

Consider these Moon Facts...

It takes 27 1/3 days for the Moon to go around Earth one time; called the Sidereal Month. From new Moon to the next new Moon takes 29 1/2 days; called the Synodic Month. Why this amount of time? Earth has also moved through space and the Moon has to catch up with that starting position. Think of our word month – from the term Moonth.

As the Moon orbits Earth, roughly the same side of the Moon faces Earth. This means that one (1) lunar rotation equals one (1) lunar revolution. Like Earth, one half (1/2 or 50%) of the Moon is always illuminated by the Sun

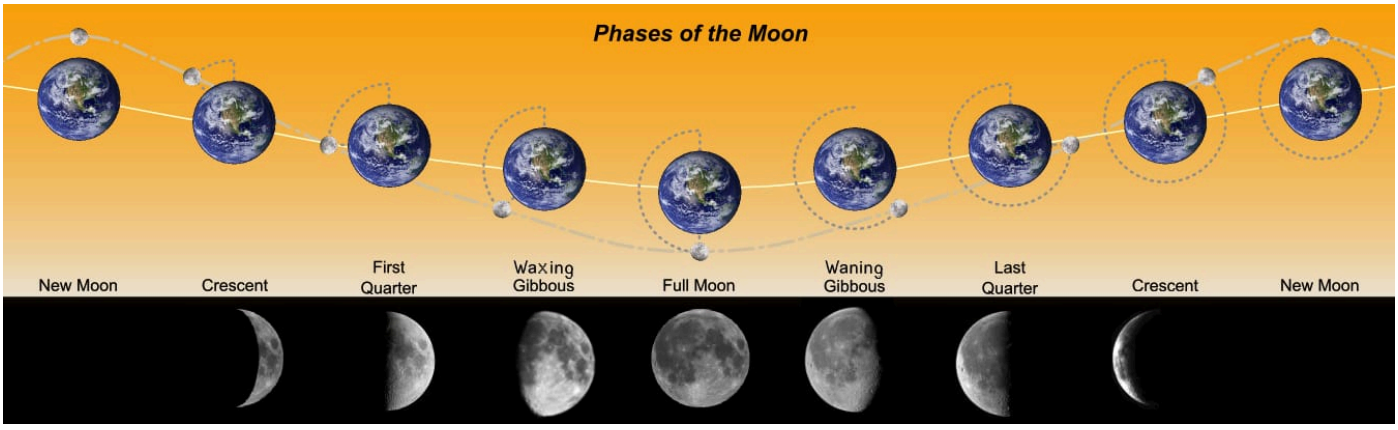
Phases of the Moon

We see the Moon go through its phases due to the position of the Earth, Moon, and Sun relative to each other, NOT due to Clouds, the Moon moving closer or the Sun farther away, or some reflection in the heavens.

Moon Phase Terminology

The descriptions are as the Moon is seen from Earth:

- **Crescent Moon** – Less than 50% illuminated
- **Gibbous Moon** – More than 50% but less than 100%
- **Waxing Moon** – "Growing" larger; New Moon up to Full Moon
- **Waning Moon** – "Growing" smaller; right past Full Moon to New Moon
- **Full Moon** – 100% illuminated- full moons rise at sunset because the Moon and Sun are on opposite sides of Earth
- **New Moon** – 0% illuminated; "no" Moon



Phases of the MoonCC BY-SA 3. 0 | Image courtesy of Wikimedia Author: Fresheneesz~commonswiki

More about the Moon...

Can we see the Moon in the daytime? Yes!

Are there times when we cannot see the Moon at all? Yes, at New Moon and sometimes right before and right after the New Moon because the thin crescent is so difficult to see.

Is there a Young Moon or an Old Moon? These are terms used to describe the Moon's phase right before (old) or after new moon (young). These are very thin crescent phases and can be difficult to see.

Does the Moon keep a constant distance from Earth? No, it varies from a Perigee of about 225,000 miles to an Apogee of about 243,000 miles. Perigee is when the Moon is closest to the Earth and Apogee is when the Moon is farthest from Earth.



Public Domain | Image courtesy of NASA.

Thy shadow, Earth, from Pole to Central Sea, Now steals along upon the Moon's meek shine In even monochrome and curving line Of imperturbable serenity. *Thomas Hardy – At a Lunar Eclipse,1926*

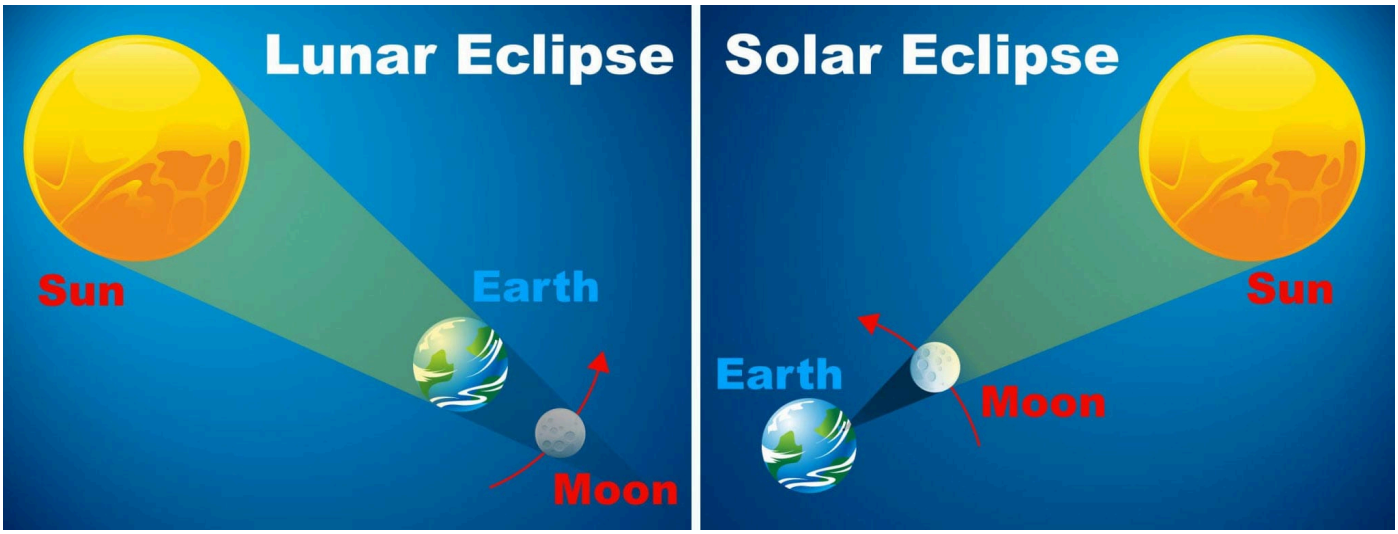
And it shall come to pass in that day, Saith the Lord God, That I will cause the Sun to go down at noon, And I will darken the Earth in a clear day. *Amos 8:9 – The Bible*

This page titled 5.2.2: The Moon in our Skies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 3.5: The Moon in our Skies by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.2.2.3: Eclipses

An eclipse is an event that happens when one body is temporarily hidden, either by passing into the shadow of another body or by having another body pass between it and the observer.



Lunar Eclipses

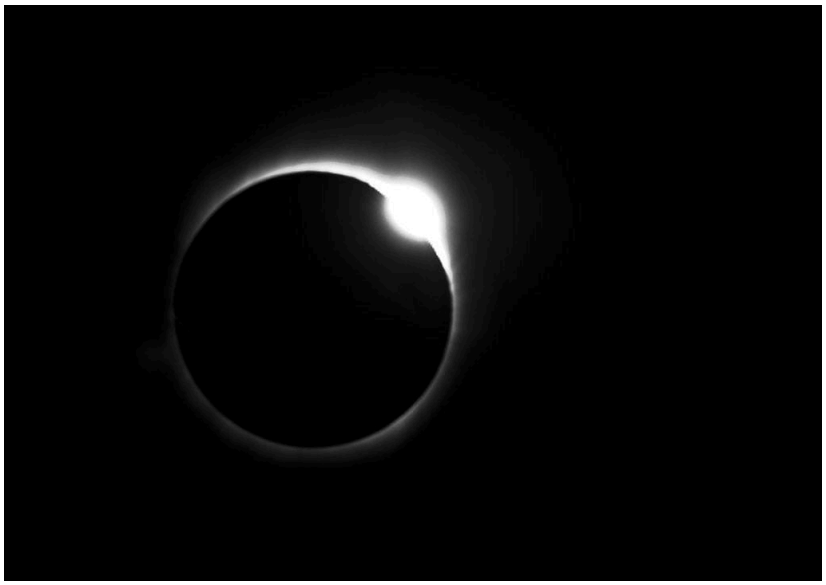
A lunar eclipse occurs when Earth lies directly between the Sun and the Moon, so that Earth's shadow falls on the Moon. A lunar eclipse can only occur at full Moon. There are three types of lunar eclipses: Penumbral, Partial, and Total. The type of lunar eclipse depends on how far the Moon travels into Earth's shadow. The entire night side of Earth can see the lunar eclipse. Total lunar eclipses can vary in color and brightness due to Earth's atmosphere and how deep the Moon passes into Earth's shadow; Red to Copper to Orange to Black. The colors can be quite varied and spectacular.

A Solar Eclipse occurs when the Moon comes between the Earth and the Sun, partially or totally blocking off the Sun from view on Earth. A solar eclipse can only occur at new Moon (whereas a lunar eclipse can only occur at full moon). There are four types of solar eclipses:

1. You must be within the eclipse "zone" to see even a partial solar eclipse.

Total Solar Eclipses

During a total solar eclipse, the Moon totally blocks all harmful solar light and radiation at totality but not during partial. Totality is the phase of the eclipse at which the Sun is completely covered by the Moon. Right before totality, you might be able to see several stars and planets as the sky begins to darken due to decreasing sunlight. At totality, daytime becomes nighttime; the Moon's shadow passes over your site and you will see colors around the horizon, called the sunrise/sunset effect. Occasionally you might see bands of black and white running across the ground right before and after totality; these are called shadow bands and are due to Earth's atmosphere. Animals often react as if it has become nighttime.



CC BY-SA 3.0 | Image courtesy of Wikimedia Author: Lutfar Rahman Nirjhar.

Right at totality and during totality, you will be able to see several things associated with totality.

The **Diamond ring** looks just like its name. It is the last bright sliver of the Sun visible as the Moon covers the Sun, or right after the end of totality as the Sun is being uncovered by the Moon.

Baily's Beads are spots of light that appear as the Moon is covering the Sun. The edge of the Moon is not smooth with mountains and valleys and sunlight goes through these valleys for a few seconds until they too are covered.

At totality, a couple of features become prominent at the now-totally eclipsed Sun. You will see pink-red spots along the edge of the Sun. These are solar prominences and surround the eclipse Sun you will see a whitish extension, the corona or Sun's outer atmosphere.

Eclipse: Lore and Legend

There are some interesting eclipse stories over the passage of civilization. For example, was Stonehenge a system to predict lunar eclipses or something else? In China, the term for an eclipse is chih, which also means to eat; something was eating the Sun. The ancient Chaldeans believed that an eclipse was a display of the Moon's anger. The Babylonians determined the eclipse's "quadrant" – geographical indicator of who would suffer the worst disasters as a result of the eclipse. The Peloponnesian War (5th century BC) was won on the heels of a lunar eclipse; the Athenians failed to retreat and the Syracusians routed them.

Maybe the most interesting and distressing was the story of Columbus and a total lunar eclipse. Columbus used a total lunar eclipse in 1504 to convince the Jamaican natives to supply him and his crew food – he threatened to keep the Moon dark unless the natives met his demands.

Today's World...

In much of the world, it is still common practice to make noise to frighten away whatever is attacking the Sun or Moon. Some peoples are also concerned that diseases are caused by eclipses. Some wells are covered in Japan to prevent celestial poisoning. Some Eskimos turn over utensils to avoid contamination. In India some people lock themselves in their homes to avoid evil rays. And in many parts of the world pregnant women are not allowed outside during an eclipse.

A Total Eclipse of the Sun

People have described total solar eclipses as magical, life-changing, and almost religious in nature...even in today's scientific and high-technology world.

Image of a total eclipse of the sun.

Public Domain

This page titled 2.2.2.3: Eclipses is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 3.6: Eclipses by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

CHAPTER 2.3: LIGHT AND SPECTRA

Red, orange, yellow, green, blue, indigo-violet mystic vision arcs
Red, orange yellow green blue indigo follow graceful violet glows
Lively prism light enchanting rainbow delights brightly dance then hides
Misty promise made sunny days ahead storm's ending I see ROY G BIV

Lorraine Margueritte Gasrel Black
Rainbow in Motion Haiku

Learning Objectives

Upon completion of this module, the student will be able to:

- Describe light and the electromagnetic spectrum
- Recognize Wien's Law and Stefan-Boltzmann Law
- Differentiate among Continuous, Emission Line, and Absorption Line spectrums
- Describe Doppler Effect, both in sound and light
- Identify the astronomical implications of spectral observations

This module covers visible light as well as other forms of the electromagnetic spectrum, which provide us information about the Universe. An object's spectrum can also provide an incredible amount of information, from the object's velocity to its distance.

2.3.1: Light

2.3.1.1: Looking at a rainbow, what do you see?

2.3.1.2: Thermal Radiation

2.3.2: Spectra

2.3.2.1: The Electromagnetic Spectrum

2.3.2.2: Spectra

2.3.2.3: Moving Objects, Spectra, and the Doppler Effect

This page titled Chapter 2.3: Light and Spectra is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 4: Light and Spectra by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.3.1: LIGHT

2.3.1.1: Looking at a rainbow, what do you see?
2.3.1.2: Thermal Radiation

This page titled 2.3.1: Light is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.3.1.1: Looking at a rainbow, what do you see?

The rainbow – with all of its myths and legends – is an indicator of the data astronomers examine.



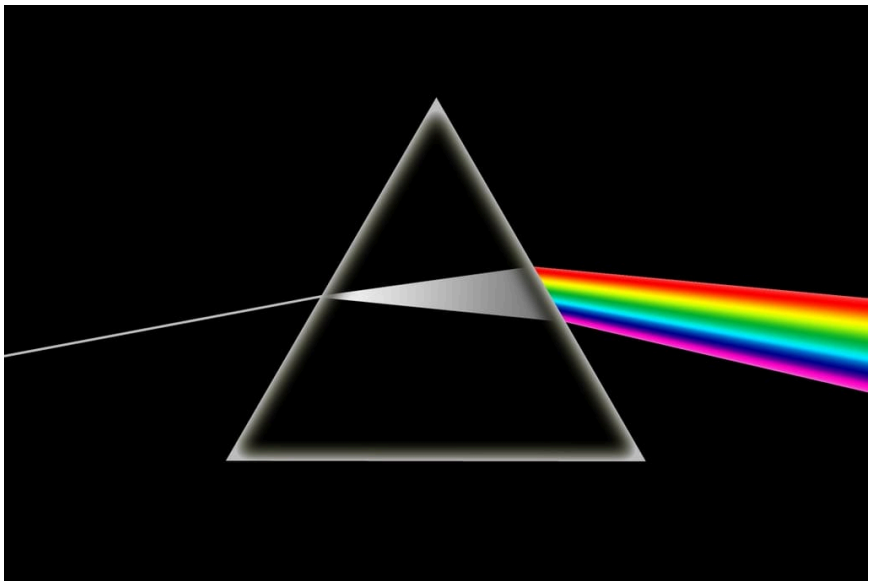
Public Domain | Image courtesy of U. S. National Oceanic and Atmospheric Administration.

Light

Light is **Radiative energy**, which can be split into a **spectrum**. A spectrum is the light's component colors or wavelengths. Light can be split into its component colors or wavelengths by using a prism or diffraction grating.

Light comes in particle bundles, called photons. These photons travel at **186,000 mi/s** or **300,000 km/s (3 X 108 m/s)**. The Speed of Light, c, is an absolute; you cannot go faster than c. (the speed of light).

Light also moves in **waves**. Particles moving in waves are referred to as an **Electromagnetic wave** ; a wave in which both electric and magnetic waves "interact" or vibrate. The **Wavelength** is the distance between adjacent peaks on the wave, and the wave's **Frequency** is the number of peaks that pass a specific point in a second.



CC BY-SA 3.0 | Image courtesy of NASA / ESA.

Image of wavelength over time. Explanation is on the page text.

Image courtesy of Florida State College at Jacksonville.

This page titled 2.3.1.1: Looking at a rainbow, what do you see? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 4.1: Looking at a rainbow, what do you see? by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.3.1.2: Thermal Radiation

An object's **temperature** is the measure of the speed at which the object's particles move. To measure temperature, astronomers prefer the **Kelvin temperature scale**. This scale is much like Celsius, but 0 ^{o} is Absolute Zero – scale starts at zero, 0K. **Wien's Law** relates the color of an incandescent object to its temperature. The law states that wavelength of the brightest color is inversely proportional to the object's temperature.

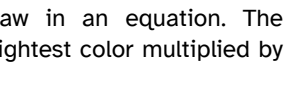
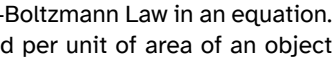


Image of Wien's Law in an equation. The wavelength of the brightest color multiplied by one divided by t.

Where:

- **Stefan-Boltzmann Law** states that the power, **P**, emitted per unit of area of an object is proportional to the fourth power of its temperature, **T**. Mathematically, this is:
  
  Image of Stefan-Boltzmann Law in an equation. Power, P, emitted per unit of area of an object is proportional to the fourth power of its temperature, T
  - **P** is the power
  - **T** is the temperature

This page titled 2.3.1.2: Thermal Radiation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 4.2: Thermal Radiation by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

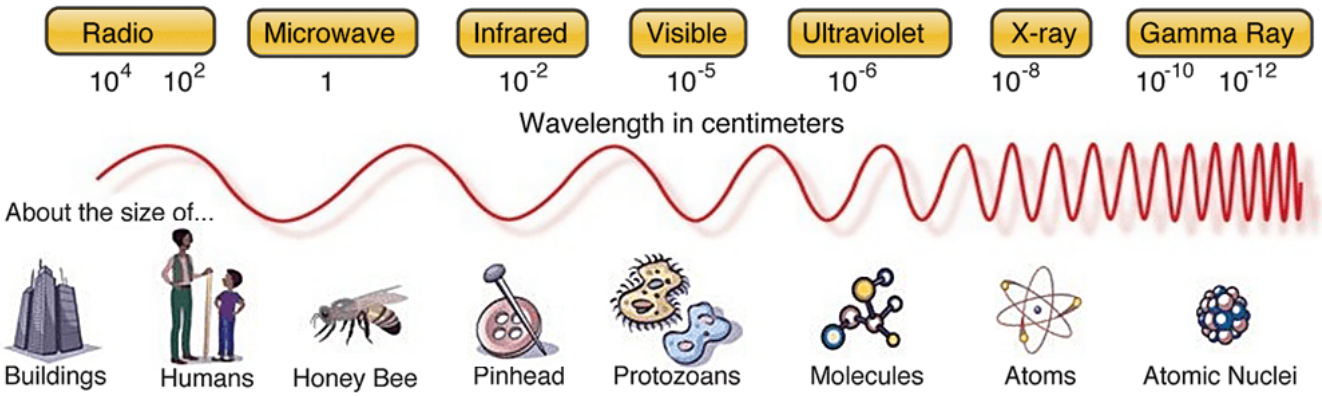
2.3.2: SPECTRA

2.3.2.1: The Electromagnetic Spectrum
2.3.2.2: Spectra
2.3.2.3: Moving Objects, Spectra, and the Doppler Effect

This page titled 2.3.2: Spectra is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.3.2.1: The Electromagnetic Spectrum

The **Electromagnetic Spectrum**, or **EMS**, is the range of frequencies of electromagnetic radiation. Electromagnetic Radiation is a fundamental phenomenon of electromagnetism, acting as waves and as particles – photons, which move through space carrying radiant energy.



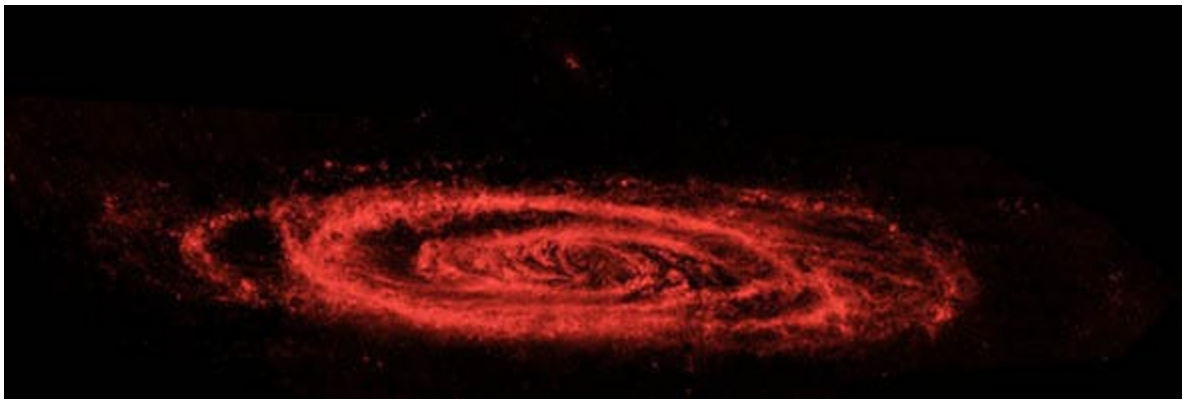
Public Domain | Image courtesy of NASA.

We see the **Visible light** portion of the EMS, wavelengths of 400 nm to 700 nm. The **wavelengths** longer than visible light include Radio, Microwave, and Infrared (IR) waves. **Visible Light** ; sometimes called **ROYGBIV**, includes Red, Orange, Yellow, Green, Blue, Indigo, and Violet. Visible light is a very small part of the EMS. **Shorter Wavelengths** than visible light include Ultraviolet (UV), X-rays, and Gamma Rays. Only radio, some IR, and visible light gets through Earth's atmosphere.

A mnemonic to remember the EMS in correct order from the longer to shorter wavelengths is **R** eal **M** en **I** n **V** irginia **U** se **X** tra **G** lue.



Andromeda Galaxy in Visible LightPublic Domain | Image courtesy of NASA / Hubble Space Telescope.



Andromeda Galaxy in IRPublic Domain | Image courtesy of NASA / Spitzer Space Telescope.

This page titled 2.3.2.1: The Electromagnetic Spectrum is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 4.3: The Electromagnetic Spectrum by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.3.2.2: Spectra

Spectra are very important to astronomy because they can show the component wavelengths and specific elements present.

Continuous Spectrum

Rectangular image of a continuous color spectrum from black to purple to blue to green to yellow to orange to red to black.

Public Domain | Image courtesy of Wikimedia Author: Jhausauer.

A continuous spectrum shows all wavelengths and is produced by hot, dense solid objects at high temperatures.

Emission Line Spectrum



Public Domain | Image courtesy of Wikimedia Author: Jhausauer.

An Emission Line spectrum produces a spectrum with **bright lines**. It occurs from heating an object, causing the electrons in the object's atoms/molecules to show at a specific wavelength.

Absorption Line Spectrum



Public Domain | Image courtesy of Wikimedia Author: Jhausauer.

An Absorption Line spectrum produces a spectrum with **dark lines**. It occurs when one object absorbs specific photons given off by a second object. For example, a light behind a cloud of gas absorbs photons at the cloud's energy level while the rest of the photons pass through the cloud.

This page titled 2.3.2.2: Spectra is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

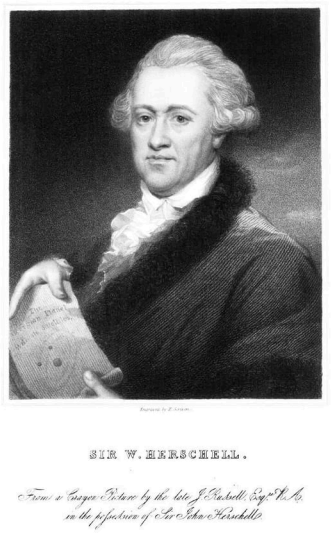
- 4.4: Spectra by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.3.2.3: Moving Objects, Spectra, and the Doppler Effect

The **Doppler Effect** is a change in an object's spectrum because the object is moving closer or away from the observer. In **sound** we experience this change in a siren's pitch. When it approaches us, the waves are being compressed, then becoming longer as it moves away. This increase in pitch can be heard as a police car approaches and then decreases as it moves farther away. **What are the Implications of Doppler Shift?**

1. **What are The Astronomical Implications and Importance of Spectra?**
    1. How do we know what a star is made up of?
        - *From the star's spectra*
    2. How do we know if an object, star, galaxy, etc., is moving towards or away from us?
        - *Look for a red or blue shift in the object's spectra*
    3. Can we tell how fast an object is moving towards or away from us?
        - *Yes, by the extent of red or blue shift*

Consider this...



Astronomer Frederick William *Herschel*Public Domain

For most of history, visible light was the single most recognized portion of the electromagnetic spectrum, EMS. The ancient Greeks noted that light traveled in straight lines and studied its properties. Over the years the study of light continued. During the 16th and 17th centuries there were conflicting theories which regarded light as either a wave or a particle. In 1800, astronomer Frederick William Herschel discovered infrared radiation while using a thermometer to measure temperatures of visible light frequencies. A year later Johann Ritter discovered ultraviolet radiation.

This page titled 2.3.2.3: Moving Objects, Spectra, and the Doppler Effect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 4.5: Moving Objects, Spectra, and the Doppler Effect by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

CHAPTER 2.4: TELESCOPE AND OBSERVING

He burned his house down for the fire insurance
And spent the proceeds on a telescope
To satisfy a lifelong curiosity
About our place among the infinities

The strongest thing that's given us to see with's
A telescope. Someone in every town
Seems to me owes it to the town to keep one.
In Littleton it may as well be me.

Robert Frost
The Star-Splitter

This module introduces you to the primary instrument used by astronomers: the telescope, as well as how the telescope works. From its historical first use to today's modern instruments, the telescope changed forever how we look at our Universe.

Learning Objectives

Upon completion of this module, the student will be able to:

- Identify the types of telescopes and telescope mounts
- Describe the differences between the refracting telescope and reflecting telescope, including both positive and negative features
- Describe the types of instruments astronomers use with the telescope and the function of each type
- Explain why space-based telescopes are of such importance to today's astronomers and the collection of information

2.4.1: The History of the Telescope

2.4.1.1: The Telescope

2.4.1.2: Early Telescopes

2.4.1.3: Telescopic Observations

2.4.2: The Workings of the Telescope

2.4.2.1: Telescope Optical Types

2.4.2.2: Telescope Mounts

2.4.2.3: Images from the Telescope

2.4.2.4: Light Pollution

2.4.3: Space-based Telescopes

2.4.3.1: Space-based Telescope

2.4.3.2: Hubble Space Telescope

This page titled Chapter 2.4: Telescope and Observing is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 5: Telescope and Observing by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.4.1: THE HISTORY OF THE TELESCOPE

The world of astronomy, which, at the time, was nearly indistinguishable from astrology, hasn't been the same since the invention of the telescope.



Galilieo's brass telescope on its stand. It is a piece of art as much as a scientific instrucment.

Image courtesy Wikipedia. Author אהאמא. Liscense.

2.4.1.1: The Telescope

2.4.1.2: Early Telescopes

2.4.1.3: Telescopic Observations

This page titled 2.4.1: The History of the Telescope is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.4.1.1: The Telescope

*"The early spectacle maker – today's equivalent of an optometrist – set up shop in the open streets. The patient would try on different 'spectacles' to see which one provided the best improvement in vision. "*The **telescope**. is an instrument designed to **collect and magnify visible light** and **see detail**. Detail refers to **angular resolution** ; think of resolving into finer details. Bigger telescopes collect more light and produce better resolution. However, the bigger the better is not always true; our atmosphere can limit a telescope's resolving abilities.



CC BY-SA 4. 0 | Image courtesy of Wikimedia Author: H. Aziz KAYIHAN.

Before the Telescope

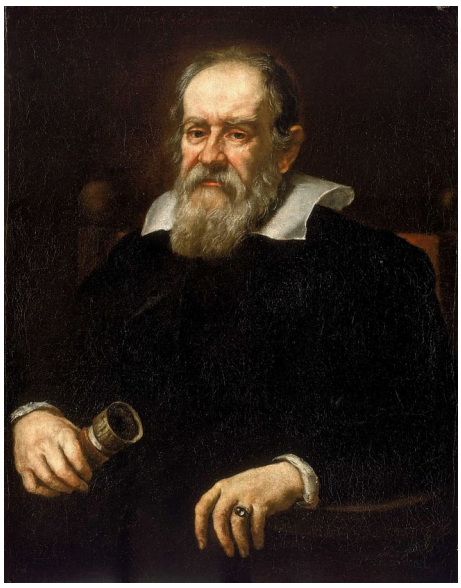
Specific discoveries led up to the invention of the telescope. Below they are divided in Year and Event.

- There are several basic terms we use when discussing telescopes. **Optics** is the science that deals with all aspects of visible light. **Primary Objective** refers to the main lens (refractor telescopes) or mirror (reflector telescopes) which gathers the incoming light. The **Eyepiece** is a lens or series of lenses which focuses the light from the telescope's primary objective for the eye. **Refraction** is the bending of waves, such as light, when it passes from one substance to another, for example, from air through glass. Refraction is the primary method of a refracting telescope. Reflection is the bending of waves, such as light or sound waves, from a surface. **Reflection** is the primary method of a reflecting telescope.

This page titled 2.4.1.1: The Telescope is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 5.1: The Telescope by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.4.1.2: Early Telescopes



Public Domain

Jan (or Han) **Lipperhey** of Holland, a spectacle maker, is given credit for inventing the first telescope in 1608 — a **refracting telescope**. He called his telescope **Kijker**, meaning "looker" in Dutch. Lipperhey's telescope was a refracting or lens-type telescope – as were all early telescopes; this used a convex lens to focus incoming light. Lipperhey thought the telescope's best use was*military* .

Lipperhey unsuccessfully tried to patent the telescope, as well as telescope binoculars. Since Lipperhey was denied a patent – probably meaning he did not invent the telescope – who did invent the telescope? Rumor has it Lipperhey's son actually discovered the correct lens combination – *he put together two lenses and "spied" a local church steeple* . Some historians argue that Lipperhey stole the idea from **Hans and Zacharias Janssen**, other spectacle makers, who supposedly built a telescope in 1595. However, historical references are not as clear on the Janssens and their telescope.

The first to use the telescope astronomically was **Galileo Galilei**. Galileo made his first telescope in 1609, based on what he had heard about the Lipperhey telescope. However Galileo's telescope was of much-better quality than Lipperhey's telescope.

"...[But] perhaps his [Galeleo] most famous invention was the telescope. Galileo made his first telescope in 1609, modeled after telescopes produced in other parts of Europe that could magnify objects three times. He created a telescope later that same year that could magnify objects twenty times. With this telescope, he was able to look at the moon, discover the four satellites of Jupiter, observe a supernova, verify the phases of Venue, and discover sunsports. His discoveries proved the Copernican system which sates that the earth and other planets revolve around the sun. Prior to the Copernican system, it was held that the universe was geocentric, meaning the sun revolved around the earth"

— The Galileo Project (1564-1642)

This page titled 2.4.1.2: Early Telescopes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 5.2: Early Telescopes by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.4.1.3: Telescopic Observations

One of Galileo's first recorded telescopic observations was of Jupiter and "three fixed stars, totally invisible by their smallness"—three of the four large moons of Jupiter. Wne now call the four large moons of Jupiter the Galilean Satellites or moons. In 1668, Sir Isaac Newton designed and built the first Reflecting Telescope. It was Galileo who first described the idea for a reflecting-type telescope. The Reflecting Telescope used a concave mirror in the place of a lens to focus incoming light.

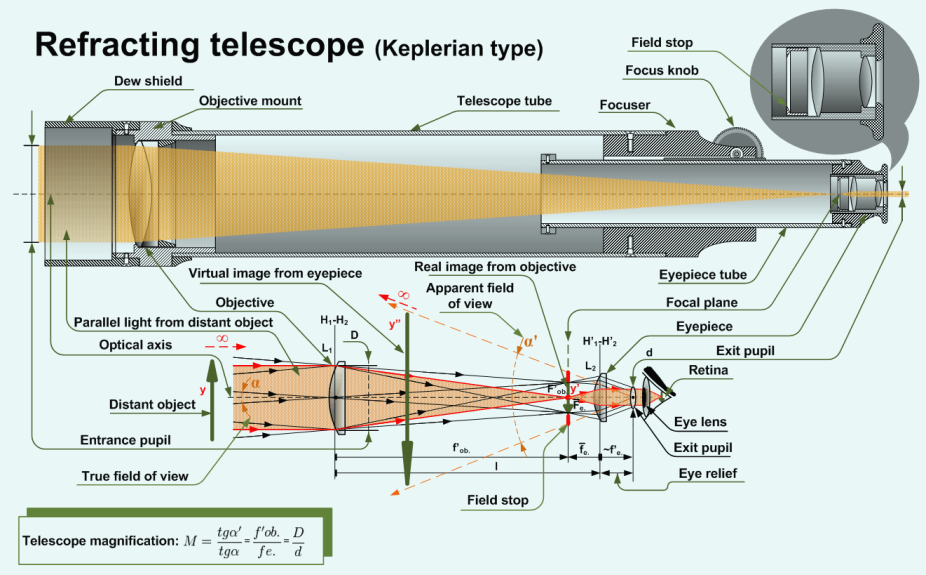
NASA's Juno spacecraft obtained this color view on June 21, 2016, at a distance of 6. 8 million miles (10. 9 million kilometers) from Jupiter. Juno will arrive at Jupiter on July 4. As Juno makes its initial approach, the giant planet's four largest moons -- Io, Europa, Ganymede and Callisto -- are visible.

Public Domain | Image courtesy of NASA.

This page titled 2.4.1.3: Telescopic Observations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.4.2: THE WORKINGS OF THE TELESCOPE

The inner workings of a telescope can range from relatively simple to extremely complex.



The inner workings and optics of a refracting telescope.
Notice that the overall length of the scope, I, is equal to the sum of the two lens' focal lengths.

Image courtesy of Wikipedia. Author: Tamasflex. License.

2.4.2.1: Telescope Optical Types

2.4.2.2: Telescope Mounts

2.4.2.3: Images from the Telescope

2.4.2.4: Light Pollution

This page titled 2.4.2: The Workings of the Telescope is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.4.2.1: Telescope Optical Types

The type of telescope primarily depends on the optical combinations used to collect the incoming light or Electromagnetic Radiation (EMR). The **Refracting Telescope** or **Refractor** uses a lens or lenses as the telescope's Primary Objective. **Binoculars** are a type of Refractor; occasionally you will find reflecting binoculars.

Graphic of Refractor

Graphic of Refractor. Text on screen contains description.

The Achromatic RefractorImage courtesy Mike Reynolds, Ph. D. of Florida State College of Jacksonville.

The **Reflecting Telescope** or **Reflector** uses a concave mirror as the telescope's Primary Objective, rather than a lens or lenses. The type of reflector depends on other system mirror(s), called the Secondary Mirror.

A **Compound** or **Catadioptric Telescope** uses a combination of Refractor and Reflector characteristics. **Radio Telescopes** are attuned to the radio end of the spectrum; it turns out that many objects produce a radio emission.

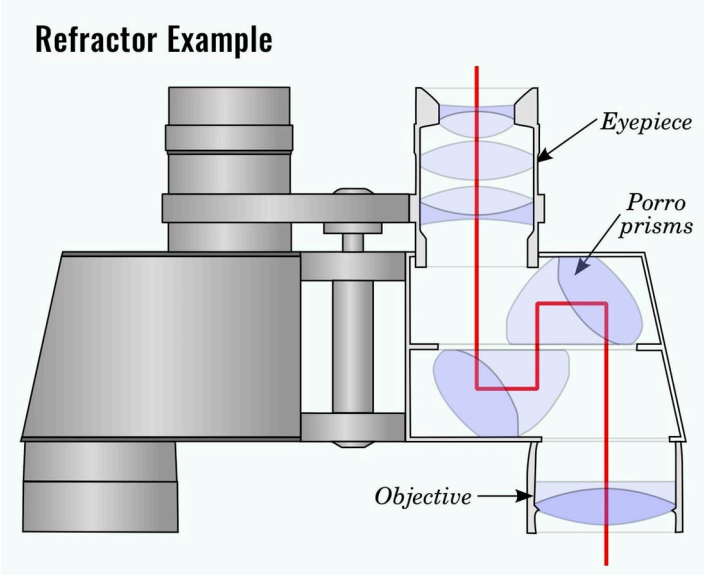
Specific Wavelength-Sensitive Telescopes are those attuned to specific regions of the spectrum; such as infrared (IR), ultraviolet (UV), and microwave.

Gamma-Ray Telescopes are used for detecting the gamma-ray part of the Electromagnetic Spectrum. These can detect very high-energy gamma-rays from radio galaxies, gamma-ray bursts from stars, supernovae events, and Quasars.

The refractor works by two lenses first gathering and directing the light (right); Galileo's refractor only used one lens. The eyepiece on the left focuses the light for the eye (or camera, spectrometer. etc.)

The binocular primarily uses refractor-like qualities with prisms to upright and fold the image.

Refractor Example



CC BY-SA 3. 0 | Image courtesy of Wikimedia Author: Turnvater Jahn.

The Newtonian reflector works first by light entering the telescope at the right and traveling to a concave mirror. Graphics of a concave and convex mirror are shown as reference points.

The light is then reflected back to a smaller, second flat mirror and reflected out of the side of the telescope to the eyepiece and the eye.

The Newtonian reflector works first by light entering the telescope at the right and traveling to a concave mirror. Graphics of a concave and convex mirror are shown as reference points.

The light is then reflected back to a smaller, second flat mirror and reflected out of the side of the telescope to the eyepiece and the eye.

Graphic of Newtonian Reflector

Graphic of Newtonian Reflector. Text on screen contains description.

The Newtonian ReflectorImage courtesy Mike Reynolds, Ph. D. of Florida State College of Jacksonville.

Reflectors use a concave mirror as its primary objective to focus the incoming light (same optical focusing effect as a convex lens). The mirror is coated on the surface of the glass, called a First Surface Mirror; the coating is usually molecules-thick coating of Aluminum or Silver. Household mirrors are coated on the back of the glass, called a Second Surface Mirror.

Reflector telescope types differ based on what happens to the light after it is reflected from the primary.

Graphic of Newtonian Reflector

Graphic of Newtonian Reflector. Text on screen contains description.

The Compound ReflectorImage courtesy Mike Reynolds, Ph. D. of Florida State College of Jacksonville.

Light enters the telescope at the right and travels to a concave mirror. The light is then reflected back to a smaller, second convex mirror. The light is then reflected back again out of the back of the telescope to the eyepiece and the eye.

Like the Reflector, uses a concave mirror as its primary objective to focus the incoming light. A corrector plate is placed at the front of the tube; it corrects for primary mirror's shortcomings. Usually compound telescopes have a shorter tube than comparable reflectors; yet they are more expensive inch-per-inch than ian reflectors.

The Compound Reflector, or Schmidt-Cassegrai, is also Called Catadioptric Telescopes.

Refractor versus Reflector

Refractor Telescope

Positive

- *Negative*
  - *Positive*
    - *Negative*
      - *Note: The refractor's positive features are the reflector's negative features and vice-versa.*



Public Domain | Image courtesy of NASA.



CC BY 3.0 | Image courtesy of Wikimedia Author: CSIRO.

This page titled 2.4.2.1: Telescope Optical Types is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 5.4: Telescope Optical Types by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.4.2.2: Telescope Mounts

CC BY 2. 5 | Image courtesy of Wikimedia Author: Marie-Lan Nguyen.

Telescope Mount is the telescope's support structure. The two main types of telescope mounts are alt-azimuth and equatorial mounts. The **Alt-azimuth** mounts move in two directions: left to right, or up and down. The word "alt-azimuth" is a combination of altitude and azimuth. **Equatorial** mounts track the apparent motion of the stars by aligning one of its axes parallel to Earth's axis.

Go-To Telescope Mounts

Go-To's telescope mounts enable the telescope to find objects in the sky using a microprocessor, encoders and motors. A Global Positioning System (GPS) as a part of the Go-To system tells the mount and microprocessor where the telescope is located, what objects are visible, and where they are in the sky.

This page titled 2.4.2.2: Telescope Mounts is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 5.5: Telescope Mounts by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.4.2.3: Images from the Telescope

Once the image is focused and magnified by the telescope, the image can be viewed several ways. **Visually** the image can be viewed by the eye. The image can be captured by a film camera (mostly historical), digital camera or video, or Charge-Couple Device (CCD); the telescope becomes the camera lens; **Digital or Photographic astrophotography**.



CC BY 2. 0 | Image courtesy of Wikimedia Author: Roland Tanglao.

A spectroscope breaks the light of the incoming image into its component wavelengths for study, called **spectroscopy**. And a **photometer** measures the amount of incoming light.



CC BY 3. 0 | Image courtesy of Wikimedia Author: Reptonix.

This page titled 2.4.2.3: Images from the Telescope is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 5.6: Images from the Telescope by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.4.2.4: Light Pollution



Public Domain | Image courtesy of Pixabay. com.

One of the biggest challenges to Earth-based telescopes is that of light pollution. Lighting interferes with the dim objects astronomers need to study. Consider how many people have actually seen our Milky Way arching across the sky. *Have you?*

Light pollution has other serious side effects. Environmentally it disrupts natural day and night patterns for many plants and animals. There are severe beach restrictions in regard to lighting in many areas due to nesting sea turtles. Another consequence of light pollution is the cost.

A number of solutions are possible. First, good night lighting is shielded; that is, directed only towards the ground. Why have a streetlight that shines in one's eyes? Second is the type of bulb used in night lighting. It has been shown that sodium vapor lamps more-efficiently light an area; both cost- and illumination-wise. As Light-emitting diode (LED) lamps are developed, better lighting solutions will follow.

This page titled 2.4.2.4: Light Pollution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 5.8: Light Pollution by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.4.3: SPACE-BASED TELESCOPES

There are great advantages to having telescopes in space, though there are limitations, too. The most notable advantage is not having to deal with all the issues the atmosphere introduces into observing. The greatest drawback is the remoteness and expense of being in, and getting into, orbit.



The James Webb Space Telescope being built. Notice that all the workers are in "clean suits". This is to keep the inner workings as debris-free as possible. Even a spec of dust could be disastrous and, when in orbit, this presents a real repair issue.
Image courtesy of Wikipedia. Author: Northrop Grumman. Liscense.



Star nursery as imaged by the James Web Space Telescope. In this image, there is the first evidence of Brown Dwarfs.

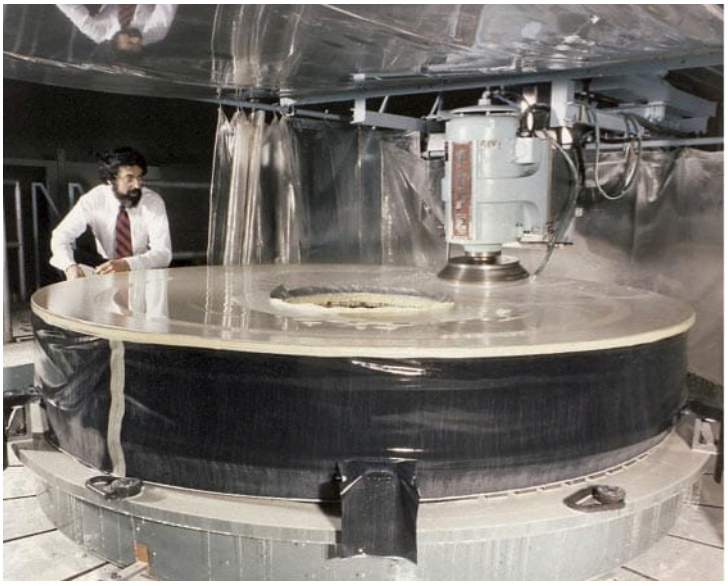
Image courtesy Wikipedia. Author: NASA's James Webb Space Telescope. Liscense.

2.4.3.1: Space-based Telescope
2.4.3.2: Hubble Space Telescope

This page titled 2.4.3: Space-based Telescopes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.4.3.1: Space-based Telescope

The idea of a space-based telescope can be traced back to the German rocket scientist Hermann Oberth, who noted in his 1923 work *Die Rakete zu den Planetenräumen (The Rocket into Planetary Space)* , how a telescope could be propelled into Earth orbit by a rocket. It was astronomer Lyman Spitzer, whose 1946 paper *Astronomical advantages of an extraterrestrial observatory* , became the basis for the project. Through the 1960s, reports and recommendations were prepared to support a Large Space Telescope, as the project was initially called. Congress approved funding for the Telescope in 1978. The Hubble Space Telescope (HST) as it was renamed to honor astronomer Edwin Hubble, was launched April 24, 1990, after numerous design and construction delays, along with flight delay due to the space shuttle Challenger accident.



Grinding of the Hubble Space Telescope Mirror, 1979Public Domain | Image courtesy of NASA.

This page titled 2.4.3.1: Space-based Telescope is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.4.3.2: Hubble Space Telescope

Within weeks of launch, the HST showed serious problems with its primary mirror, which affected the precise focus of the telescope. A fix was designed and astronauts about Space Shuttle Endeavor in December 1993 captured and repaired the near-sighted telescope. The repair was somewhat simple though—to equip the HST with glasses. The fix worked and images returned from the Hubble Space Telescope have amazed and intrigued both astronomers and the general public ever since.



Release of the Hubble Space Telescope from Space Shuttle Endeavour, 1990Public Domain | Image courtesy of NASA.



Blast-off of the space shuttle Endeavour and the Hubble Space Telescope, April 24, 1990Public Domain | Image courtesy of NASA.

This page titled 2.4.3.2: Hubble Space Telescope is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 5.10: Hubble Space Telescope by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

CHAPTER 2.5: SOLAR SYSTEM FORMATION AND OTHER STELLAR SYSTEMS

This module introduces you to our solar system and how astronomers believe it formed, based on what we see around us and around other star systems in formation. We will also look at the discovery of other planets around other stars–called *exoplanets*.

Learning Objectives

Upon completion of this module, the student will be able to

- Explain the nebular theory of the solar system formation
- Describe the evidence which supports the nebular theory
- List the IAU's criteria to be a planet, and how this changed the status of Pluto
- Explain what an exoplanet is, as well as the history of their discoveries
- List examples of unusual stellar systems
- Explain brown dwarfs and how they differ from stars and planets
- Discuss stellar system habitable zones and what that has to do with life

2.5.1: Formation

2.5.1.1: Our Solar System

2.5.1.2: The Nebular Theory

2.5.1.3: What is a Planet?

2.5.1.4: The Nebular Theory - Other Important Evidence

2.5.1.5: The Nebular Theory - Proplyds

2.5.1.6: Systems and Extrasolar Planets

2.5.1.7: The Discovery of Extrasolar Planets

2.5.1.8: The Kepler Mission

2.5.1.9: What We Have Learned about Stellar Systems and Extrasolar Planets?

2.5.1.10: The Circumstellar Habitable Zone

2.5.1.11: Some Interesting Exoplanets

2.5.1.12: Extrasolar Planets

2.5.2: The Rocky Planets

2.5.2.1: What do you think?

2.5.2.2: Comparing Planets

2.5.2.3: Rocky Planets

2.5.2.4: Craters

2.5.2.5: Planet Mercury

2.5.2.6: Planet Venus

2.5.2.7: Planet Earth

2.5.2.8: Earth's Moon

2.5.2.9: Planet Mars

2.5.2.10: Snow on Mars

2.5.2.11: A Martian Controversy
2.5.2.12: Martian Moons
2.5.2.13: Characteristics of the Solar System's Rocky Planets

2.5.3: The Gas Giant Planets

2.5.3.1: What do you think?
2.5.3.2: Gas Giant Planets
2.5.3.3: Rings
2.5.3.4: Planet Jupiter
2.5.3.5: Jupiter's Satellites
2.5.3.6: Planet Saturn
2.5.3.7: Saturn's Rings
2.5.3.8: Saturn's Satellites
2.5.3.9: Planet Uranus
2.5.3.10: Planet Neptune
2.5.3.11: Gas Giant Planets and Select Satellites Overviews

2.5.4: Minor Bodies of the Solar System

2.5.4.1: Trans-Neptunian objects, the Kuiper Belt, and the Oort Cloud
2.5.4.2: The Oort Cloud and Kuiper Belt
2.5.4.3: What do you think?
2.5.4.4: Minor Bodies
2.5.4.5: Asteroids
2.5.4.6: Brightest Asteroids Visible from Earth
2.5.4.7: Asteroid Classification
2.5.4.8: Meteoroids, Meteors, and Meteorites
2.5.4.9: Meteorites
2.5.4.10: The Impact of Meteorites
2.5.4.11: Comets
2.5.4.12: Comet Characteristics
2.5.4.13: Images of Comets
2.5.4.14: Comets in History
2.5.4.15: Scientific Advances
2.5.4.16: Halley's Comet
2.5.4.17: A Comet Impacts a Planet
2.5.4.18: The Centaurs- Are they Asteroids or Comets?
2.5.4.19: Pluto
2.5.4.20: Eris

2.5.5: The Sun

2.5.5.1: What do you think?
2.5.5.2: Fission and Fusion
2.5.5.3: Our Star, the Sun

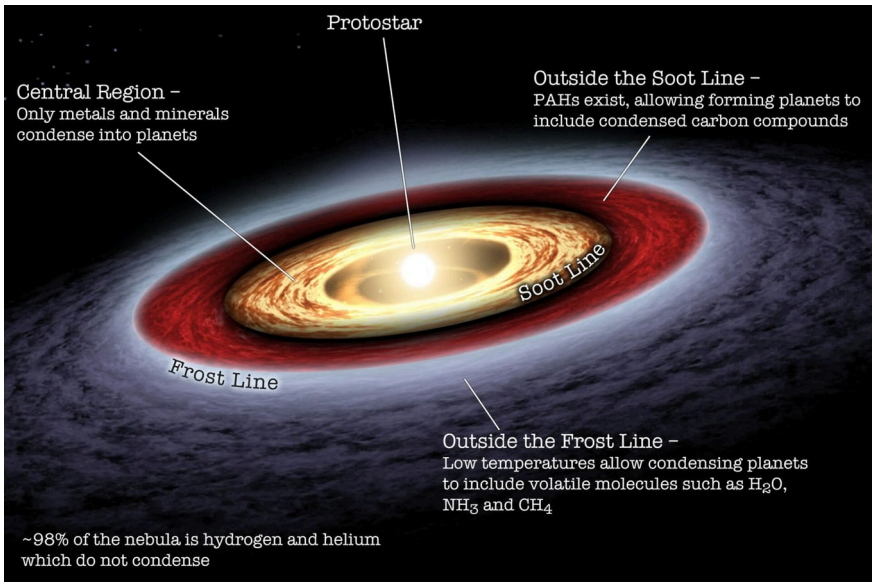
2.5.5.4: How the Sun Works
2.5.5.5: A “Quick Guide” to Solar Fusion—The Proton-Proton Cycle
2.5.5.6: Solar Regions
2.5.5.7: Major Solar Features
2.5.5.8: Studying the Sun
2.5.5.9: The Dark Side of the Sun
2.5.5.10: The National Solar Observatory
2.5.5.11: NASA’s Heliophysics Science Division

This page titled Chapter 2.5: Solar System Formation and Other Stellar Systems is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6: Solar System Formation and Other Stellar Systems by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.1: FORMATION

The solar system is though to begin as a nebular that slowly condenses down into a centra star, or star system, with orbiting planets and minor bodies.



Artist's rendition of a protostar system.

Image courtesy of Wikipedia. Author: NASA/JPL-Caltech. Liscense.

2.5.1.1: Our Solar System

2.5.1.2: The Nebular Theory

2.5.1.3: What is a Planet?

2.5.1.4: The Nebular Theory - Other Important Evidence

2.5.1.5: The Nebular Theory - Proplyds

2.5.1.6: Systems and Extrasolar Planets

2.5.1.7: The Discovery of Extrasolar Planets

2.5.1.8: The Kepler Mission

2.5.1.9: What We Have Learned about Stellar Systems and Extrasolar Planets?

2.5.1.10: The Circumstellar Habitable Zone

2.5.1.11: Some Interesting Exoplanets

2.5.1.12: Extrasolar Planets

This page titled 2.5.1: Formation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

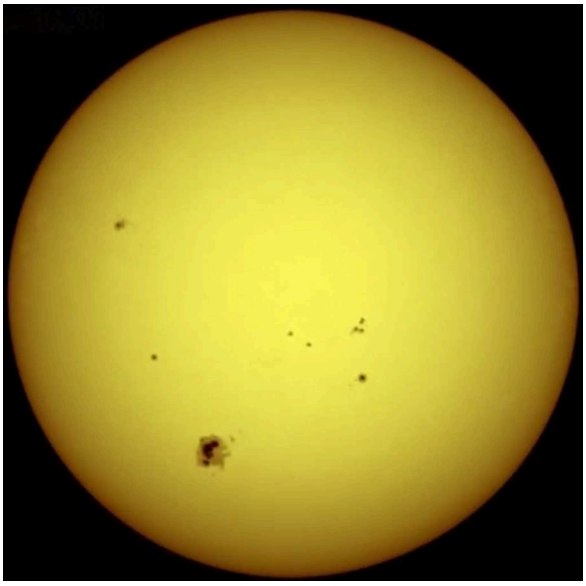
2.5.1.1: Our Solar System

Astronomers believe the solar system is about **4. 6 billion years old**. We base that idea through the study of meteorites and "dating" of these objects and observations of our Sun, as well as similar stars.

The Solar System is made up of one central star, eight (or nine, or ten...) known planets, satellites orbiting the planets, and miscellaneous debris; minor bodies; asteroids, meteoroids, comets, and dust, and what is known as the **Kuiper Belt Objects** and the **Oort cloud**.

Sun

One Central Star—Sun



Public Domain | Image courtesy of NASA / ESA.

Planets

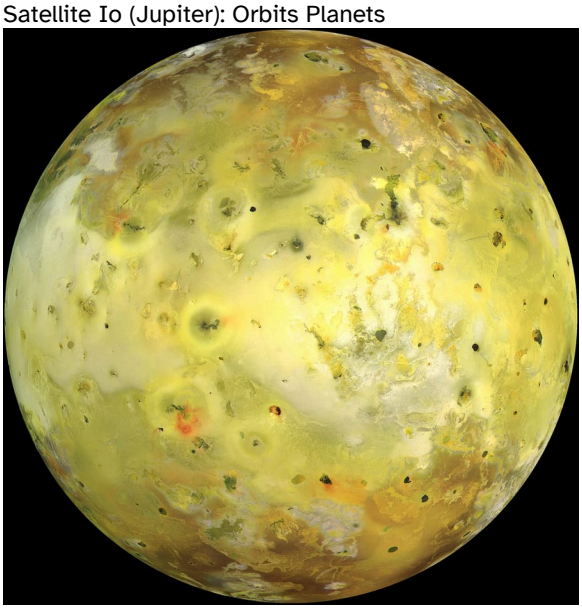
Rocky Planet Earth– Eight (or nine, or ten...) known Planets



Public Domain | Image courtesy of NASA / ESA.

Satellites

Satellite Io (Jupiter): Orbits Planets



Public Domain | Image courtesy of NASA / ESA.

Minor Bodies

Asteroid-951 Gaspra



Public Domain | Image courtesy of NASA / ESA.
Comet Hale-Bopp



CC BY-SA 3.0 | Image courtesy of Wikimedia Authors: E. Kolmhofer, H. Raab; Johannes-Kepler-Observatory, Linz, Austria.

This page titled 2.5.1.1: Our Solar System is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6.1: Our Solar System by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.1.2: The Nebular Theory

So...how did the solar system form and end up with all these different types of objects? Currently the best theory is the **Nebular Theory**. This states that the solar system developed out of an interstellar cloud of dust and gas, called a **nebula**. This theory best accounts for the objects we currently find in the Solar System and the distribution of these objects. The Nebular Theory would have started with a cloud of gas and dust, most likely left over from a previous supernova. The nebula started to collapse and condense; this collapsing process continued for some time. The Sun-to-be collected most of the mass in the nebula's center, forming a **Protostar**.

A protostar is an object in which no nuclear fusion has occurred, unlike a star that is undergoing nuclear fusion. A protostar becomes a star when nuclear fusion begins. Most likely the next step was that the nebula flattened into a disk called the **Protoplanetary Disk** ; planets eventually formed from and in this disk.

Three processes occurred with the nebular collapse:

Gravitational Collapse: Mutual gravitational attraction pulls the diffuse gas and dust inward toward a denser center. This converts gravitational potential energy into heat, causing the center to warm up significantly and eventually form a protostar (the Sun).

Conservation of Angular Momentum: As the cloud shrinks inward, it begins to spin faster and faster to conserve its angular momentum (similar to an ice skater pulling in their arms).

Conservation of Angular Momentum: As the cloud shrinks inward, it begins to spin faster and faster to conserve its angular momentum (similar to an ice skater pulling in their arms).

This page titled 2.5.1.2: The Nebular Theory is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.1.3: What is a Planet?

The International Astronomical Union (IAU) ruled in August 2006 that to be a planet an object must meet three criteria:

1. It must have enough mass and gravity to gather itself into a sphere
2. It must orbit the sun
3. It must reign supreme in its own orbit, having “cleared the neighborhood” of other competing bodies

Is there a controversy?

Yes. This ruling demoted Pluto from a planet to a dwarf planet. According to the new rule, Pluto has not ‘reigned supreme in its own orbit, ‘clearing the neighborhood’ of other competing bodies.

The International Astronomical Union (IAU) has sanction over astronomical definitions and rules. Most of these are minor decisions, like adding leap seconds to a year. The decision – stating what a planet is – is a major decision. The *Pluto Planet* debate had been brewing for several years.

Yet this was not the IAU’s first vote on the subject. One week prior to the “final” planet definition vote, the IAU’s Planet Definition Committee endorsed the following:

1. It must have enough mass and gravity to gather itself into a sphere
2. It must orbit the sun

What were the implications of this *first definition?*

Pluto was still a planet, so was its moon **Charon**, the asteroid **Ceres**, and the recently discovered Eris. For about a week, the solar system would have had twelve planets instead of nine. The IAU rejected this definition and the now-infamous definition accepted.

What are the implications of the *new planet definition?*

There are now **eight planets** in our solar system, along with a new classification of objects called **Dwarf Planets**. Currently there are three Dwarf Planets: Pluto, Ceres (previously classified as an asteroid), and Eris. A subclass of Dwarf planets is called the **Plutoids** ; these ice dwarfs are trans-Neptunian (past the planet Neptune) planets that orbit the Sun and are large enough to be somewhat spherical in shape. It’s important to **note that these rulings only apply to our solar system, not to other stellar systems with planets**.

This decision is still hotly contested some ten years later. With the arrival of the **New Horizons** spacecraft at Pluto and its five known moons in 2015, astronomers will take a closer look at this icy world.



Public Domain | Image courtesy of NASA / ESA.

This page titled 2.5.1.3: What is a Planet? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6.3: What is a Planet? by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.1.4: The Nebular Theory - Other Important Evidence

The types of objects found within the solar system provide significant clues and evidence to support the Nebular Theory. First, the types of Planets and their distributions: with the Rocky planets being close to the Sun, and Gas Giants planets being far from the Sun, Dwarf Planets or Plutoids, a class of Dwarf planets, are found far from the Sun. Comets, asteroids, and meteorites recovered on Earth also provide a number of clues and evidence of Nebular-type development. And the motions of most solar system objects orbit and rotate in an organized fashion. There are a few exceptions to what we would expect to find. For example, we would not expect to find a planet with a large moon, specifically Earth and the Moon. Uranus is tilted on its side and 'rolls' around the Sun. And Venus rotates in retrograde (backwards) as it orbits the Sun, contrary to the other planets in the solar system.

This page titled 2.5.1.4: The Nebular Theory - Other Important Evidence is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

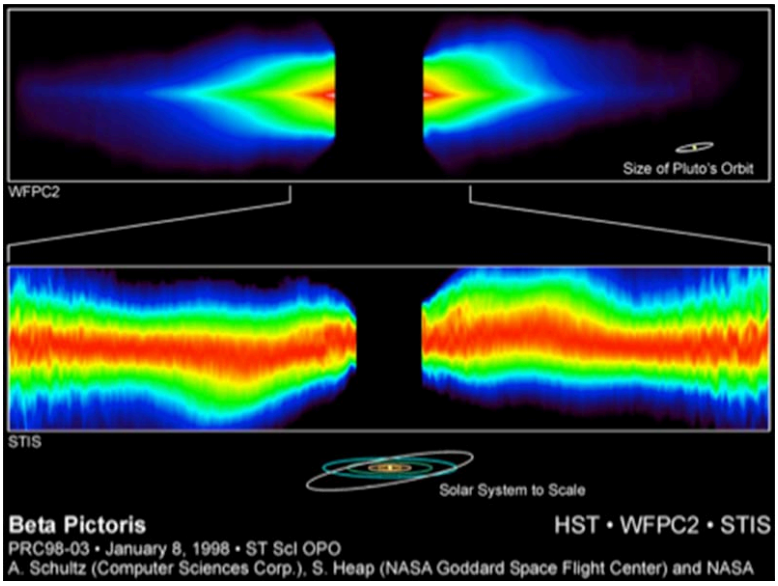
2.5.1.5: The Nebular Theory - Proplyds

What Evidence do we have of a Nebular Theory-type development?

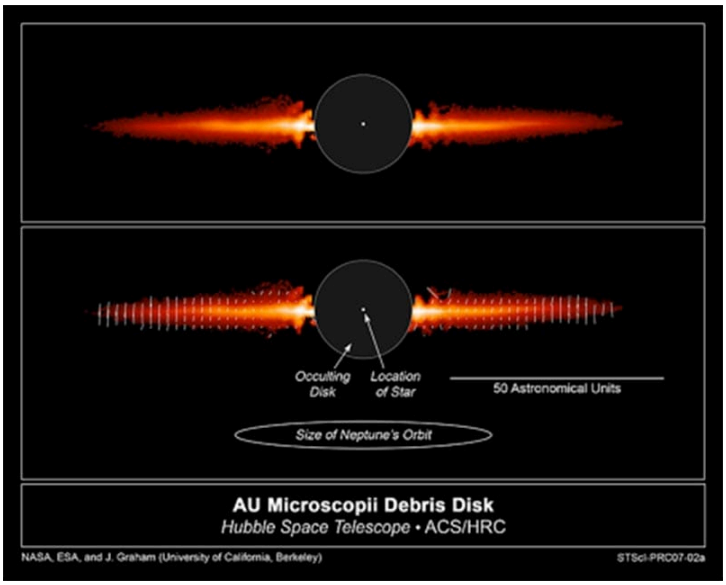
We have observed discs of gas and dust around other stars. We can also see evidence of stars and planets forming in clouds of gas and dust; young planet systems in the making are called Proplyds. Other disk-forming evidence found throughout the Universe includes spiral galaxies. Computer modeling is used to model formation of stellar systems, like our solar system.



NASA Hubble Space TelescopePublic Domain | Image courtesy of NASA / ESA.



NASA Hubble Space TelescopePublic Domain | Image courtesy of NASA / ESA.

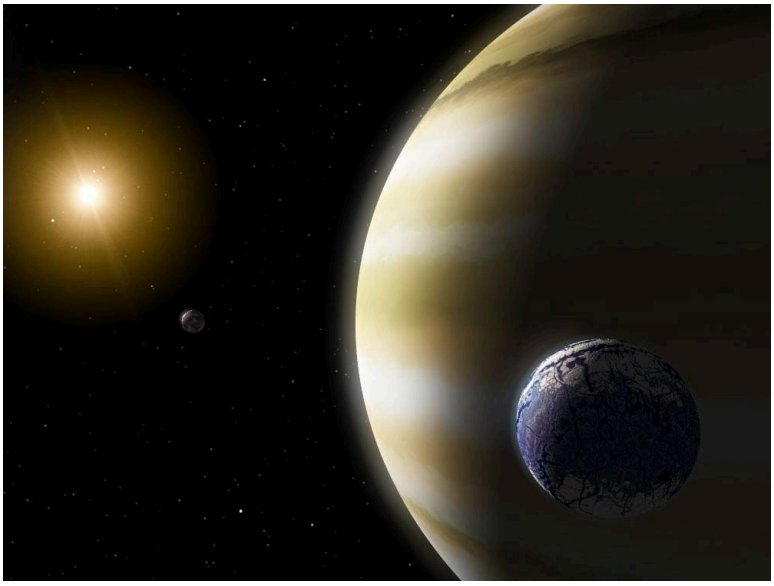


NASA Hubble Space TelescopePublic Domain | Image courtesy of NASA / ESA.

This page titled 2.5.1.5: The Nebular Theory - Proplyds is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6.5: The Nebular Theory - Proplyds by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.1.6: Systems and Extrasolar Planets



Public Domain | Image courtesy of NASA.

Are their other planets orbiting other stars?

Yes. We have found evidence of **Extrasolar Planets** (or **Extra-Solar** ... *2 words*). These are planets orbiting other stars outside our solar system, thus the term “Extrasolar. ” Extrasolar planets are also called **Exoplanets**, the more-commonly used term.

Instead of calling these solar systems, they are called **Stellar Systems**. Why? Because the solar system is the proper name for our sun, its planets, satellites, asteroids, comets, debris, etc.

This page titled 2.5.1.6: Systems and Extrasolar Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6.6: Systems and Extrasolar Planets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.1.7: The Discovery of Extrasolar Planets

Dr. Geoff Marcy, UCSF-UCB (University of California San Francisco – University of California Berkeley) led the research efforts in the late 1980s to develop extrasolar planet search techniques. The main question was: What do you look for when trying to see a planet around another star?

Image of an exoplanet caught on the move. It is Beta Pictoris b, about nine times
as far from its star Beta Pictoris as Earth is from the Sun.
Image courtesy of ESO/A. -M. Lagrange.
The image above shows an exoplanet caught on the move. Beta Pictoris is blocked from view because it is too bright to see it *and* the planet *Beta Pictoris b* at the same time.

Several techniques have been developed:

1. Technology has allowed both amateur and professional astronomers the ability to make such observations from here on Earth. And eventually technology should allow astronomers to directly view exoplanets.

The Discovery of Extrasolar Planets

As of fall 2014, astronomers have found **1,822** planets in 1,137 planetary systems and 467 multiple planet systems. Seven planets have been discovered around the star HD10180, the largest known system outside our solar system. And currently there are thousands of candidate exoplanets.

So what are these exoplanets like?

Most are *much* bigger than Earth, more like Jupiter and even much larger than Jupiter. Many are extremely hot; they orbit quite close to their star or stars. And many of these will “lose” to their parent stars and crash into their star as the star’s gravity pulls them in.

This page titled 2.5.1.7: The Discovery of Extrasolar Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6.7: The Discovery of Extrasolar Planets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.1.8: The Kepler Mission



The image above shows a size comparison of the exoplanets Kepler-20e and Kepler 20f with Venus and Earth. Image courtesy of NASA/ Ames/JPL-Caltech.

The **Kepler Mission** is NASA's first mission capable of finding Earth-size and smaller planets around other stars using an orbiting space telescope. The Kepler Space Telescope has been monitoring 100,000 stars similar to our Sun since its launch in March 2009. The Kepler Space Telescope, named for astronomer Johannes Kepler, looks for planets orbiting stars by looking for changes in a star's brightness.

Kepler's primary mission is to find Earth-size exoplanets. Detecting the massive gas giant exoplanets – called **Super Jupiters** – has been relatively easy with other methods. So the Kepler Space Telescope watches transits and measures the reduction in the star's light due to the planet transiting in front of its star.

This image shows a size comparison of the exoplanets Kepler-20e and Kepler 20f with Venus and Earth.

This page titled 2.5.1.8: The Kepler Mission is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6.8: The Kepler Mission by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.1.9: What We Have Learned about Stellar Systems and Extrasolar Planets?

How rare are these stellar systems? We do not know yet... but many stellar systems are not made up of just one star. **Binary Stars** are systems in which physically associated star systems are made up of two stars. It appears that about half of all stars are binary star systems. And **Multiple Star Systems** have more than two stars. We have found and observed numerous multiple star systems. One such system we see in Scorpius – called **β Scorpii** – has at least **7 stars**. Binary and multiple star systems present challenges for intelligent life.

This page titled 2.5.1.9: What We Have Learned about Stellar Systems and Extrasolar Planets? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.1.10: The Circumstellar Habitable Zone

The area around a star within which a planet or planets with sufficient mass and atmospheric pressure can support liquid water at the planet's surface is called the **circumstellar habitable zone** ; also called the **Goldilocks Zone**. First theorized in 1953, numerous exoplanets have been discovered within this Zone. Not all biologists and astronomers agree that a planet must exist in the circumstellar habitable zone to support life or even intelligent life; many point to **extremophiles** found on Earth; organisms which flourish in extreme Earth conditions that would be harmful to most life.

It appears that one in five stars have an Earth-sized planet in a circumstellar habitable zone. There could be as many as 40 billion circumstellar habitable zone planets in the Milky Way Galaxy.

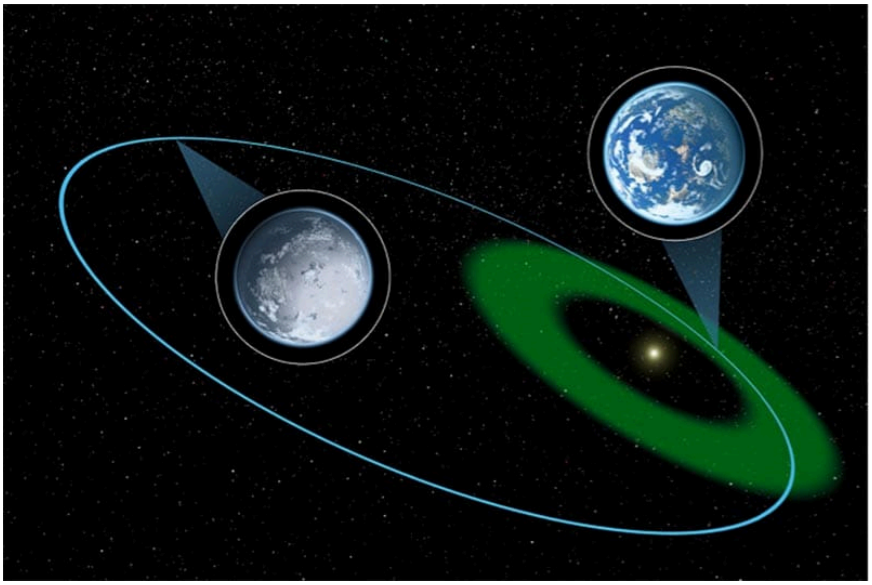


Image courtesy of NASA/JPL-CaltechThe image above shows an Eccentric Habitable Zone – a planet with a very-elliptical orbit that goes from a circumstellar habitable zone on the right to what would be considered uninhabitable during most of its orbit.

This page titled 2.5.1.10: The Circumstellar Habitable Zone is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6.10: The Circumstellar Habitable Zone by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.1.11: Some Interesting Exoplanets

TrES-4 was discovered when looking for transiting planets. It is about 70 percent bigger than Jupiter but less massive. TrES-4 has the density of approximately 0. 2 grams per cubic centimeter, or that of balsa wood. The **Gliese 581 System** is a red dwarf star with a mass about 1/3 that of the Sun. It is located 20. 4 light years from Earth and is among the 100 closest stars to Earth. Gliese 581 has at least four planets:

- **HD80606b** is a strange planet about four times the mass of Jupiter. It comes so close to its star, at its perihelion point HD80606b would have a surface temperature of 1,200oC.

  Then there is the strange planet, **Wasp-12b**. It takes this planet 1. 1 earth days to go around its star one time; in other words, you would celebrate your birthday every 26. 4 hours. Wasp-12b's surface temperature is estimated to be 1,500oC, and it is slowly being *eaten* by its star.

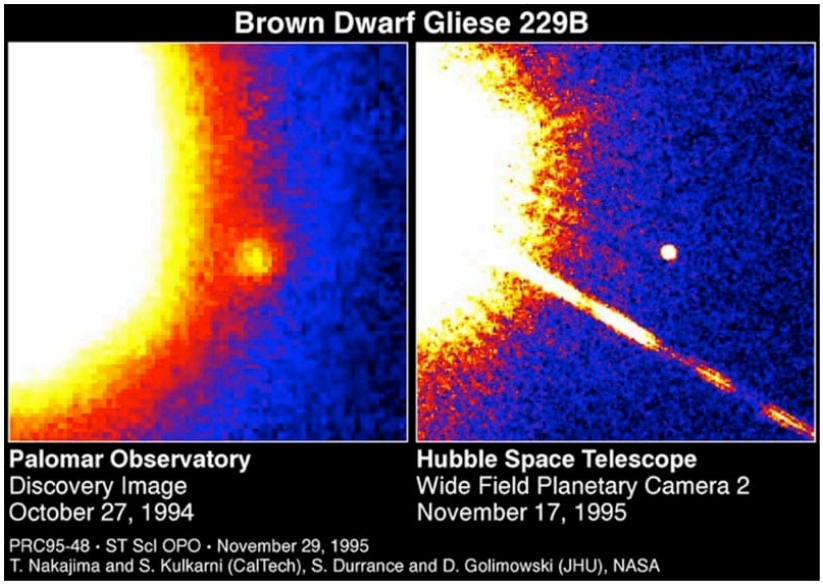
Stellar System Stars and Life

Many biologists and astronomers believe that you have to have the right type of star, as well as the right type of planet, and the right distance from its star for intelligent life.

This page titled 2.5.1.11: Some Interesting Exoplanets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6.11: Some Interesting Exoplanets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.1.12: Extrasolar Planets



These images show Brown Dwarf Gliese 229B, orbiting its star Gliese 229. Gliese 229B is in the white circle on each image. The left image was the discovery image, taken with the Palomar Telescope in California. The right image was taken with the Hubble Space Telescope. Public Domain | Image courtesy of NASA.

The extrasolar planets tell us a lot about our own solar system's formation and development. Yet, there are some significant issues at this point. First, there appears to be numerous exoplanets much larger than Jupiter. Why is this so? And many of these extrasolar planets orbit much closer to their stars than our Gas Giants, which are much further from the Sun. The thinking here is that these Gas Giants "migrated" closer to their stars during the development of their stellar systems.

Brown Dwarfs are sub-stellar objects with a mass below that necessary to maintain hydrogen nuclear fusion reactions in their cores, as do stars. Sizes of these brown dwarfs – up to 80 Jupiter masses – range from very large gas giant planets to just below the mass necessary to "turn on" as a star. Are brown dwarfs' planets? Stars? Failed stars? Two extrasolar planets have been discovered orbiting brown dwarfs.

Consider this...

Giordano Bruno (1548-1600) was an Italian Dominican Friar, poet, philosopher, astronomer, and mathematician. His theories about the Universe were considered extreme for his day, eventually leading to his being burned at the stake. Bruno thought the stars were distant suns being orbited by their own planets. He also thought these planets might even harbor life of its own. Bruno also believed the Universe was infinite.

There are countless suns and countless earths all rotating around their suns...

— Friar Giordano Bruno – (1584)

This page titled 2.5.1.12: Extrasolar Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 6.12: Extrasolar Planets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2: THE ROCKY PLANETS

"How inappropriate to call this planet Earth when clearly it is Ocean..." - Arthur C. Clark, 1917-2008

Learning Objectives

Upon completion of this module, the student will be able to:Upon completion of this module, the student will be able to:

- Identify the characteristics of Rocky Planets
- Define craters and cratering
- Identify the primary characteristics of each rocky planet
- Identify the similar characteristics of each rocky planet
- Describe a circumstellar habitable zone
- Discuss the characteristics of the Moon, Deimos, and Phobos

2.5.2.1: What do you think?

2.5.2.2: Comparing Planets

2.5.2.3: Rocky Planets

2.5.2.4: Craters

2.5.2.5: Planet Mercury

2.5.2.6: Planet Venus

2.5.2.7: Planet Earth

2.5.2.8: Earth's Moon

2.5.2.9: Planet Mars

2.5.2.10: Snow on Mars

2.5.2.11: A Martian Controversy

2.5.2.12: Martian Moons

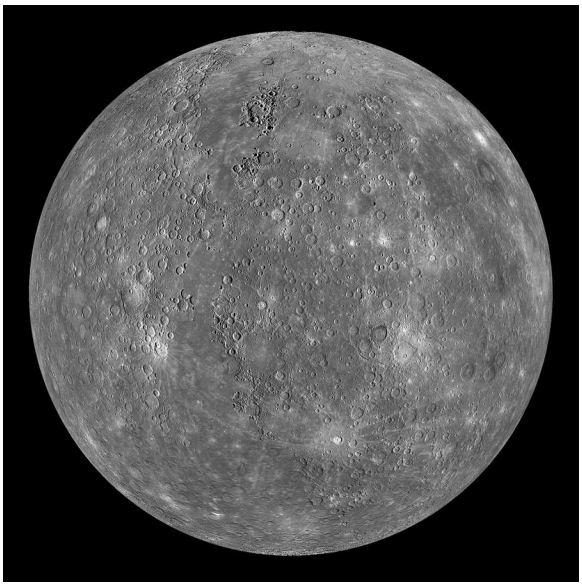
2.5.2.13: Characteristics of the Solar System's Rocky Planets

This page titled 2.5.2: The Rocky Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7: The Rocky Planets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.1: What do you think?

Imagine that you are approaching each of our Solar System's Rocky Planets in a spaceship for the first time. Looking at photos of each of the rocky planets, what similarities do you see at first glance?



MercuryPublic Domain | Image courtesy of NASA / JPL-CALECH.

Image of planet Venus.

VenusPublic Domain | Image courtesy of NASA / JPL-CALECH.



EarthPublic Domain | Image courtesy of NASA / ESA.

Image of planet Mars.

MarsPublic Domain | Image courtesy of NASA / JPL-Caltech.

This page titled 2.5.2.1: What do you think? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.1: What do you think? by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.2: Comparing Planets

We will look at our Solar System and its worlds from a comparison perspective, called **Comparative Planetology**, rather than focus on a lot of numbers and data. To compare the planets, they will be grouped into two major types: the **Rocky** or **Terrestrial (Earth-Like) planets**, and the **Gas Giants** or **Jovian (Jupiter-like) planets**. Recall, the focus of this module will be on the Rocky Planets. The next module will focus on the Gas Giants.

The table, **Rocky Planets Vs. Gas Giant Planets**, provides a broad snapshot of the differences between the Rocky Planets and the Gas Giant Planets.

Rocky Planets Vs. Gas Giant Planets

Rocky Planets: Mercury-Venus-Earth-Mars

Characteristics:

- Distance from the Sun — Close
- Revolution Length — Short
- Surface Type — Solid surface
- Diameter — Small
- Craters — Many
- Water — All except Venus
- Rings — None
- Satellites — Few or none

Gas Giant Planets: Jupiter-Saturn-Uranus-Neptune

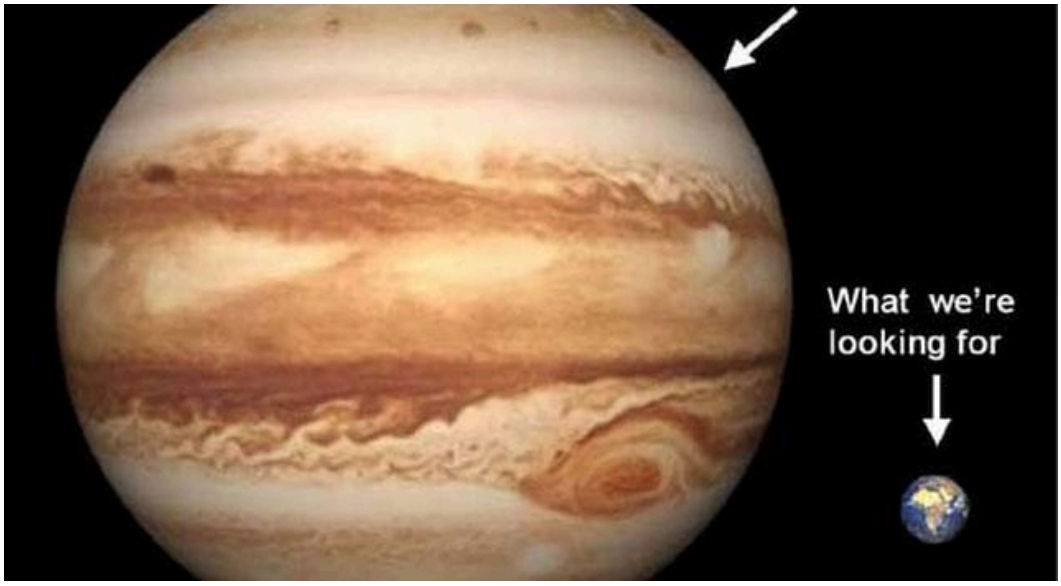
Characteristics:

- Distance from the Sun — Farther out
- Revolution Length — Long
- Surface Type — Gas; Probably no surface or very small icy core
- Diameter — Large
- Craters — None (gas)
- Water — Yes; in the atmosphere (gas)
- Rings — Yes; from one to multiple rings
- Satellites — Many

This page titled 2.5.2.2: Comparing Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.2: Comparing Planets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.3: Rocky Planets



Public Domain | Image courtesy of NASA.

Rocky Planets are relatively small when compared to the Gas Giants or many of the recently discovered Exoplanets. Rocky Planets have solid surfaces, which exhibit impacts through multiple craters. There are no rings, few or no moons, and evidence of past or current tectonic, volcanic activity. Rocky planets are sometimes referred to as the **Terrestrial Planets** (Earth-like).

This page titled 2.5.2.3: Rocky Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.3: Rocky Planets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.4: Craters

One characteristic we will find throughout the Solar System is bodies that have solid surfaces and exhibit impact or craters. **Craters** are depressions and scars on a Solar System body's surface. The word crater comes from the Greek for bowl; these are bowl-like features. Most cratering is due to impacts by space debris, called Minor Bodies. This includes asteroids, comets, and smaller debris. Planets and satellites can also have craters due to volcanic activity. Crater sizes can span from microscopic to hundreds of miles across. Occasionally an impact not only produces a crater, it produces a ray. **Rays** are impact "splash;" imagine dropping a rock into flour.

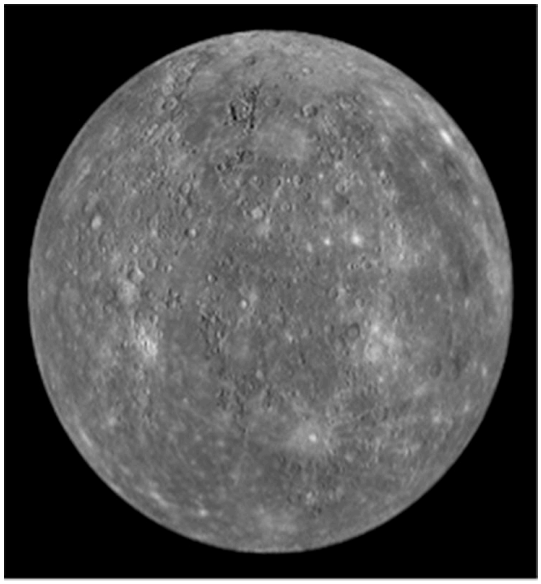


Image courtesy of Mike Reynolds, Ph. D. of Florida State at Jacksonville.

This page titled 2.5.2.4: Craters is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.2.5: Planet Mercury

Mercury is the smallest and closest planet to the Sun. It is actually smaller than two of the Solar System's moons—Ganymede (Jupiter) and Titan (Saturn). Because Mercury has no real atmosphere, it has extreme high and low temperatures. Examining the planet, Mercury looks a lot like the Moon.

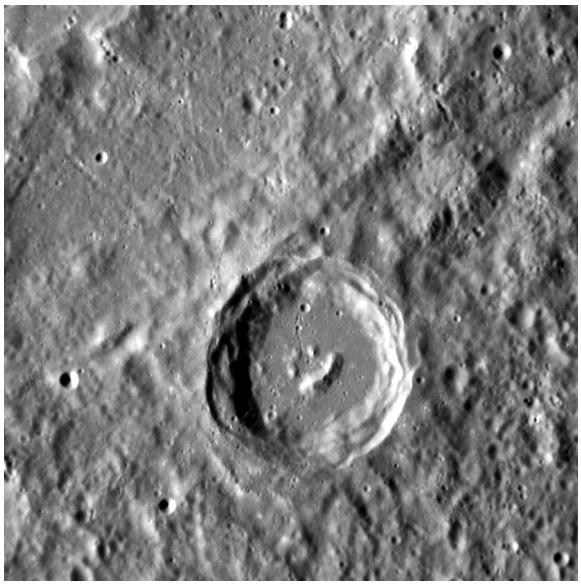


MercuryPublic Domain | Image courtesy of NASA / JPL-CALECH.

Planet Mercury at a Glance

Characteristic – Current State

- Impact Craters – Yes
- Tectonic Craters – Yes; not active
- Volcanoes – Yes; not active
- Atmosphere – No real atmosphere
- Water – Yes; water and ice in craters at the poles
- Erosion – No
- Dunes – No
- Polar Caps – No
- Satellites – No
- Life – Unknown



Craters on MercurynCC BY 2. 0 | Image courtesy of NASA Goddard Space Flight Center from Greenbelt, MD, USA.

This page titled 2.5.2.5: Planet Mercury is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.5: Planet Mercury by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.6: Planet Venus

Venus is the second planet from the Sun. It is often called Earth's twin because they are similar in size and both have cloudy atmospheres. It is also the closest planet Earth approaches, and besides the Moon, usually the second brightest object in the night sky. Venus is often called the Morning Star or the Evening Star because it is very bright when seen in the morning or evening sky.

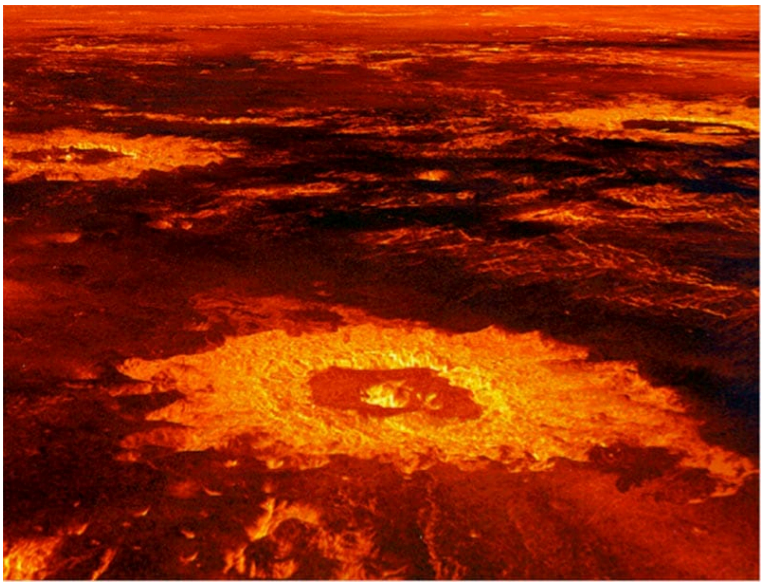
Planet Venus at a Glance

Characteristic — Current State

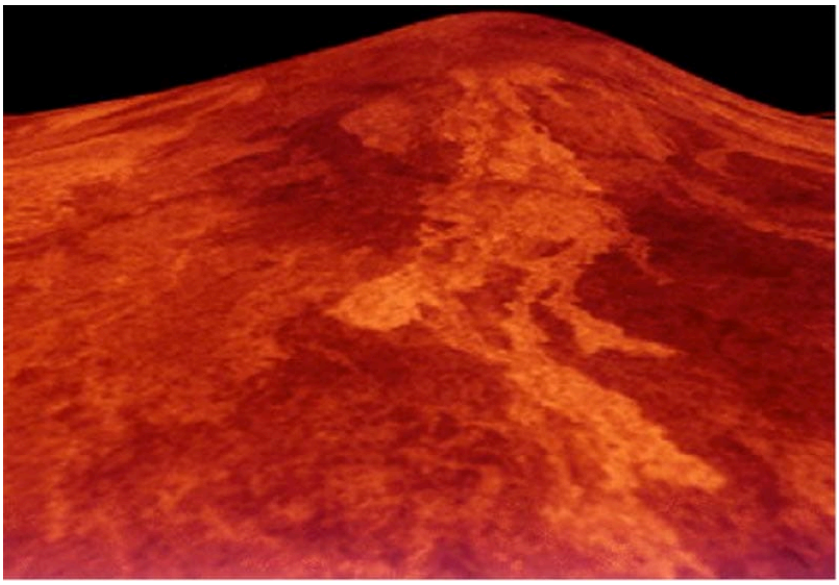
- Impact Craters – Yes
- Tectonic Craters – Yes
- Volcanoes – Yes; over 1,600 volcanoes – most on one body in the Solar System
- Atmosphere – Carbon Dioxide (CO_2), Sulfuric Acid (H_2SO_4), High Pressure, Clouds, Odd Polar Vortex
- Water – No
- Erosion – Yes
- Dunes – Yes
- Polar Caps – No
- Satellites – No
- Life – Unknown

Image of the planet Venus.

VenusPublic Domain | Image courtesy of NASA.



Crater on the surface of VenusPublic Domain | Image courtesy of NASA.



Sif Mons Volcano on VenusPublic Domain | Image courtesy of NASA.

Note: The image of Venus is blue here because it is a filtered image. In visible light, Venus looks like this:

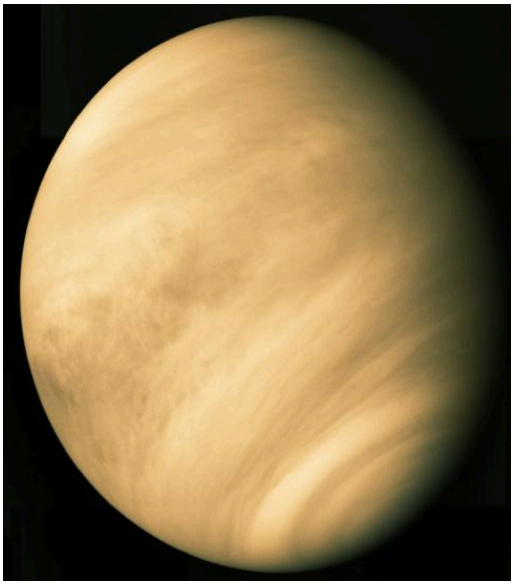
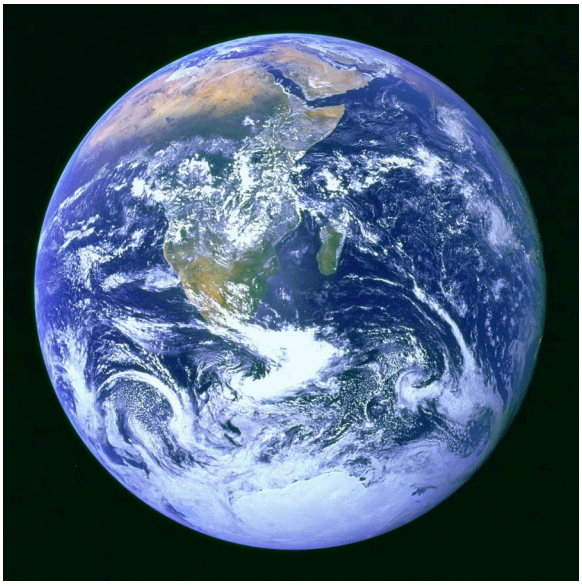


Image used courtesy of the European Space Agency.

This page titled 2.5.2.6: Planet Venus is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.6: Planet Venus by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.7: Planet Earth



EarthPublic Domain | Image courtesy of NASA.

Earth is the largest of the Solar System's rocky planets and the **only known planet with liquid water**. Water makes up about 75% of Earth's surface.

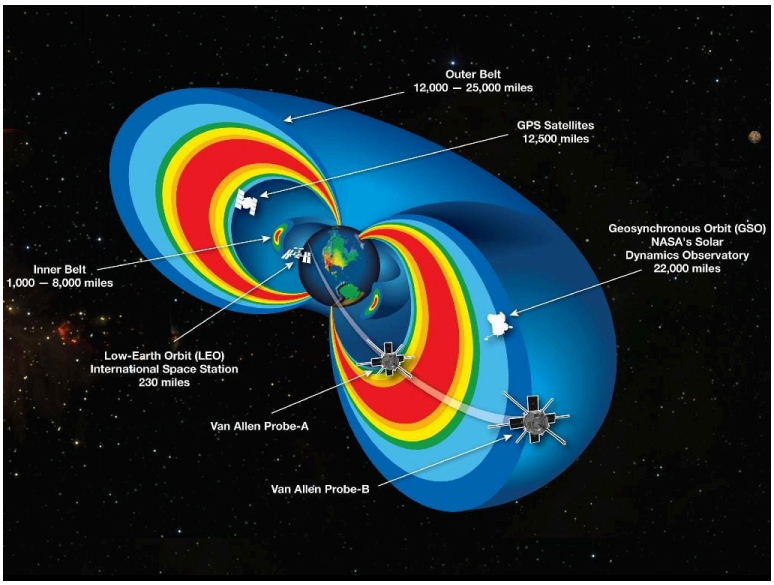
Daily changes in the level of the water – known as **tides** – are due to the Moon and Sun's gravitational influences. These gravitational effects raise tidal bulges in the oceans. The size of a specific tide primarily depends on the Sun-Moon orientations relative to Earth. Tidal interaction between Earth and Moon causes the Earth's rotation to slow and the Moon to be locked in a synchronous orbit – revolution is equal to rotation. Some "rocking" of the Moon back and forth occurs, called **librations**.

Forces that Shape Earth: Earth's Geological Processes

Impact Cratering was very intense during the early period of the Solar System; this also shaped Earth into its characteristics.

Other forces that shape Earth include **erosion, weathering, plate tectonics, and volcanism.**

- **Erosion, which is a continuous process,is** the wearing away of the surface by water, atmosphere, mechanical, and chemical processes.
- **Weathering is** a gradual physical and chemical wearing away of rocks and surface material.
- **Plate tectonics is** the motion of a body's plates driven by internal stresses.
- **Volcanism is** the eruption of molten rock from a body's interior onto its surface.



Public Domain | Image courtesy of NASA.

Earth has a dynamic and active **atmosphere**. Overall, Earth has the most dynamic weather in our Solar System. Our atmosphere is not as thick as the Gas Giants or even Venus (Rocky Planet), yet it is like looking through 30 feet of water.

Earth also has a strong **Magnetic Field**, which is unusual for a Rocky Planet. It is believed to be due to Earth's molten core undergoing rotation and convection (giving off heat). Mercury and the Moon exhibit very weak magnetic fields, whereas Venus has none and Mars virtually none; yet there is evidence of a magnetic field in Mar's early history.

The **Van Allen Radiation Belts** are charged Particle Belts made up of ions and electrons, which were predicted by James Van Allen, and discovered in 1958 by the US Satellite Explorer 1. These belts protect Earth as intense solar particles from the Sun's solar wind strikes Earth. These solar wind particles can occasionally be seen as **Aurora**, as the charged solar particles strike Earth's poles.

Earth is the only known planet with life. Earth's orbit is within the **circumstellar habitable zone** – area in a star's orbit where ideal conditions exist for life. If Earth was 5% closer to the Sun, it would be Venus-like, and 20% farther out from the Sun, Earth would be Mars-like. Earth also exists in a system that has the right type of star, only one star, good location within the Milky Way Galaxy, the right type of galaxy, etc. Astronomers, biologists, and physicists have identified over 800 identified factors necessary for intelligent life.

Planet Earth at a Glance

Characteristic — Current State

- Impact Craters — Yes
- Tectonic Craters — Yes
- Volcanoes — Yes
- Atmosphere — Nitrogen (N), Oxygen (O_2), Clouds, Rain, Snow
- Water — Vapor, Liquid, and Ice → Primarily a water planet
- Erosion — Yes
- Dunes — Yes
- Polar Caps — Yes
- Satellites — One, Moon
- Life — Yes

This page titled 2.5.2.7: Planet Earth is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.7: Planet Earth by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.8: Earth's Moon



Earth rise from Lunar OrbitPublic Domain | Image courtesy of NASA.

The Earth-Moon system is sometimes referred to as a double or twin planet, due to the size of the Moon when compared to Earth. As with all of the planets and other moons, we see the Moon due to reflected sunlight. Lunar temperature varies from 253oF in the day to -387oF at night.

How did the Moon form? There are a number of theories—from capture of the Moon by Earth, formation of the Moon in the same area as Earth, and the blob-spinning-of-early Molten Earth. However, the most widely accepted theory is the **Giant Impact Theory**. This theory posits that a Mars-sized rogue planet that astronomers call Orpheus struck early Earth. The evidence seems to highly favor the giant impact theory over others. Evidence of lava flows and a small iron lunar core are at the heart of this theory.



1st Quarter MoonCC BY-SA 2. 0 | Image courtesy of orrey Pines – San Diego – California – USA.

Perhaps surprisingly, water has been found on the Moon. Water ice deposits have been found in the coldest spots of the Moon's South Pole craters that are not exposed to sunlight. And astronomers have discovered the spectral signal of water and/or hydroxyl (HO-) ion from three different spacecraft. Estimates are that one ton of the lunar surface top layer would hold about 32 ounces of

water.



Large Lunar Craters Eratosthenes (upper left) and Copernicus (center right)Image courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

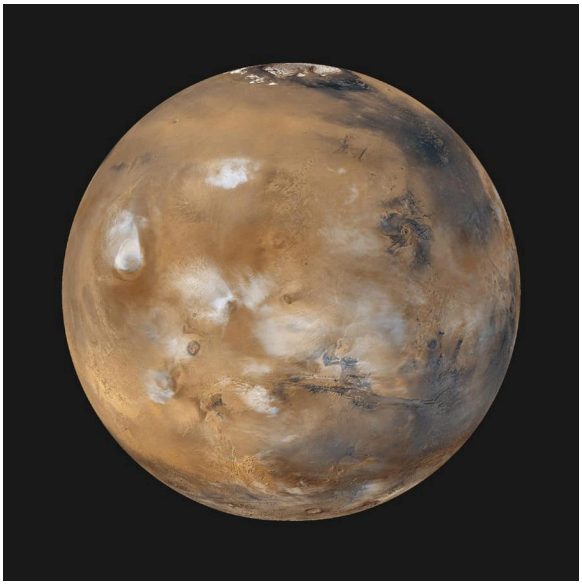


Full *Moon*Image courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

This page titled 2.5.2.8: Earth's Moon is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.8: Earth's Moon by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.9: Planet Mars



Mars – note the polar cap (top), cloudsPublic Domain | Image courtesy of NASA.

Mars is the fourth planet from the Sun. It is historically called the **Red Planet** because of its reddish appearance in Earth's night sky. Mars has some Earth-type characteristics, such as Polar Caps, clouds, and water in at least the ice or solid phase. It was very active volcanically at one time. Numerous spacecraft have successfully explored Mars from orbit and the surface. It is by far the most-visited planet (by spacecraft) in our Solar System, besides Earth.

Differences between Martian North and South Polar Caps

North Polar Cap:

- Primarily composed of water ice, H2O
- Size of Greenland
- Seasonal
- Goes through a variety of change

South Polar Cap:

- Primarily composed of carbon dioxide ice, CO_2
- 1/10 the size of the North Polar Cap
- Always there
- On one side of Pole

The differences in the Martian polar caps are probably due to different polar climates. It's important to note that Earth's Polar Caps are also different; the North Pole is floating ice versus the South Pole as a continent.

This page titled 2.5.2.9: Planet Mars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.9: Planet Mars by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.10: Snow on Mars

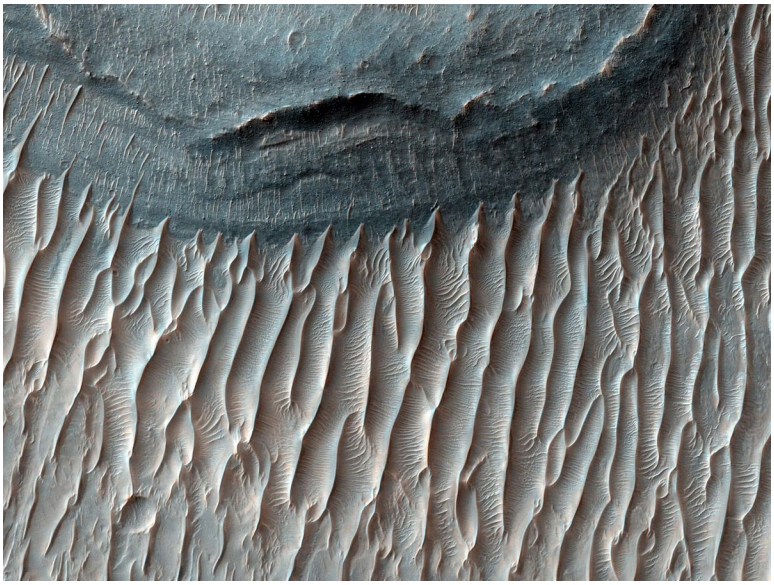
The Phoenix Mars Lander detected snow falling from Martian clouds in 2009. The snow was detected from clouds about 2. 5 miles above the spacecraft's landing site. Data show the snow vaporizing before reaching the ground. Information from the Phoenix Mars Lander also confirmed that a hard subsurface layer at its far-northern site contains water and ice.

Planet Mars at a Glance

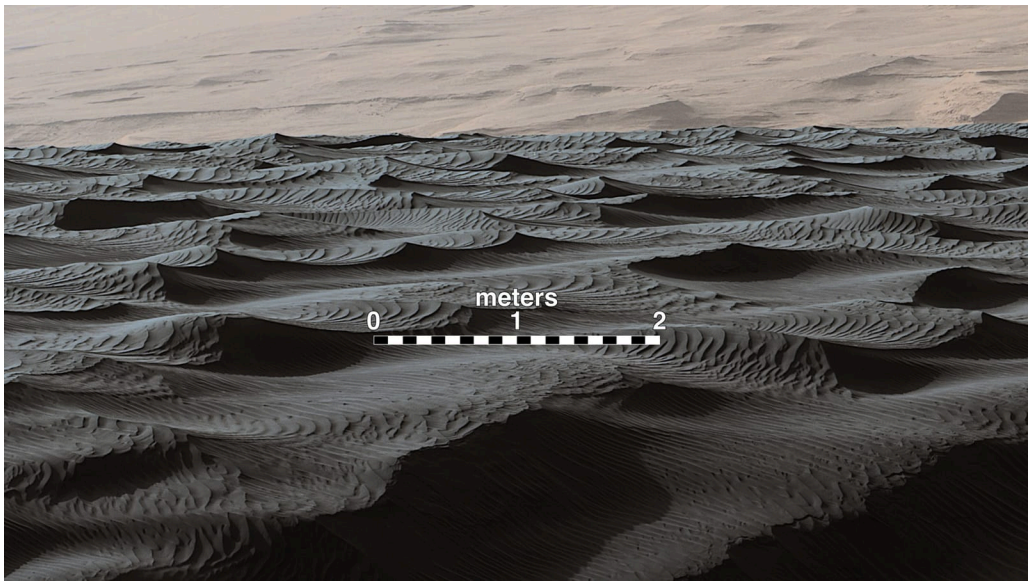
Characteristic — Current State

- Impact Craters — Yes; a very large impact crater **Borealis Basin** (largest known in the Solar System), 6,600 miles across
- Tectonic Craters — Yes, still active?
- Volcanoes — Yes, not active; One of the largest volcanoes in the Solar System; **Olympus Mons**
- Atmosphere — Carbon Dioxide (CO_2), Low Pressure, Clouds, Snow, Dust devils, Dust storms, Slight traces of Methane (CH_4)
- Water — Both liquid & water ice
  - Subsurface liquid water aquifers?
  - 2% water in soil
- Erosion — Yes
- Dunes — Yes
- Polar Caps — Yes
- Satellites — Two; Deimos and Phobos
- Life — Unknown; we are actively searching for life

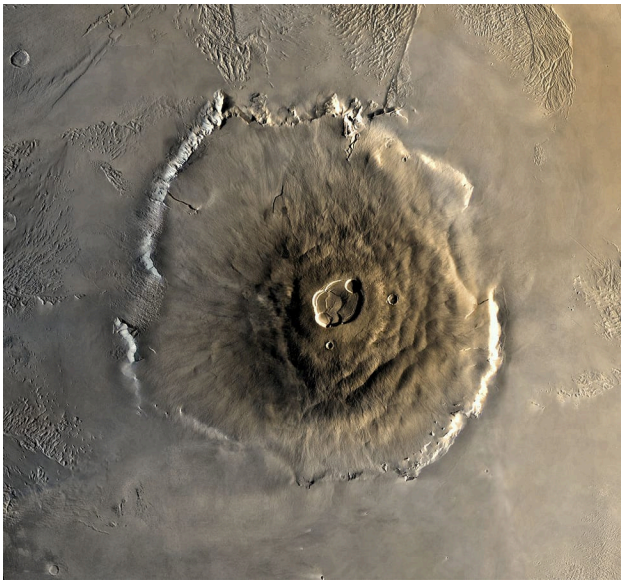
Public Domain | Image courtesy of NASA.



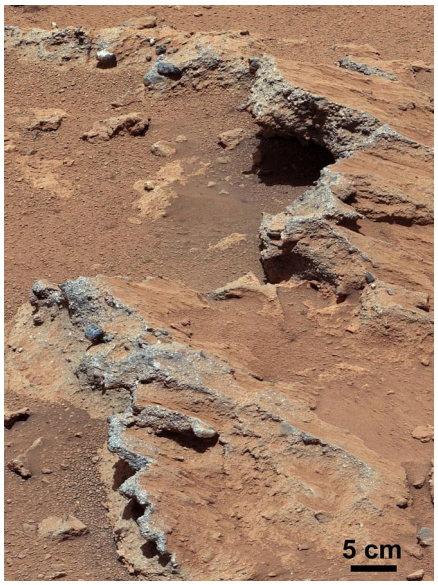
Ius Chasma Floor: The layered deposits consist of dark basalt lava flows and bright sedimentary layers; likely to be from atmospheric dust, sand, or alluvium from an ancient water source. Mars Reconnaissance Orbiter, 9/18/2008Public Domain | Image courtesy of NASA.



Sand Dunes on Mars: Endurance Craters are dramatic dune fields on the crater floor. Dune crests have accumulated more dust than the flanks of the dunes and the flat surfaces between the dunes. Images courtesy of NASA



Martian Volcano Olympus Mons. Public Domain | Image courtesy of NASA.



The surface of Mars. NASA Viking SpacecraftPublic Domain | Image courtesy of NASA.

This page titled 2.5.2.10: Snow on Mars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.10: Snow on Mars by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.11: A Martian Controversy

Before spacecraft and more-advanced equipment, we depended on the eye and photography to decipher and understand what we saw in the Universe. Mars was – and it still is – interesting because it is so close to Earth and somewhat "Earth-like. " Some astronomers focused on the planets and specifically Mars. **Percival Lowell** was one such astronomer.

Lowell Observatory in Flagstaff, Arizona

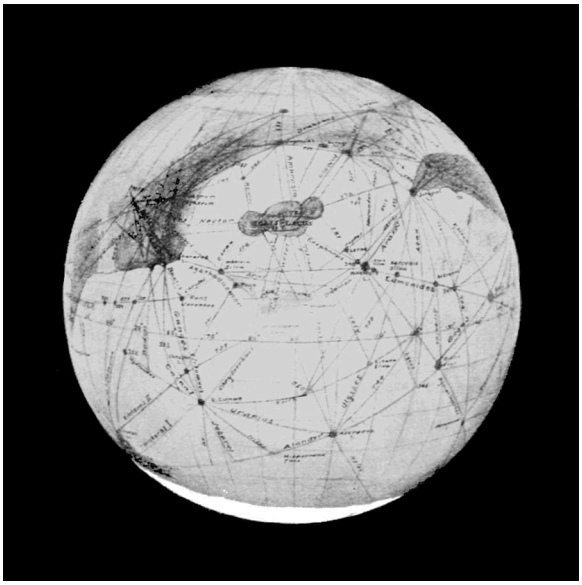
In the late 1800s, Percival Lowell was the first to seek site with good air clarity and superior seeing (steady air or lack of turbulence). After testing many different site tests, Lowell gave the lumber town of Flagstaff, AZ (population of 1,000) high marks. He selected a hill 300 feet above town and built an observatory, which is still in use today.

Lowell initiated Mars Mania; he sensationalized what he perceived as Martian **canals**. The canals were first described by the Italian astronomer Giovanni Schiaparelli during the close Martian approach in 1877. Schiaparelli gave them the Italian name 'canali. ' The translation was taken as canals; canals refer to something made by someone, something, but not Earth-natural.... Lowell pressed the idea that these canals were made by intelligent beings – Martians. There were books and articles that promoted the Martian canals as well as Martian seasonal vegetation growth. This, and surrounding controversy, stimulated interest in Red Planet. And in 1898, science fiction author H. G. Wells wrote "The War of the Worlds," furthering the controversy.

Other observers also saw the Martian Canals. Lowell astronomer Earl Slipher's book, *The Photographic Story of Mars* , shows adjacent direct photographs and drawings. In 1962, Slipher asked readers to judge:

History ... shows that every skilled observer ... has had no great difficulty of seeing and convincing himself of the reality of the canals. Photographs have recorded traces of so many of the canals and oases ... that they should remove all doubt of the reality — E. C. Slipher (1962).

So what happened? Were the canals real? When spacecraft first started exploring Mars in 1965, Mariner 4, no canals were seen. It is now believed that the eye fooled the astronomer, he connected features on Mars that were there – but the canals were not. It was like connecting the dots, an optical illusion.



Lowell's 1916 Map of Mars. The fine lines are the Martian Canals. Public Domain

This page titled 2.5.2.11: A Martian Controversy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.11: A Martian Controversy by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.12: Martian Moons

Mars has two small moons, **Deimos and Phobos**. The names are taken from the Greek for Dread (Deimos) and Fear (Phobos). They were discovered by Asaph Hall in 1877.

Phobos is the larger of the two moons (barely larger than 16 miles at its biggest point) and orbits Mars once every 11 hours and 6 minutes. Both moons are odd-shaped and very small. Both Phobos and Deimos were most-likely captured asteroids or the result of a collision with Mars. As the images show, both moons have craters. Neither moon has an atmosphere.

Image of Deimos.

Photo of Deimos. Public Domain | Image courtesy of NASA/JPL-Caltech/University of Arizona.

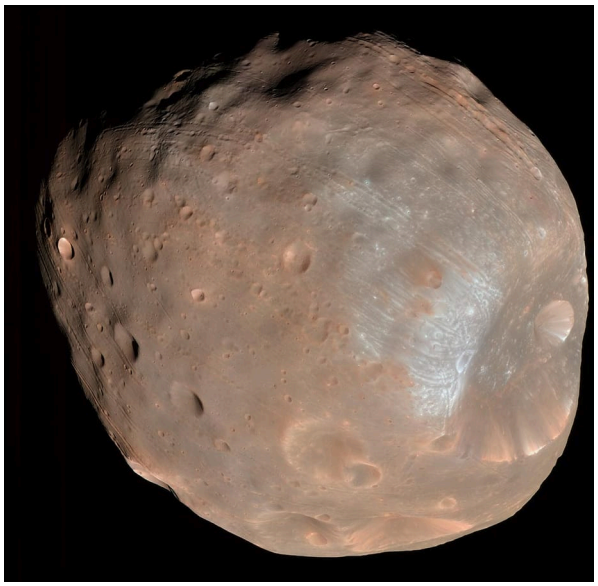


Photo of Phobos. Public Domain | Image courtesy of NASA. Taken by NASA's Mars Reconnaissance Orbiter in 2008.

This page titled 2.5.2.12: Martian Moons is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.12: Martian Moons by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.2.13: Characteristics of the Solar System's Rocky Planets

Comparison of the Rocky Planets

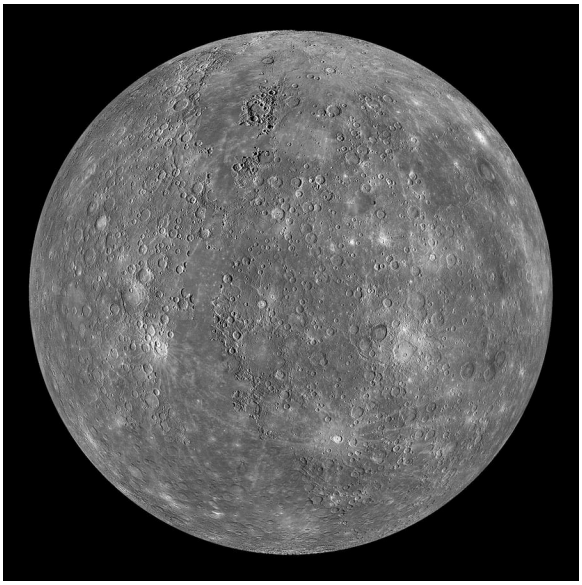
Mercury

Characteristic — Current State

- Impact Craters — Yes
- Tectonic Craters — Yes; not active
- Volcanoes — Yes; not active
- Atmosphere — No real atmosphere
- Water — Yes; water and ice in craters at the poles
- Erosion — No
- Dunes — No
- Polar Caps — No
- Satellites — No
- Life — Unknown

The Bottom Line...

- Looks like the Moon; heavily cratered
- No real atmosphere
- No seasons



Public Domain | Image courtesy of NASA / JPL-CALECH.

Venus

Characteristic — Current State

- Impact Craters — Yes
- Tectonic Craters — Yes
- Volcanoes — Yes; over 1,600 volcanoes – most on one body in the Solar System
- Atmosphere — Carbon Dioxide (CO_2), Sulfuric Acid (H_2SO_4), High Pressure, Clouds, Odd Polar Vortex
- Water — No

- Erosion — Yes
- Dunes — Yes
- Polar Caps — No
- Satellites — No
- Life — Unknown

The Bottom Line...

- Called Earth's twin
- Extreme Greenhouse effect
- Rains sulfuric acid
- Polar Vortex

Image of Venus.

Public Domain | Image courtesy of NASA / JPL-CALECH.

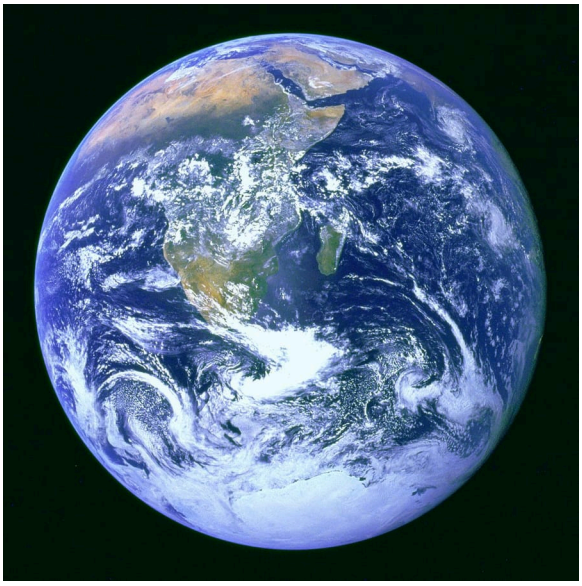
Earth

Characteristic — Current State

- Impact Craters — Yes
- Tectonic Craters — Yes
- Volcanoes — Yes
- Atmosphere — Nitrogen (N), Oxygen (O_2), Clouds, Rain, Snow
- Water — Vapor, Liquid, and Ice → Primarily a water planet
- Erosion — Yes
- Dunes — Yes
- Polar Caps — Yes
- Satellites — One, Moon
- Life — Yes

The Bottom Line...

- Water Planet
- Big moon – double planet
- Life everywhere



Public Domain | Image courtesy of NASA / ESA.

Moon

Characteristic — Current state

- Craters
- Evidence of past volcanic, tectonic activity
- No atmosphere
- Water ice found in a lunar south pole crater
- Probably formed from a collision with Earth
- Looks like Mercury...



CC BY-SA 3. 0 | Image courtesy of Wikimedia Author: Gregory H. Revera.

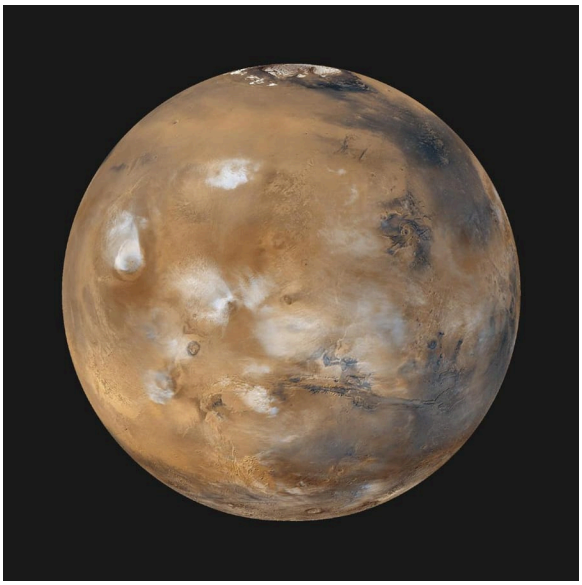
Mars

Characteristic — Current State

- Impact Craters — Yes; a very large impact crater **Borealis Basin** (largest known in the Solar System), 6,600 miles across
- Tectonic Craters — Yes, still active?
- Volcanoes — Yes, not active; One of the largest volcanoes in the Solar System; **Olympus Mons**
- Atmosphere — Carbon Dioxide (CO_2), Low Pressure, Clouds, Snow, Dust devils, Dust storms, Slight traces of Methane (CH_4)
- Water — Both liquid & water ice
  - Subsurface liquid water aquifers?
  - 2% water in soil
- Erosion — Yes
- Dunes — Yes
- Polar Caps — Yes
- Satellites — Two; Deimos and Phobos
- Life — Unknown; we are actively searching for life

The Bottom Line...

- Is liquid water still flowing?
- Very light compared to Venus and Earth (less dense)
- Life?

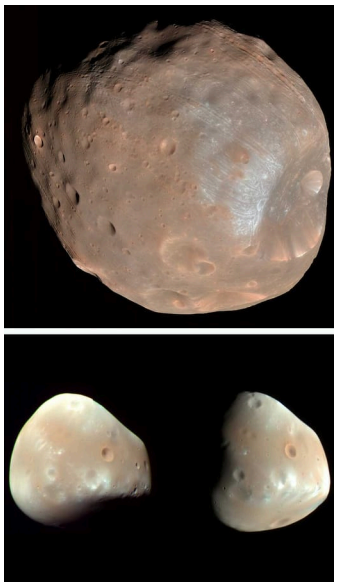


Public Domain | Image courtesy of NASA / JPL-CALECH.

Deimos and Phobos

Characteristic — Current state

- Craters
- No Atmosphere
- Very small
- Captured asteroids or result of collisions with Mars



Public Domain | Images courtesy of NASA/JPL-Caltech/University of Arizona.

Consider this...

How did elementary school kids used to remember the order of the planets from the Sun? A simple mnemonic: My Very Educated Mother Just Served Us Nine Pizzas! (Mercury Venus Earth Mars Jupiter Saturn Uranus Neptune Pluto) But with the demotion of Pluto as a full planet, the pizzas went away, so now it is My Very Educated Mother Just Served Us Nachos!

Note: The image of Venus is blue due to being a filtered image.

This page titled 2.5.2.13: Characteristics of the Solar System's Rocky Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 7.13: Characteristics of the Solar System's Rocky Planets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.3: THE GAS GIANT PLANETS

"Since 'tis certain that Earth and Jupiter have their Water and Clouds, there is no reason why the other Planets should be without them. I can't say that they are exactly of the same nature with our Water; but that they should be liquid their use requires, as their beauty does that they be clear. "

"This Water of ours, in Jupiter or Saturn, would be frozen up instantly by reason of the vast distance of the Sun. Every Planet therefore must have its own Waters of such a temper not liable to Frost. "

Christian Huygens– Discovered Saturn's moon Titan in 1655

Learning Objectives

Upon completion of this module, the student will be able to:Upon completion of this module, the student will be able to:

- Identify the characteristics of the Gas Giant planets
- Describe the formation of rings
- Identify the primary characteristics of each Gas Giant planet
- Compare the characteristics of each Gas Giant planet is similar
- Identify the characteristics of Gas Giant satellites

This module presents the Gas Giant Planets in our solar system: Jupiter, Saturn, Uranus, and Neptune, and their moons. We will examine characteristics of each Gas Giant planet, both similar and dissimilar characteristics.

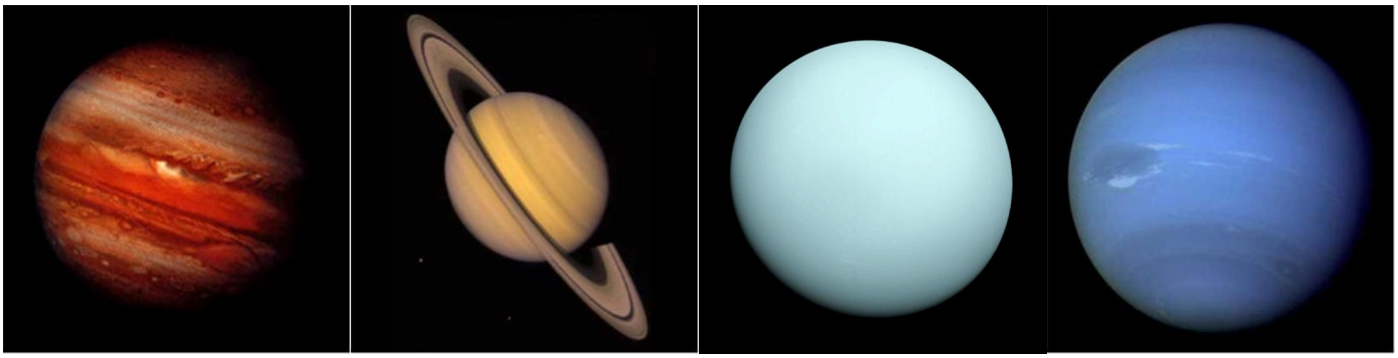
2.5.3.1: What do you think?
2.5.3.2: Gas Giant Planets
2.5.3.3: Rings
2.5.3.4: Planet Jupiter
2.5.3.5: Jupiter's Satellites
2.5.3.6: Planet Saturn
2.5.3.7: Saturn's Rings
2.5.3.8: Saturn's Satellites
2.5.3.9: Planet Uranus
2.5.3.10: Planet Neptune
2.5.3.11: Gas Giant Planets and Select Satellites Overviews

This page titled 2.5.3: The Gas Giant Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 8: The Gas Giant Planets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.3.1: What do you think?

Again, imagine you are a first-time visitor to our solar system. You come in and pass the Gas Giant planets on your way to the inner solar system and the Rocky Planets. At first glance, what differences would you see between the Gas Giant planets and the Rocky Planets?



JupiterPublic, Saturn, Uranus, Neptune Domain | Image courtesy of NASA.

This page titled 2.5.3.1: What do you think? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 8.1: What do you think? by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.3.2: Gas Giant Planets

Like the Rocky Planets, the Gas Giant planets – sometimes referred to as the **Jovian Planets** (Jupiter-like) – also have a number of general characteristics. Gas Giant planets are large in size when compared to the Rocky Planets.

Gas Giant planets are made up of primarily Gases – thus their name – with basically the same compounds and elements, but differing percentages. These worlds are not primarily composed of rock or other solid matter as the Rocky Planets.

Some other Gas Giant characteristics include short revolution periods (when compared to the Rocky Planets), many moons, rings, and farther from the Sun in our solar system thus the longer periods of revolution.

This page titled 2.5.3.2: Gas Giant Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 8.2: Gas Giant Planets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.3.3: Rings

One characteristic we will find with the Gas Giant planets is rings. A **ring** is a disk of dust or small object orbiting a planet or other body. It is hypothesized that rings can form in one of three ways.

1. Material from the solar system's original protoplanetary disk that was within the planet's Roche limit, thus not forming a moon.
    - The **Roche limit** is the minimum distance to which a satellite can approach its primary body without being torn apart by tidal forces due to gravity.
2. A moon that broke apart due to Roche limit tidal stresses.
3. Debris from the moon that broke up due to a large impact.

Different forces can influence rings and ring structures. These include other rings, satellites orbiting the planet, and the planet itself.

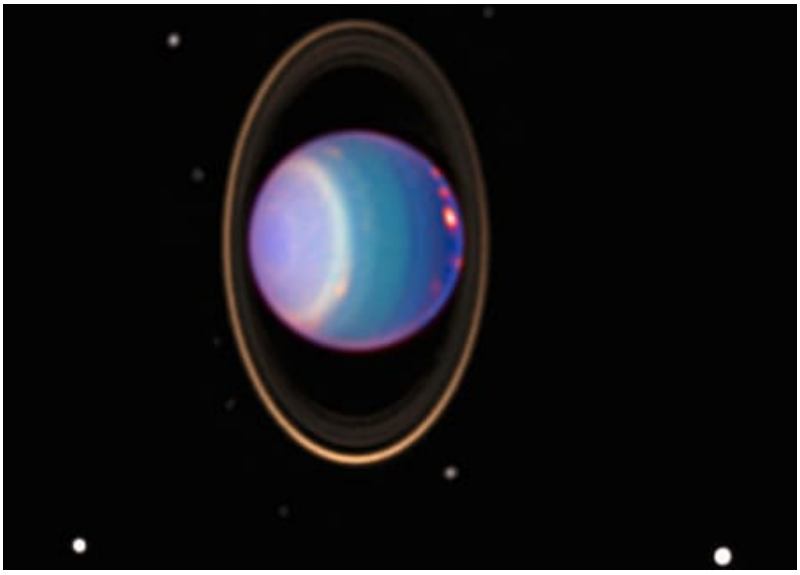
There are some objects that have rings but are not Gas Giants. For example, one asteroid has been discovered with rings; this is thought to be due to a collision or collisions. The Martian moon Deimos will eventually break up due to the gravitational tidal forces of Mars and Deimos' orbit getting too close to Mars; one hypothesis is that a ring will be formed from Deimos around Mars at that time.

Image of Enhanced color image of Saturn's rings.

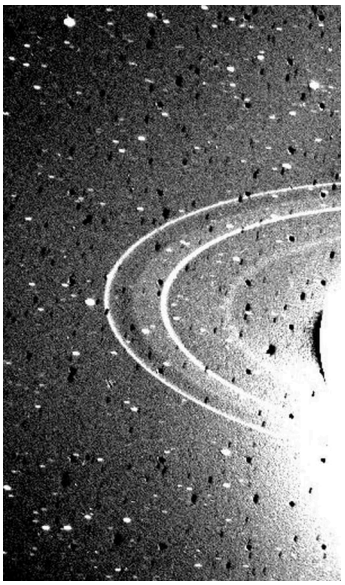
Saturn's rings via NASA VoyagerPublic Domain | Image courtesy of NASA.

Image of Black and white image capturing Planet Jupiter's single ring.

Jupiter's single ringPublic Domain | Image courtesy of NASA.



Uranus and some of its rings via NASA Hubble Space TelescopePublic Domain | Image courtesy of NASA.



Neptune’s rings via NASA Voyager 2Public Domain | Image courtesy of NASA.

This page titled 2.5.3.3: Rings is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 8.3: Rings by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.3.4: Planet Jupiter

Jupiter is the largest planet in the solar system; all of the other planets could fit inside Jupiter. One of its major atmospheric features is the **Great Red Spot** – cyclonic in nature and twice Earth's diameter. The Great Red Spot was first seen by Galileo in 1610 and has been seen to change shape and color over the past 400 years.

Lightning and thunderstorms have been observed in Jupiter's upper atmosphere.

Jupiter has some other interesting features, including:

- It gives off more heat than it receives.
- It has a small, dense rocky core surrounded by hydrogen and some helium even though it is primarily a Gas Giant planet. Some speculate the core is solid hydrogen.
- It is a natural radio source, like the Sun and other stars.
- It has aurora events, much like Earth.

Jupiter at a Glance

Characteristics to Compare

Atmosphere

- Hydrogen (H_2), Helium (H_e), Methane (CH_4), Ammonia (NH_3), Water (H_2O)

Rotation

- Short rotation — 9. 8 hours

Revolution

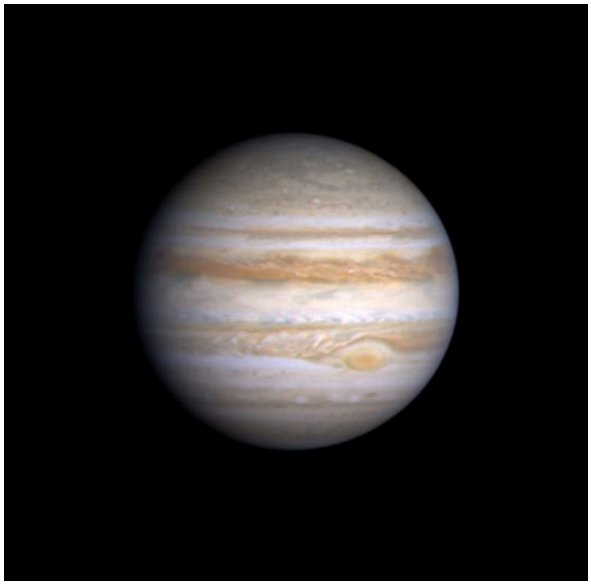
- Long revolution — 11. 86 years

Rings

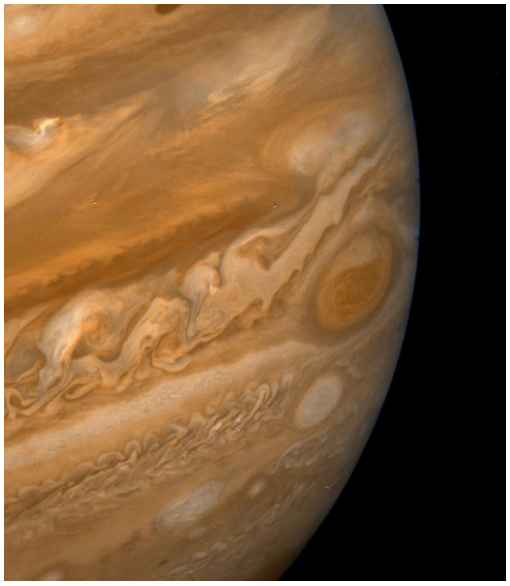
- One thin ring detected by the Voyager 2 spacecraft

Magnetic Field

- A very strong magnetic field; 20,000 times that of Earth's



The planet Juptier via NASA Cassini SpacecraftPublic Domain | Image courtesy of NASA.



Jupiter's Great Red Spot via NASA Voyager 2Public Domain | Image courtesy of NASA.



Jupiter’s Great Red Spot with turbulent cloud tops via NASA Galileo SpacecraftPublic Domain | Image courtesy of NASA.

Image showing lightning in Jupiter’s upper atmosphere. These mosaics of Jupiter's night side show the Jovian aurora at approximately 45 minute intervals as the auroral ring rotated with the planet below the spacecraft. The images were obtained by the Solid State Imaging (SSI) system on NASA's Galileo spacecraft. during its eleventh orbit of Jupiter. The auroral ring is offset from Jupiter's pole of rotation and reaches the lowest latitude near 165 degrees west longitude. The aurora is hundreds of kilometers wide, and when it crosses the edge of Jupiter, it is about 250 kilometers above the planet.

Lightning in Jupiter’s upper atmospherePublic Domain | Image courtesy of NASA.

This page titled 2.5.3.4: Planet Jupiter is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 8.4: Planet Jupiter by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.3.5: Jupiter's Satellites

Jupiter has four Galilean moons. These are the moons that Galileo studied with his first built telescope. They are, from biggest to smallest:

- Ganymede
  - Largest and most massive moon of the solar system.
  - Orbits Jupiter roughly every 7 days.
  - Has an internally generated magnetic field.
  - Has an underground ocean.
  - Tidally locked with Jupiter, which means it keeps the same face towards the planet at all times.
- Callisto
  - Third largest moon of the solar system.
  - Orbits Jupiter rounding every 16.5 days.
  - Composed about equally of rock and ice.
  - Most heavily cratered surface in the solar system.
  - Tidally locked with Jupiter.
- Io
  - Innermost of the four Galilean moons.
  - Orbits Jupiter roughly every 42.5 hours.
  - The most volcanically active body in the solar system.
  - Yellow in color due to sulfur on surface.
  - Tidally locked with Jupiter.
- Europa
  - Slightly larger than Earth's Moon.
  - Orbits Jupiter roughly every 3.5 days.
  - Liquid water ocean under a layer of frozen surface.
  - Exhibits water plumes, or water geyser-like activity.
  - Tidally locked with Jupiter.



From left to right: Ganymede, Callisto, Io, and Europa. Images courtesy of NASA.

This page titled 2.5.3.5: Jupiter's Satellites is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.3.6: Planet Saturn

Saturn is "the ringed planet. " Although all four Gas Giant planets have rings, Saturn's are by far the most-spectacular and intricate. Saturn's upper atmosphere is not as turbulent when compared to Jupiter's upper atmosphere; it does not show as many markings and spots as Jupiter. The atmosphere shows an odd polar hexagon, due to Saturn's atmospheric jet stream. There are some other interesting Saturnian characteristics. First, the density of Saturn is so low that it would float in water – *if you could find a bathtub big enough to hold Saturn.* Much like Jupiter, Saturn probably has a very small rocky core surrounded by hydrogen and helium. And also like Jupiter, Saturn also experiences aurora and lightning.

Saturn at a Glance

Characteristics to Compare

Atmosphere:

- Predominantly hydrogen and helium

Rotation:

- 10.5 hours

Revolution:

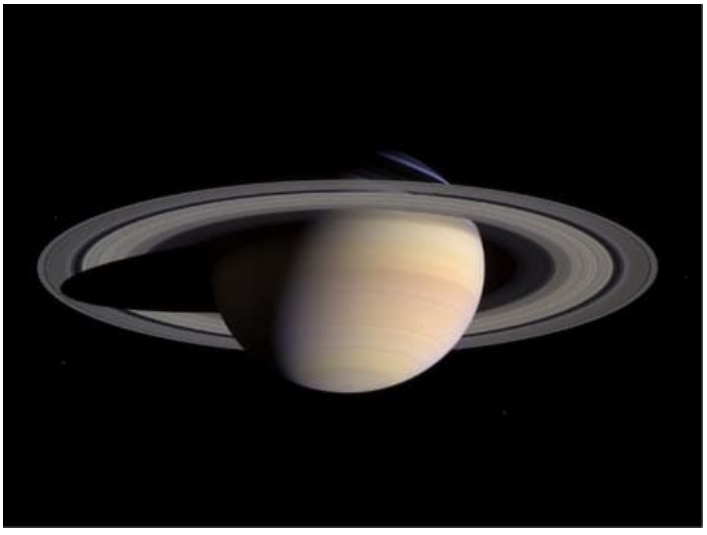
- 29.5 years

Rings:

- Extensive ring system

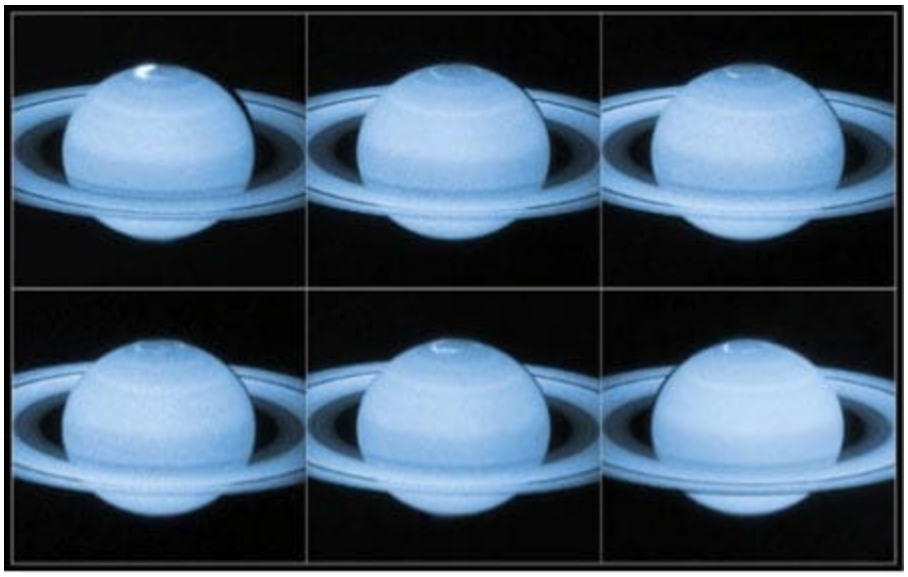
Magnetic Field:

- Yes



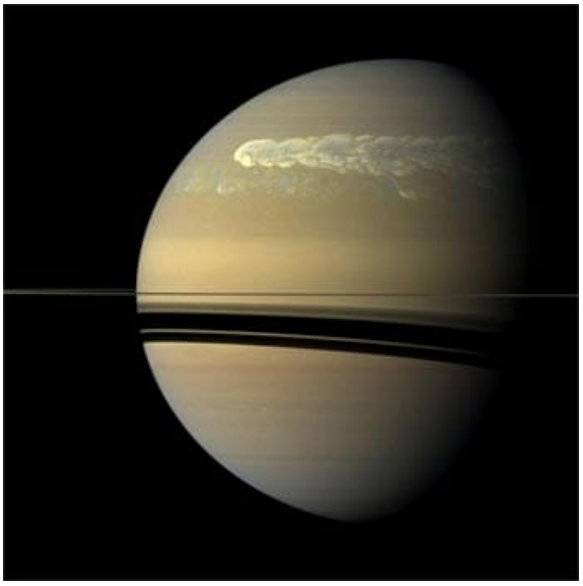
Saturn as imaged by Cassini from below the rings.

Image courtesy of NASA



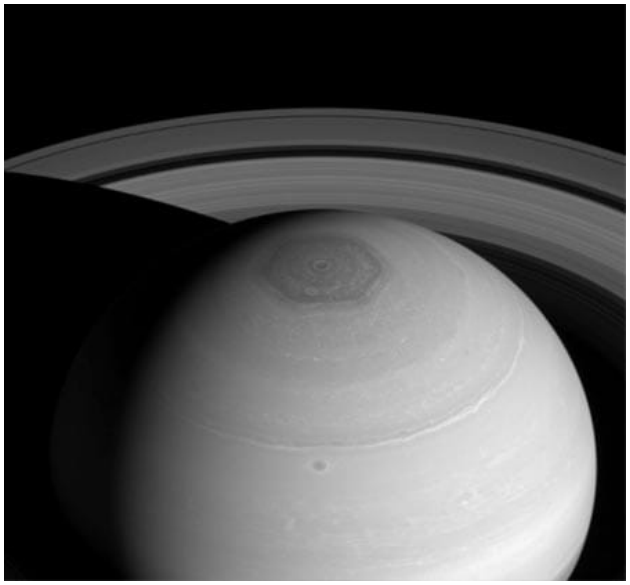
Saturn as imaged with a high pass filter by the Hubble Space Telescope. Aurora can be seen at the pole.

Image courtesy of NASA



Saturn's atmosphere as imaged by Cassini.

Image courtesy of NASA



North Pole of Saturn showing its hexagonal shape, as imaged by Cassini.

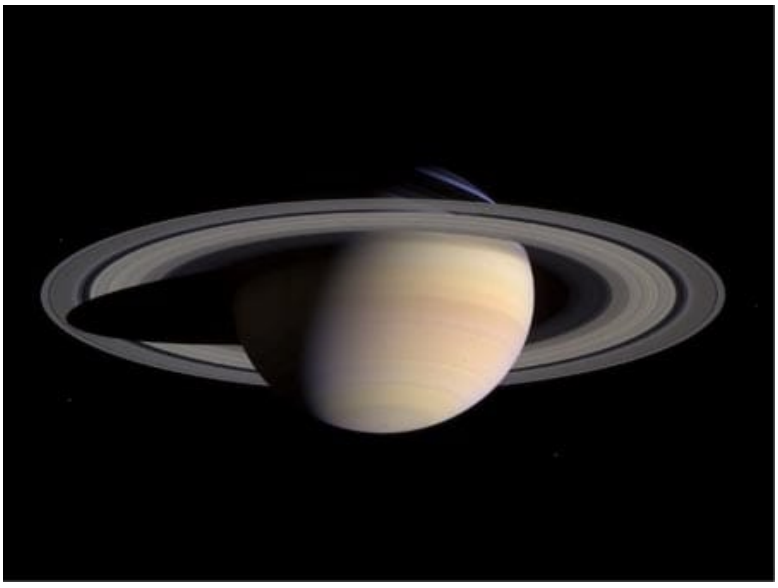
Image courtesy of NASA

This page titled 2.5.3.6: Planet Saturn is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

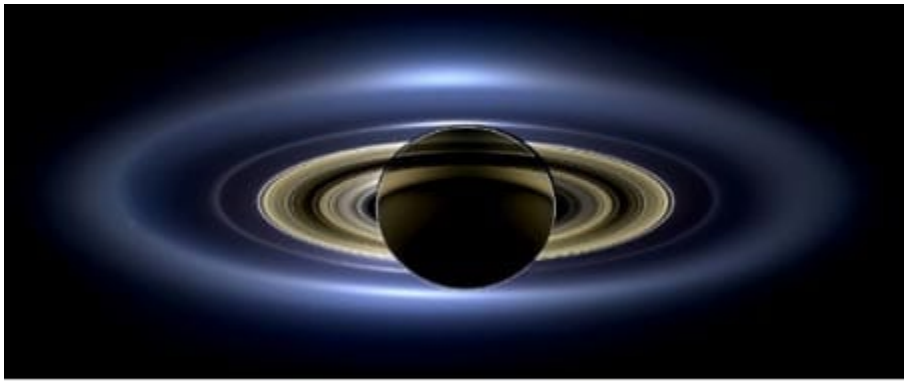
2.5.3.7: Saturn's Rings

Saturn's rings are made up of particles that range from dust-size to meters in diameter. The rings are mostly water and ice in composition, with some rocky silicate materials. Galileo Galilei first observed them in 1610; he could not discern them as rings and referred to them as the *ears of Saturn* . In 1655, astronomer Christian Huygens hypothesized that Saturn had a ring, based on his telescopic observations. Giovanni Cassini suggested in 1655 that Saturn's ring was actually made up of numerous smaller rings with gaps or divisions between these smaller rings. Saturn exhibits seven major rings with multiple divisions and gaps in the rings. The outer F ring is twisted and knotted like a piece of taffy; this is due to interactions of the F ring with Saturn's moon Prometheus. There are small moons, called **shepherd** or **sheepherder moons**, which orbit near the outer edges of a ring or within gaps in the rings.

These sheepherder moons maintain a sharply defined edge to the ring. Radial features in the B ring, called spokes, were suggested by Steve O'Meara (Earth observations) and confirmed by spacecraft. These **spokes**, seasonal in nature, appear to be caused by electrostatic forces. A quite-distant ring of dust was discovered in 2009, called the Phoebe ring, and the Saturnian moon Phoebe that is not visible in Earth-based telescopes.



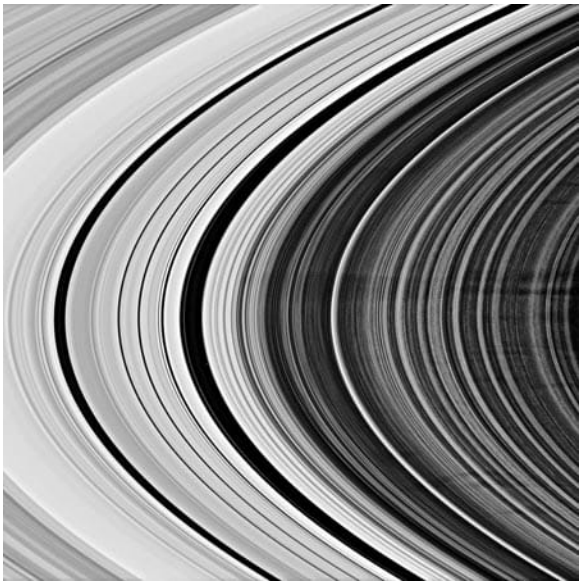
Saturn's rings via NASA Hubble Space TelescopePublic Domain | Image courtesy of NASA.



Saturn's rings, with Saturn blocking the Sun via NASA Cassini SpacecraftPublic Domain | Image courtesy of NASA.



Saturn's F ring, with the moon Prometheus pulling material from the ring via NASA Cassini SpacecraftPublic Domain | Image courtesy of NASA.



Spokes in Saturn's B ring via NASA Cassini SpacecraftPublic Domain | Image courtesy of NASA.

This page titled 2.5.3.7: Saturn's Rings is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 8.7: Saturn's Rings by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

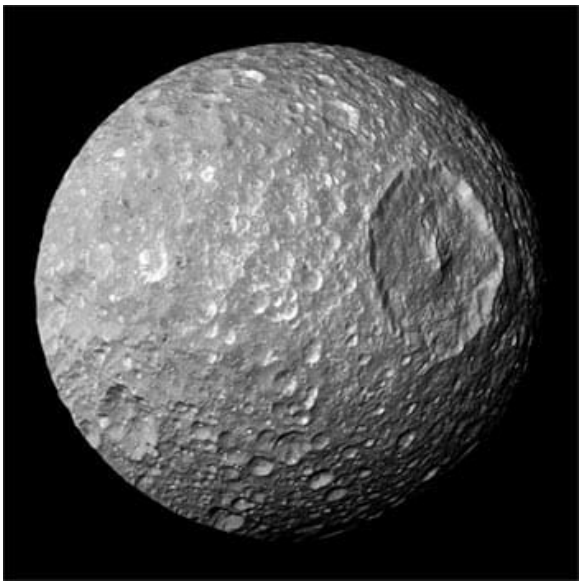
2.5.3.8: Saturn’s Satellites

While Saturn has one significant moon, Titan, several others are worth interest and were explored by the Cassini mission.

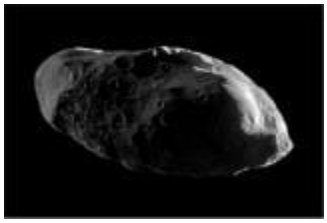
- Enceladus
  - Sixth largest moon of Saturn.
  - Most reflective body in the solar system due to snow-covered surface.
  - Orbits Saturn in just over a day.
  - Exhibits hydrothermal activity.
  - Tidally locked with Saturn.
- Mimas
  - Innermost of Saturns major moons.
  - Heavily cratered with one large, prominent crater, named Herschel after the moon's discoverer, as a distinguishing feature.
  - Mimas is responsible for clearing the material from the Cassini Division, the gap between Saturn's two widest rings, the A Ring and B Ring.
  - May have a subsurface ocean, but more observation, exploration, and research is needed.
  - Tidally locked with Saturn.
- Prometheus
  - Most likely composed of ice.
  - Is a shepherd satellite for the inner edge of Saturn's narrow F Ring.
  - Orbits Saturn in roughly 14 hours.
  - Believed to be tidally locked with Saturn.
- Titan
  - Largest moon of Saturn.
  - Second largest moon of solar system (behind Jupiter's Ganymede).
  - Has a thick atmosphere of nitrogen and methane.
  - Rains liquid methane.
  - Has hydrocarbon lakes.
  - Huygens probe landed on its surface; only other moon in the solar system we have landed on.



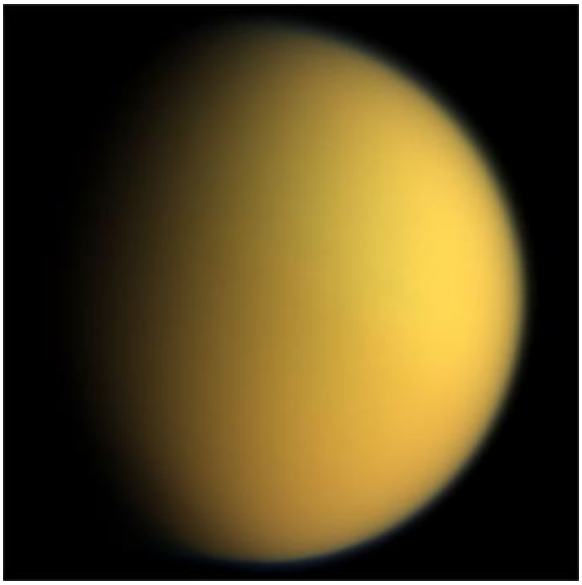
Enceladus as imaged by Cassini. Image courtesy of NASA.



Mimas as imaged by Cassini showing prominent crater. Image courtesy of NASA.



Prometheus as imaged by Cassini. Image courtesy of NASA.



Titan as imaged by Cassini. Image courtesy of NASA.



Surface of Titan as imaged by the Huygens probe. Image courtesy of NASA.

This page titled 2.5.3.8: Saturn's Satellites is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.3.9: Planet Uranus

Uranus was the first planet found with a telescope. Uranus was discovered by Sir William Herschel in 1781, even though others might have recorded Uranus prior to Herschel and not realized what they were seeing. Uranus is often referred to as Neptune's twin, because of similarities in their sizes, colors, and other characteristics.

Uranus rotates on its side. It is believed that this odd rotation is due to a collision with its moon Miranda. The planet shows little atmospheric activity, and actually looks like a green billiard ball. If Uranus has a core at all, it is a very small rocky core.

In its atmosphere, Uranus has more water, ammonia, and methane ices in its atmosphere than Jupiter and Saturn; so it is often referred to as an **Ice Giant**. It is the coldest planet in the solar system. Like Jupiter and Saturn, Uranus also exhibits aurora activity. The planet has 13 rings, composed of dark particles made of dust to less than a meter in diameter. Uranus has 27 known moons (as of August 2014); not as many as Jupiter or Saturn, but definitely its own system of worlds which orbit the planet.

Uranus at a Glance

Characteristics to Compare

Atmosphere

- Hydrogen (H_2), Helium (H_e), Methane (CH_4), Ammonia (NH_3), Water (H_2O)

Rotation

- Short rotation — 17. 3 hours retrograde (backwards)

Revolution

- Long revolution — 84. 01 years

Rings

- An intricate system of rings:
  - 13 rings total; first 9 found via airborne telescope

Magnetic Field

- Yes, tilted about 45°. Reason for tilt unknown

Uranus Satellite of Distinction

- Titania
  - Largest moon of Uranus.
  - First studied in modern age by Voyager 2.
  - Orbits Uranus in about 9 days.
  - Orbit lies completely within Uranian magnetosphere.
  - Tidally locked with Uranus.



Uranus. The blue color is from the methane in the atmosphere. Image courtesy of NASA.



A black and white image taken of Titania by Voyager 2. Image courtesy of NASA.

This page titled 2.5.3.9: Planet Uranus is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.3.10: Planet Neptune

Neptune was the first planet discovered through mathematical prediction. Urbain Le Verrier and John Couch Adams predicted the position in 1846. Neptune is smaller than Uranus, but denser. Like Uranus, Neptune is also referred to as an Ice Giant, as well as Uranus' twin. It too may have a very small rocky core. Neptune also exhibits more atmospheric activity than Uranus. Its Great Dark Spot is a storm similar to Jupiter's Great Red Spot, and about the same size as Earth. Neptune also has atmospheric scooters that are white cloud groups. It was so named because it moved faster than the Great Dark Spot. Its atmosphere sustains the highest winds in the solar system, in excess of 1,200 miles per hour. Like the other Gas Giant planets, Neptune also exhibits aurora activity.

Neptune has five rings, which are composed of dark particles. The rings appear to be broken, more like arcs or ringlets. There are 14 known moons (as of August 2014).

Neptune at a Glance

Characteristics to Compare

Notable Feature

- Great Dark Spot

Atmosphere

- Primarily of hydrogen and helium with hydrocarbons (predominantly methane) and traces of nitrogen.

Rotation

- Rotates approximately every 16 hours.

Revolution

- Approximately 165 years.

Rings

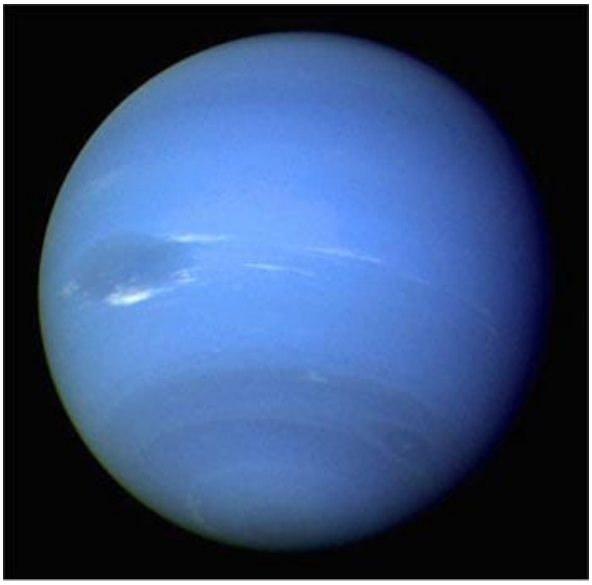
- Yes, three main rings.

Magnetic Field

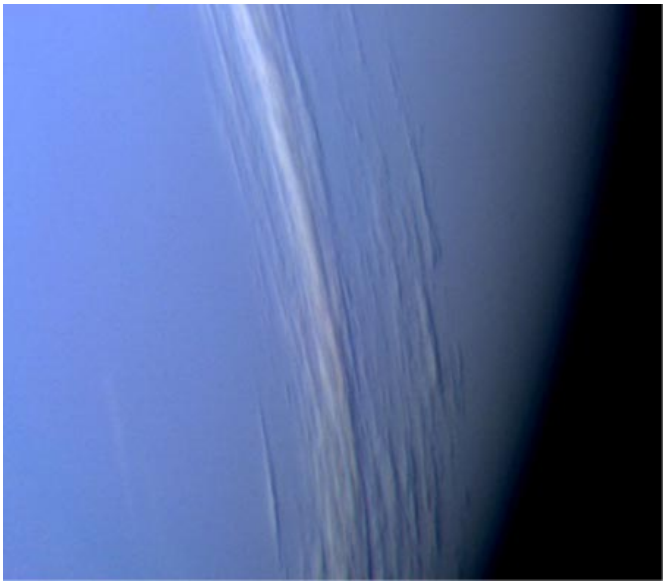
- Yes; Not aligned with rotational poles.

Neptune Satellite of Distinction

- Tritan
  - Largest moon of Neptune.
  - Orbits in a retrograde orbit, opposite to the rotation of the planet, the only large moon in the solar system to do so.
  - May have a subsurface ocean.
  - Tidally locked to Neptune.



The Great Dark Spot of Neptune. Image courtesy of NASA.



Elongated cloud formations on Neptune. Image courtesy of NASA.



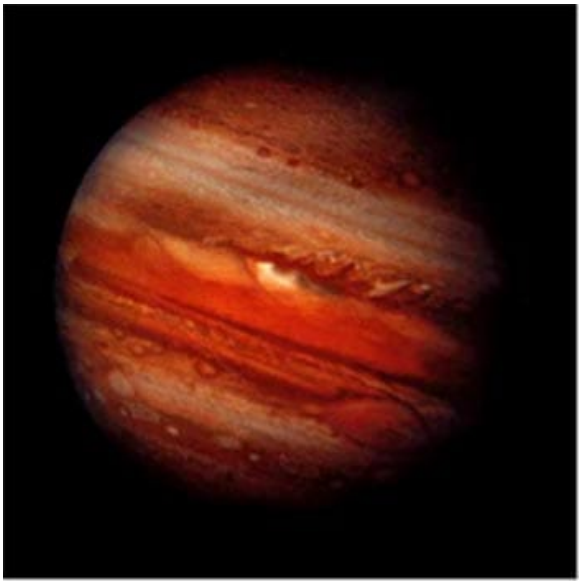
Tritan, the major moon of Neptune. Image courtesy of NASA.

This page titled 2.5.3.10: Planet Neptune is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.3.11: Gas Giant Planets and Select Satellites Overviews

Solar System's Gas Giant Planets at a Glance

Jupiter



Public Domain | Image courtesy of NASA.

Characteristics

Atmosphere

- Hydrogen (H_2), Helium (H_e), Methane (CH_4), Ammonia (NH_3), Water (H_2O)

Rotation

- Short rotation — 9. 8 hours

Revolution

- Long revolution — 11. 86 years

Rings

- One thin ring detected by the Voyager 2 spacecraft

Magnetic Field

- A very strong magnetic field; 20,000 times that of Earth's

Satellites

- 67

The Bottom Line...

- Largest solar system planet
- Great Red Spot

- Gives off more heat than receives
- Small, dense rocky core surrounded by H_2 and H_e
- Radio source
- Aurora
- Lightning and thunderstorms

Saturn

Image of Saturn.

Public Domain | Image courtesy of NASA.

Characteristics

Atmosphere

- Hydrogen (H_2), Helium (H_e), Methane (CH_4), Ammonia (NH_3), Water (H_2O)

Rotation

- Short rotation — 10. 6 hours

Revolution

- Long revolution — 29. 46 years

Rings

- An intricate system of rings:
  - Spokes
  - Twisted F ring
  - Giant outer ring

Magnetic Field

- Yes

Satellites

- 62

The Bottom Line...

- Spectacular / intricate rings
- Would float in water
- Small, dense rocky core surrounded by H2 and He
- Polar hexagon
- Aurora
- Lightning

Uranus



Public Domain | Image courtesy of NASA.

Characteristics to Compare

Atmosphere

- Hydrogen (H_2), Helium (H_e), Methane (CH_4), Ammonia (NH_3), Water (H_2O)

Rotation

- Short rotation — 17. 3 hours retrograde (backwards)

Revolution

- Long revolution — 84. 01 years

Rings

- An intricate system of rings:
  - 13 rings total; first 9 found via airborne telescope

Magnetic Field

- Yes, tilted about 45°. Reason for tilt unknown

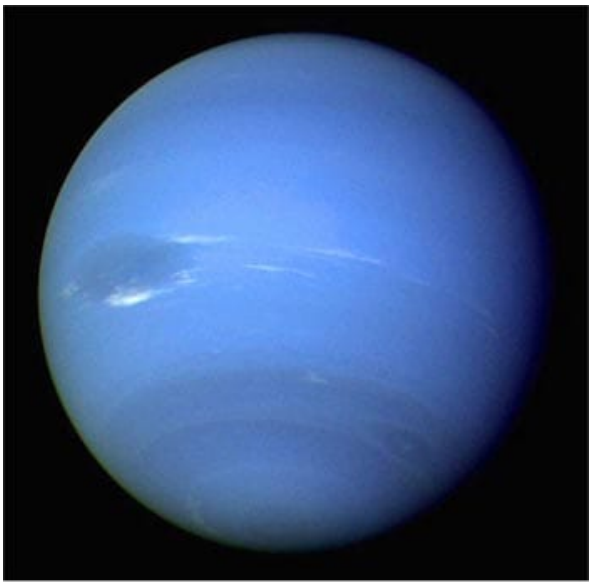
Satellites

- 27

The Bottom Line...

- 1st planet discovered with a telescope
- *Ice Giant*
- Very small rocky core
- Little atmospheric activity
- Looks like a billiard ball
- Aurora
- Neptune's twin

Neptune



Public Domain | Image courtesy of NASA.

Characteristics to Compare

Atmosphere

- Hydrogen (H_2), Helium (H_e), Methane (CH_4), Ammonia (NH_3), Water (H_2O)

Rotation

- Short rotation — 16. 1 hours

Revolution

- Long revolution — 164. 8 years

Rings

- A partial, twisted series of 5 rings confirmed by the Voyager 2 spacecraft

Magnetic Field

- Yes, tilted about 45°. Reason for tilt unknown

Satellites

- 14

The Bottom Line...

- 1st planet predicted position mathematically
- *Ice Giant*
- Very small rocky core
- Great Dark Spot, Scooters
- Looks like a billiard ball
- Aurora
- Uranus' twin

Consider this...

We consider the Gas Giant planets to be huge, especially Jupiter. But when compared to other stellar systems, they are small – even "king of the planets," Jupiter. Many of these hot Jupiters are much closer to their star than our Gas Giant planets and dwarf Jupiter in size. And it is theorized that some of these so-called "Super Jupiters" might be failed stars, that is, not quite enough mass for the nuclear fusion process. Nuclear fusion is where two particles collide at high speed and form a particle of a higher mass and a lot of energy.



Public Domain | Image courtesy of NASA.

This page titled 2.5.3.11: Gas Giant Planets and Select Satellites Overviews is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 8.11: Gas Giant Planets and Select Satellites Overviews by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4: MINOR BODIES OF THE SOLAR SYSTEM

Learning Objectives

Upon completion of this module, the student will be able to:Upon completion of this module, the student will be able to:

- Differentiate between the types of minor bodies found in our solar system
- List the characteristics of asteroids, including Near-Earth Asteroids (NEOs) and Potentially-Hazardous Asteroids (PHAs)
- Differentiate among meteoroids, meteors, and meteorites
- List the characteristics of three types of meteorites
- List the characteristics of comets
- List the characteristics of centaurs
- Explain the discoveries of Pluto and Eris
- List the characteristics of the Trans-Neptunian objects, Kuiper Belt and the Oort Cloud

This module explores the minor bodies of the solar system: asteroids, comets, and dwarf planets, as well as meteors and meteorites.

2.5.4.1: Trans-Neptunian objects, the Kuiper Belt, and the Oort Cloud

2.5.4.2: The Oort Cloud and Kuiper Belt

2.5.4.3: What do you think?

2.5.4.4: Minor Bodies

2.5.4.5: Asteroids

2.5.4.6: Brightest Asteroids Visible from Earth

2.5.4.7: Asteroid Classification

2.5.4.8: Meteoroids, Meteors, and Meteorites

2.5.4.9: Meteorites

2.5.4.10: The Impact of Meteorites

2.5.4.11: Comets

2.5.4.12: Comet Characteristics

2.5.4.13: Images of Comets

2.5.4.14: Comets in History

2.5.4.15: Scientific Advances

2.5.4.16: Halley's Comet

2.5.4.17: A Comet Impacts a Planet

2.5.4.18: The Centaurs- Are they Asteroids or Comets?

2.5.4.19: Pluto

2.5.4.20: Eris

This page titled 2.5.4: Minor Bodies of the Solar System is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9: Minor Bodies of the Solar System by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.1: Trans-Neptunian objects, the Kuiper Belt, and the Oort Cloud

Trans-Neptunian objects (TNO) are any solar system minor planet that orbits the sun at a greater average distance than Neptune. Pluto is now considered a TNO, as is Eris. As of July 2014, over 1,500 trans-Neptunian objects have been cataloged and of these, some 200 have been designated as dwarf planets. From Earth, astronomers study TNO heat emissions, colors, and spectra. The **Kuiper Belt** is a region beyond the orbit of Neptune at 30 Astronomical Units (AU) to about 50 AU from the sun. It is sometimes called the **Edgeworth–Kuiper Belt**. The Kuiper Belt is much larger than the Asteroid Belt. It's about twenty times as wide and twenty to two hundred times as massive. Kuiper Belt objects, called **KBOs**, are composed of rock and metal, like the asteroids, but also frozen ices like ammonia, methane, and water. Trans-Neptunian objects are Kuiper Belt objects (KBOs), but KBOs are not TNOs because the distance range of KBOs from the sun is much farther out in the solar system.

At least two moons – Saturn's Phoebe and Neptune's Triton – are believed to be captured KBOs. To date, over 1,000 KBOs have been discovered and around 100,000 KBOs are suspected to exist.

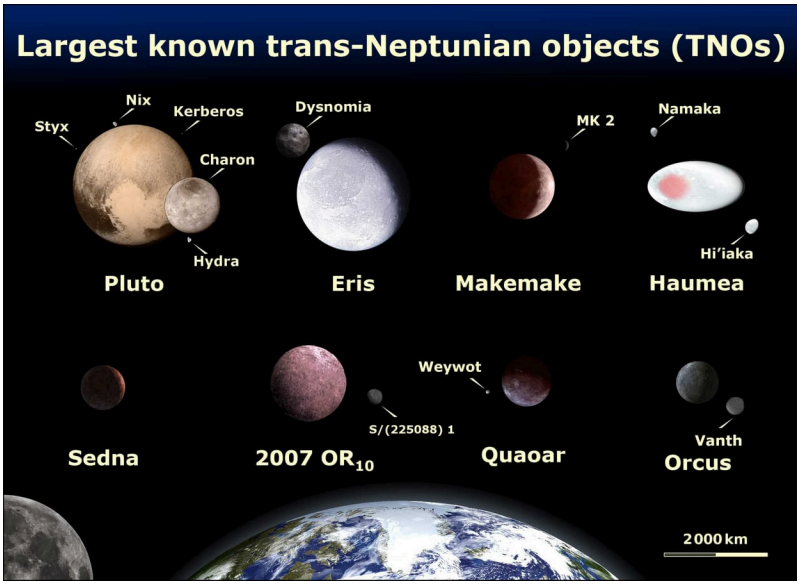
KBOs are fragments from the original solar system's protoplanetary disc that did not combine to form the more-massive and larger bodies. Astronomers believe that belts of fragments like the solar system's Kuiper Belt are present around other stellar systems being studied.

This page titled 2.5.4.1: Trans-Neptunian objects, the Kuiper Belt, and the Oort Cloud is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.1: Trans-Neptunian objects, the Kuiper Belt, and the Oort Cloud by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.2: The Oort Cloud and Kuiper Belt

The **Oort Cloud** is a hypothesized spherical cloud of icy objects up to 50,000 AU from the sun. This spherical cloud is the primary source of long-period comets. No Oort Cloud objects have been found, to date. Any object within the Oort Cloud would be called an **Oort Cloud Object, (OCO)**. The Oort Cloud would represent the physical boundary of the solar system.



CC BY-SA 3.0 | Image courtesy of NASA and Wikimedia Author: Lexicon.

This page titled 2.5.4.2: The Oort Cloud and Kuiper Belt is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.2: The Oort Cloud and Kuiper Belt by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.3: What do you think?

Ancient people feared comets. Knowing nothing about science or basic astronomy, can you imagine seeing this object—Comet Hale-Bopp, suddenly appear in the skies, larger and brighter than anything you have ever seen? How would the appearance of a bright comet have possibly affected you?

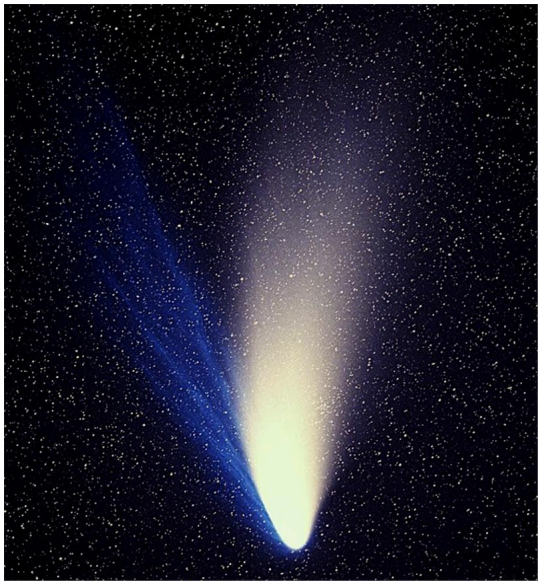


Image courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

This page titled 2.5.4.3: What do you think? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.4.4: Minor Bodies

Minor Bodies are the smaller, often non-spherical objects found in the solar system. The minor bodies include the asteroids, comets, meteoroids, the groups of objects referred to as the Kuiper Belt Objects, Trans-Neptunian Objects, and Oort Cloud Objects, and other space junk and dust. Minor bodies are significantly smaller in size when compared to the sun and planets. Each of these objects usually has its own orbit, is mostly irregular in shape, and was formed from leftover solar nebula material.

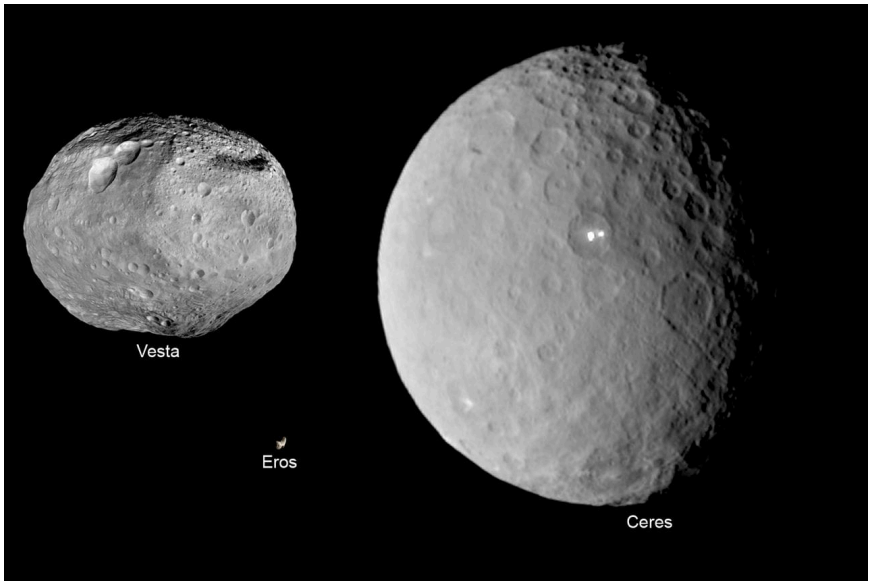
Minor Planets specifically refer to the dwarf planets, asteroids, centaurs, Kuiper belt objects, and other trans-Neptunian objects.

This page titled 2.5.4.4: Minor Bodies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.4: Minor Bodies by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.5: Asteroids

The **asteroids** are a group of the solar system's minor bodies that are much like the Rocky Planets in composition: carbon-rich, silicate materials (rock), and/or metal. The word Aster-oid means star-like. Asteroids are also called **minor planets** (not to be confused with dwarf planets). The larger asteroids are called **planetoids**. Italian astronomer Giuseppe Piazzi discovered the first asteroid, Ceres, in 1801. Ceres, which orbits between Mars and Jupiter, was first thought to be a planet. Later that same year, Pallas was discovered in basically the same orbit as Ceres. And by 1807 two additional asteroids had been discovered: Pallas and Vesta. At this time, some 625,000 asteroids have been identified, and astronomers are finding new asteroids all of the time. Most of the asteroids orbit between Mars and Jupiter, an area called the **asteroid belt**.



Public Domain

Ceres is the largest asteroid found to date, and was reclassified as a Dwarf Planet in 2006. Ceres appears to be fairly spherical in shape, which is unusual for asteroids in general. Ceres is 950 kilometers or 590 miles in diameter, and makes up a third of the mass of the asteroid belt.

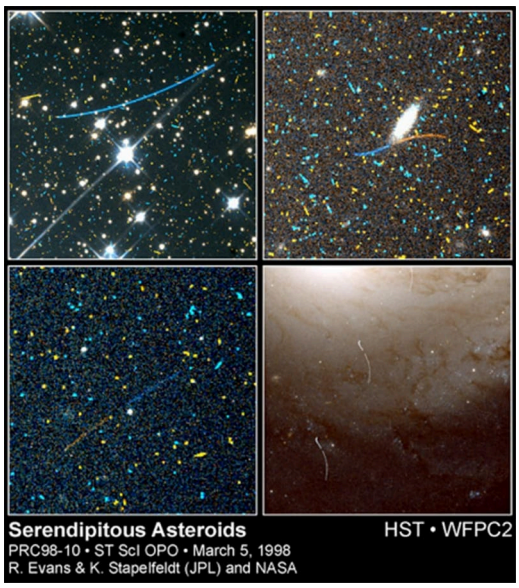
There are a number of interesting **asteroid characteristics**. Asteroids are most-likely leftover material from the formation of the solar system. Most asteroids are quite irregular in shape, with sizes ranging from around 0. 6 mile to 590 miles (Ceres). Asteroids can have satellites. Ida, with its moon Dactyl, was the first discovered pair. Currently more than 150 asteroids are known to have a small companion satellite; some have two moons.

The asteroids are the parent body sources for most meteorite impacts on Earth and on other solar system bodies. Ninety-five percent of the original asteroids are no longer around; most were impactors. An **impactor** is an object that strikes, or impacts, another object. Asteroid impacts cause **catastrophism** : the sudden shaping of a planet's or satellite's surface due to an asteroid impact.

This page titled 2.5.4.5: Asteroids is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.5: Asteroids by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.6: Brightest Asteroids Visible from Earth



Public Domain | Image courtesy of NASA.

(4) Vesta

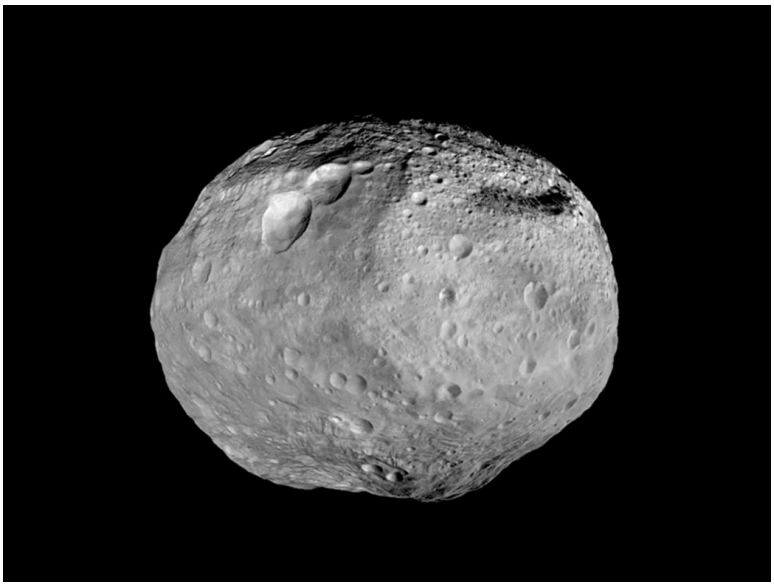


Image of the asteroid (4) VestaPublic Domain | Image courtesy of NASA/JPL-Caltech/UCAL/MPS/DLR/IDA.

Maximum Visual Magnitude

- 5. 1

Diameter in Miles

- 326

(2) Pallas

Maximum Visual Magnitude

- 6. 4

Diameter in Miles

- 338

(1) Ceres [Dwarf Planet]

Maximum Visual Magnitude

- 6. 7

Diameter in Miles

- 590

(7) Iris

Maximum Visual Magnitude

- 6. 7

Diameter in Miles

- 149 x 124 x 124

(433) Eros

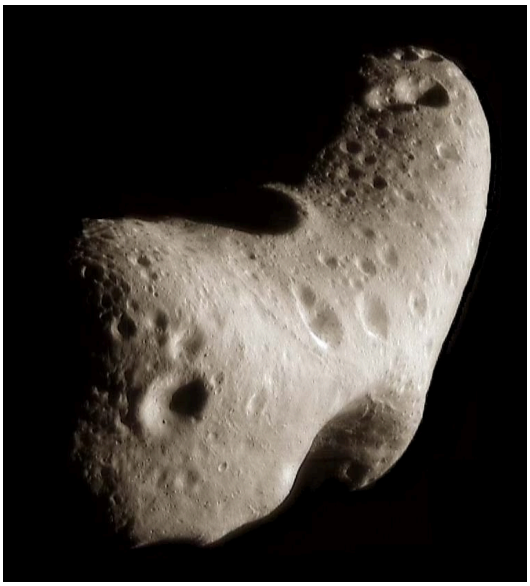


Image of the asteroid (433) ErosPublic Domain | Image courtesy of NASA.

Maximum Visual Magnitude

- 6. 8

Diameter in Miles

- 21 x 7 x 7

This page titled 2.5.4.6: Brightest Asteroids Visible from Earth is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.6: Brightest Asteroids Visible from Earth by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.7: Asteroid Classification

There are three major groups of asteroids, based on where they orbit in the solar system.

Main Asteroid Belt

Asteroids orbiting between Mars and Jupiter. This contains the majority of known asteroids and is estimated to have 1 to 2 million asteroids and a significantly higher number of smaller pieces (meteoroids and dust). These were most likely formed due to the gravitational attraction of Jupiter and experienced many collisions.

Trojan Asteroids

Asteroids share an orbit with a larger planet. The Trojan Asteroids do not collide with the planet because of where the Trojan asteroids orbit the sun in relation to the parent planet. Jupiter has the largest number of Trojan asteroids discovered to date.

Near-Earth Asteroids (NEAs)

Finally, there are the **Near-Earth Asteroids or NEAs** ; these asteroids pass close to Earth. About 10,000 NEAs have been discovered to date. And over 1,400 have been classified as **Potentially Hazardous Asteroids or PHAs** ; those asteroids that could pose a threat to Earth. PHAs come closer to Earth than about 5 million miles.

PHAs are being discovered all of the time, so the number of known PHAs continue to grow. None of the known potentially hazardous asteroids are on a collision course with Earth... **with one possible exception**. Astronomers employ the **Torino Scale** for categorizing the impact hazard associated with PHAs; the scale ranges from 0 to 10, with 0 being no chance and 10 being 100% chance of collision.

Asteroid 99942 Apophis was discovered in 2004, originally it was called 2004 MN4. The diameter of Apophis is approximately 325 meters or 1,066 feet. Apophis will pass dangerously close to Earth in 2029 and 2036. Observations and orbital calculations show there is no chance of impact during either of these close passages.

This page titled 2.5.4.7: Asteroid Classification is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.7: Asteroid Classification by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.8: Meteoroids, Meteors, and Meteorites

Three terms that are often mistakenly used interchangeably are meteoroids, meteors, and meteorites. **Meteoroids** are materials orbiting in space and are smaller than asteroids.

Meteors are the streaks we see in the sky, commonly called *shooting* or *falling stars* ; also can be **fireballs** and/or **bolides**. Most **meteor showers** are comet "dust" left behind as a comet orbits the sun that Earth runs into this trail of dust as we orbit the sun.

Meteorites are materials that makes it to the surface of a planet, satellite, etc.

"I would more easily believe that (a) Yankee professor would lie than that stones would fall from heaven. "

—Thomas Jefferson (1808)

This page titled 2.5.4.8: Meteoroids, Meteors, and Meteorites is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.8: Meteoroids, Meteors, and Meteorites by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.9: Meteorites

Meteorites are representatives of the early solar system. There are three types of meteorites. **Stones or Stony meteorites.** These are composed of silicates and other materials. Most stony meteorites contain metal in the form of iron and nickel. And stone meteorites are by far the most common type; some 94% of recovered meteorites are stony. Stone meteorites also include very rare types of meteorites, including lunar, Martian, and possibly even comets.

Iron meteorites, or Irons. These are composed of iron (Fe) and nickel (Ni), with some cobalt (Co) and other trace elements and minerals. Iron is the predominant metal, and the type of Iron meteorite depends on the ratio of iron to nickel. No Earth rock has iron and nickel in combination. Thus a 'rock' recovered with iron and nickel together is a meteorite. Irons are very dense meteorites, in comparison to stony meteorites.

Stony Irons. As the name indicates, they are a combination of iron and stony meteorite properties and the rarest type or class of meteorite. The **Pallasites**, a class within the Stony Irons, are simply spectacular. When sliced and polished, you can see a metal matrix of iron-nickel that surrounds crystals of olivine – peridot in gem terms.

This page titled 2.5.4.9: Meteorites is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.9: Meteorites by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.10: The Impact of Meteorites

Meteorites can cause damage upon impact, from minor to significant. Examples of three major impacts causing significant damage in the last 100 years are included below.

Tunguska

Place

- Tunguska, Krasnoyarsk | Krai, Russia

Date

- June 30, 1908

The Impact

- An air burst comet or asteroid impact damaging over some 2,150 square kilometers or 830 square miles of forested area. No meteorites were ever recovered.

Sikote-Alin

Place

- Sikhote-Alin, Primorye | Russia

Date

- February 12, 1947

The Impact

- A brighter-than-the-sun fireball announced the arrival of an air burst impactor, showering some 28 tons of meteoritic iron shrapnel on the Sikhote-Alin Mountains.

Chelyabinsk Oblast

Place

- Chelyabinsk Oblast | Russia

Date

- February 15, 2013

The Impact

- 1,491 people were injured and over 7,200 buildings damaged by the air burst from this impactor, estimated to weigh 10,000 tons; numerous stone meteorites were found.

This page titled 2.5.4.10: The Impact of Meteorites is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.10: The Impact of Meteorites by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.11: Comets

Often called dirty snowballs, **comets** are really more like snowy dirtballs. Presently, there are around 5,000 known comets; with new comets being discovered all of the time. The number of known comets is much smaller in number than the known number of asteroids. Many comets orbit in the **Kuiper Belt** and the **Oort Cloud**. Comet orbits are often highly elliptical, so much so that these orbits are sometimes described as cigar-shaped. Comet orbits around the sun can be a short period (<200 years) or a long period (>200 years).

Comets are named after their discoverers, which is a tradition that spans centuries. And hundreds of years ago some European astronomers made their living discovering comets.

We have seen a comet impact a planet; Comet Shoemaker-Levy 9 impacted Jupiter with 21 cometary fragments in July 1994. Gene and Carolyn Shoemaker and David Levy discovered the comet.

"Comets are like cats. They both have a tail and they both have a mind of their own..."

—David Levy (1994)

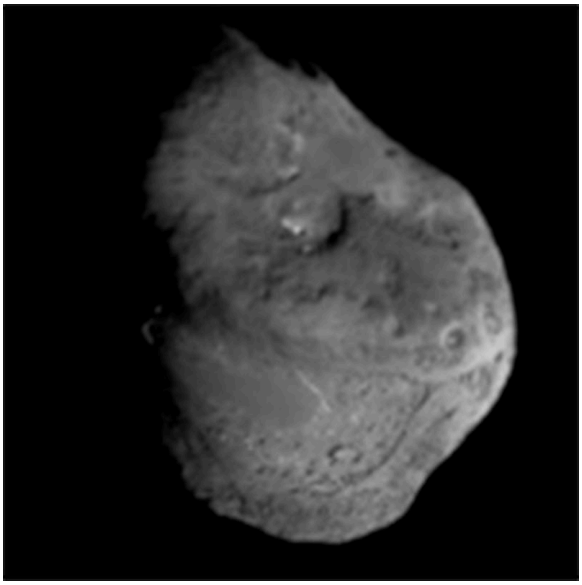
This page titled 2.5.4.11: Comets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.11: Comets by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.12: Comet Characteristics

Comets are composed of Carbon dioxide (CO2), water ices, silicates, and organic material. The materials at the surface of the comet are of a talc-like fineness, as seen up close on Comet Tempel-1. Comets are the sources for the majority of our meteor shower “material;” the fine dust comets leave behind like bread crumbs as they orbit the sun.

The comet’s **nucleus** is the solid, core structure of a comet. The **coma** or **head** is the fuzzy haze that surrounds the comet’s nucleus. The **tail** may be made of gas, dust, or both. The **coma** and **tail**are what we see from Earth. The sizes of comet nucleus have been observed up to around 30 kilometers or 20 miles in diameter. When a comet gets close to the sun, they begin to heat up and outgas, making for the coma and sometimes a tail.



Tempel-1 cometPublic Domain | Image courtesy of NASA.

The behavior of comets is often unpredictable, due to their interactions with the sun, and other solar system dynamics. Comets, which have made numerous passages around the sun, appear to produce less dust and gas, thus less-impressive tails. Often first-time inner solar system comets can break up when passing near the sun, which was the unfortunate fate of Comet ISON in 2013.

The **true nucleus** of a comet has been seen six times: Halley (1986), Borrelly (2001), Wild-2 (2004), Tempel-1 (2004 and 2005), Hartley-2 (2010), and Churyumov-Gerasimenko (2014). If all goes according to plan, the European Space Agency’s Rosetta spacecraft will release a lander, called Philae, in November 2014. Philae will land on Comet Churyumov-Gerasimenko, providing us with a wealth of data.

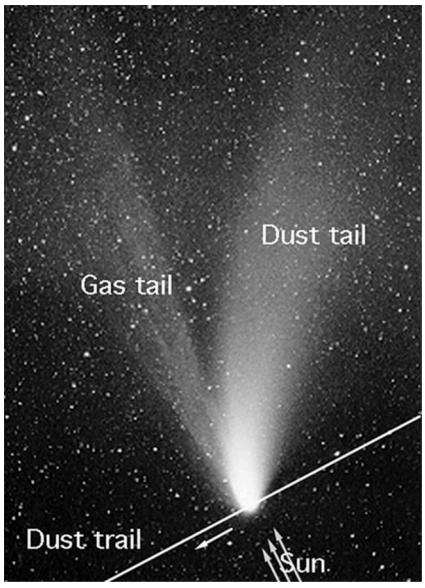
This page titled 2.5.4.12: Comet Characteristics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.12: Comet Characteristics by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.13: Images of Comets



Public Domain | Image courtesy of NASA.



Public Domain | Image courtesy of NASA.



CC SY-SA 3. 0 | Image courtesy of Wikimedia Author: Soerfm.

Image of Collision of Comet Tempel-1's nucleus and the Deep Impact probe.

Public Domain | Image courtesy of NASA.

This page titled 2.5.4.13: Images of Comets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.4.14: Comets in History

Comets have often been written about and portrayed in history. Some of the earliest comet observations were recorded by the Chinese on oracle bones. Oracle bones were pieces of turtle shells and bones used for a form of divination. Aristotle thought and proposed that comets were an atmospheric phenomena, in other words, not in our solar system or in space.

Comets were considered bad omens until the 16th century. These omens would include crop failures, diseases, deaths of royalty, or other such catastrophic events. Pliny the Elder (23-79 AD), a Roman natural philosopher, thought comets were the cause of political issues and death. Halley's Comet made a passage at the same time during the Battle of Hastings in 1066.

This page titled 2.5.4.14: Comets in History is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.14: Comets in History by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.15: Scientific Advances

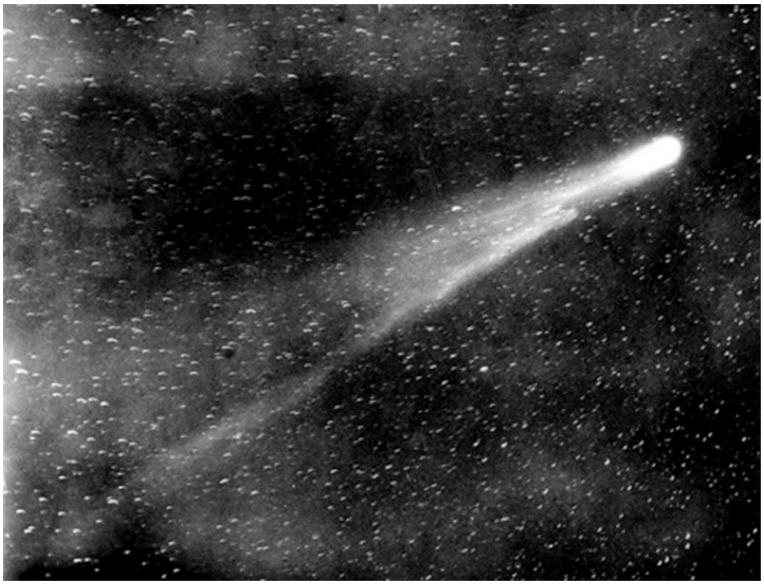
In the more-advanced and scientific approaches to comets, Tycho Brahe measured the parallax of a comet in 1577, which showed it was outside the atmosphere of the Earth. Isaac Newton derived his Law of Universal Gravitation, for which he showed the occasional parabolic orbits of objects. He was able to correlate a parabolic orbit to a comet in 1680. Edmund Halley applied Newton's Law of Universal Gravitation to a series of comet appearances over some 350 years. Three of these comet appearances had similar orbits; Halley hypothesized it was the same comet. Based on this, Halley predicted the comet would return in 1758 or 1759; the comet did appear as Halley predicted and was named Halley's Comet.

This page titled 2.5.4.15: Scientific Advances is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.4.16: Halley's Comet

Perhaps the best-known comet, Halley's Comet, has been a regular visitor through the solar system for thousands of years. Halley is pronounced like Valley ... hæli. The comet has been observed and recorded since at least 240 BC. Records of Halley's passages were made by the Chinese, Babylonians, and medieval Europeans. Recall, Edmund Halley was the first astronomer to determine that Halley's comet was the same comet that passed by Earth every 75-76 years. The 1910 Earth passage of Halley's Comet brought much scientific revelation, and much fear. Cyanogen gas was discovered by astronomers during the 1910 passage, leading to mass panic among the public. People were selling comet pills, comet umbrellas, and gas masks.

Halley's Comet was the first comet to be visited by spacecraft, in 1986; Vega 1 (USSR) and Giotto (European Space Agency). And, interestingly, it turns out that in the last 2,000 years the least-visible or poorest passage of Halley's Comet to Earth was in 1986. The next close passage of Halley's Comet will be July 2061; so mark your calendars!



Halley's Comet in *1910*Public Domain

This page titled 2.5.4.16: Halley's Comet is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.16: Halley's Comet by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.17: A Comet Impacts a Planet

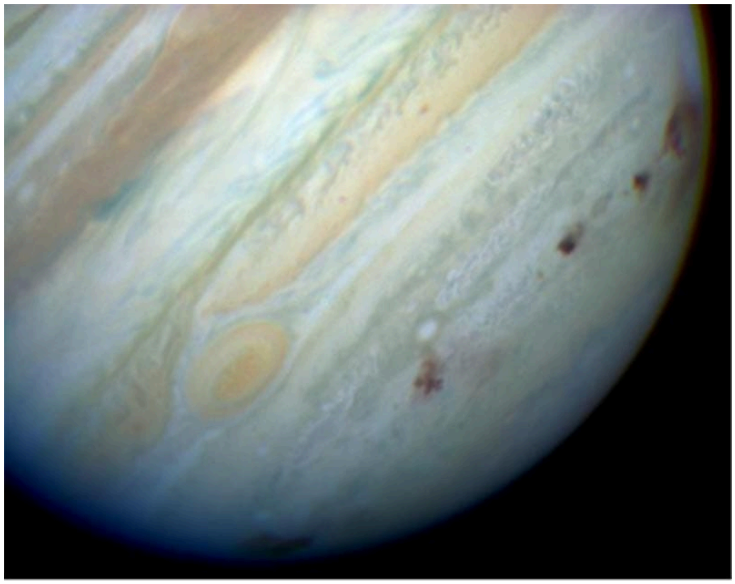
Except for seeing meteors enter our atmosphere, no contemporary astronomer had ever observed an impact on another world until 1994. **Comet Shoemaker–Levy 9, S-L 9,** was discovered in March 1993 by astronomers Eugene and Carolyn Shoemaker and David Levy. Carolyn's first impression of the discovery photo was it "looked like a squashed comet. " This description was due to the fact that the comet had been broken apart by the gravitational pull of Jupiter. The largest cometary fragment was 2 kilometers or 1. 2 miles across.

A NASA Hubble Space Telescope (HST) image of comet Shoemaker-Levy 9, taken on May 17, 1994, with the Wide Field Planetary Camera 2 (WFPC2) in wide field mode. When the comet was observed, its train of 21 icy fragments stretched across 1. 1 million km (710 thousand miles) of space, or 3 times the distance between Earth and the Moon. This required 6 WFPC exposures spaced along the comet train to include all the nuclei. The image was taken in red light. The comet was approximately 660 million km (410 million miles) from Earth when the picture was taken, on a mid-July collision course with the gas giant planet Jupiter.

Public Domain | Image courtesy of NASA.

After determining the comet's orbit, it was found that Shoemaker–Levy 9 fragments would impact Jupiter over several days, 22-26 July 1994. The question became if the impacts would be visible to astronomers and orbiting satellites, like the Hubble Space Telescope.

Over the five-day period, twenty-one (21) distinct impacts were observed by satellites like the Hubble Space Telescope, as well as Earth-based telescopes. Fireballs resembling an atomic bomb mushroom cloud were seen at Jupiter's horizon. Dark spots were seen in the upper atmosphere of Jupiter after impact; these were visible even in small amateur telescopes. Some of these spots were as big as Earth. The black impact features were visible on Jupiter for months; some likened them to a black eye. Due to the S-L 9 series of impacts, astronomers were better able to explain rows or chains of craters found on the Moon and other objects, like Jupiter's moon Ganymede.



NASA Hubble Space TelescopePublic Domain | Image courtesy of NASA.



NASA Hubble Space TelescopePublic Domain | Image courtesy of NASA.



NASA Galileo SpacecraftPublic Domain | Image courtesy of NASA.

This page titled 2.5.4.17: A Comet Impacts a Planet is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

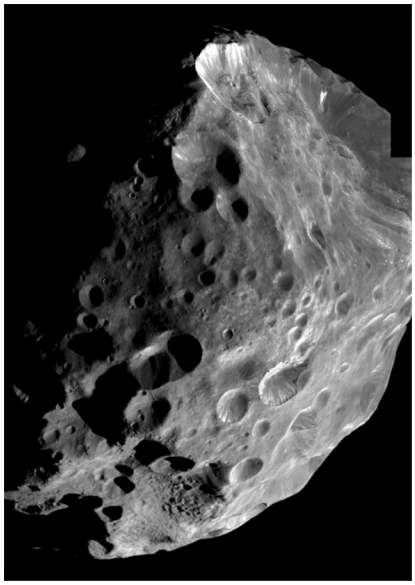
- 9.17: A Comet Impacts a Planet by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.18: The Centaurs- Are they Asteroids or Comets?

In the past few years, astronomers were closely observing asteroids that had characteristics of comets, and comets, which seemed more like asteroids. Astronomers determined they had found a new group of objects: the Centaurs. These objects are named after the mythological beings that were a mixture of human and horse. **Centaurs** behave with characteristics of both asteroids and comets; the formerly clean line between comets and asteroids was gone.

It is thought that perhaps 44,000 centaurs exist. Most centaurs have been found between Jupiter and Neptune, crossing the orbit or orbits of one of the Gas Giants. Several varying characteristics have been observed with identified centaurs being studied. What has been found so far is water ice on their surfaces, exhibiting comet-like tails, and the fundamental make-up of asteroids and not comets. There may be some relation between the centaurs and the Kuiper Belt Objects.

Probable the most notable centaur is **Chiron** ; an approximately 110-kilometer-wide object that displays characteristics of both an asteroid and the nucleus of Halley's Comet. Saturn's moon Phoebe might be a captured centaur. No confirmed centaur has been photographed up close.



Public Domain | Image courtesy of NASA.

This page titled 2.5.4.18: The Centaurs- Are they Asteroids or Comets? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

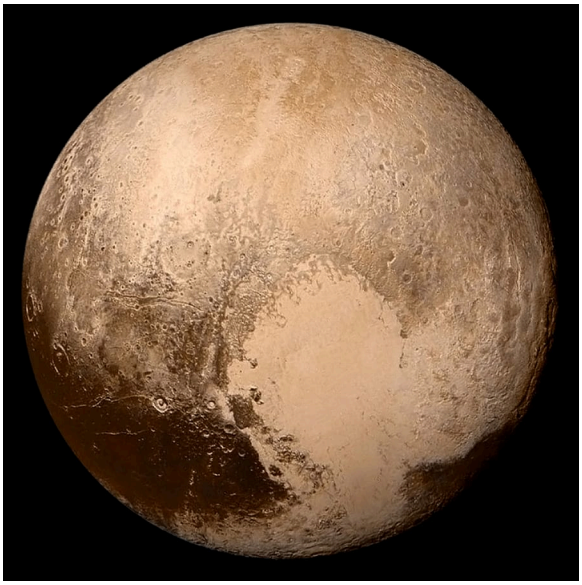
- 9.18: The Centaurs- Are they Asteroids or Comets? by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.19: Pluto

Pluto enters the solar system's stage. The then 9th planet was discovered in 1930 by Clyde Tombaugh at Lowell Observatory. The first two letters of **PL** uto are also for **P** ercival **L** owell, who mathematically predicted the location. In Roman mythology, Pluto (Greek: Hades) is the god of the underworld. This seemed like a fitting name for a planet so distant from the sun. Pluto has been controversial in recent years. Is Pluto a planet? Is it a comet or a centaur? Pluto appears to outgas like a comet and it appears to have more of a carbon monoxide, CO, atmosphere. Why does it have a moon, **Charon**, which is nearly the same size as Pluto, and also has four smaller moons, making five moons total? Pluto and Charon orbit around each other; a true double-dwarf planet system.

Pluto has a highly-elliptical orbit, more like a comet or asteroid than a planet. It is so elliptical that Pluto's orbit crosses inside the orbit of Neptune for a short period. And the orbit of Pluto is highly tilted against the plane of the solar system's planets.

Many of these questions will be answered in 2015 as the NASA spacecraft **New Horizons** arrives at Pluto after a nearly-decade long trip from Earth.



Public Domain | Image courtesy of NASA.

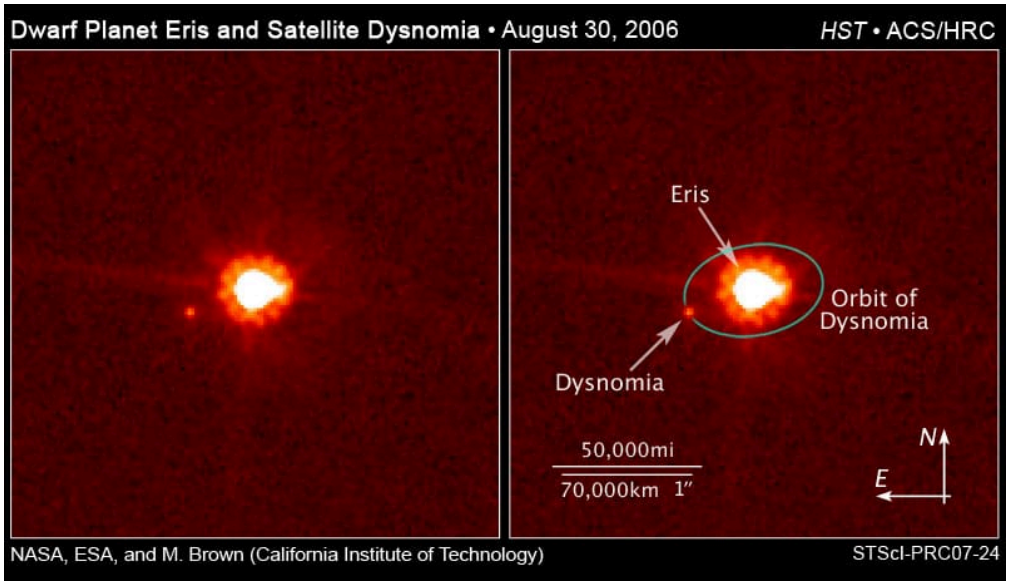
This page titled 2.5.4.19: Pluto is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.19: Pluto by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.4.20: Eris

Planet X: 2003 UB313, Eris was discovered by Dr. Mike Brown on October 21, 2003. After multiple observations in July 2005, the discovery of Eris was confirmed and formally announced. Eris is about 14. 5 billion kilometers from the sun and is in a very highly tilted orbit compared to the other planets. Eris is about 2,326 kilometers in diameter, whereas Pluto is 2,338 kilometers. Like Pluto, Eris is probably a very large comet or ice ball with a moon. Eris is named after the Greek Goddess of chaos and strife. A moon has been discovered orbiting Eris, named Dysnomia (Eris' daughter who was the spirit of lawlessness). Dysnomia is about 250 kilometers in diameter.

When first discovered, Brown gave Eris the name Xena, and its moon Dysnomia the name Gabrielle. Later the International Astronomical Union adopted the names Eris and Dysnomia.



Public Domain | Image courtesy of NASA.

This page titled 2.5.4.20: Eris is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 9.20: Eris by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5: THE SUN

“Here Comes the Sun
Here comes the sun
Here comes the sun,
and I say, It’s all right
Little darling
It’s been a long, cold lonely winter
Little darling
It feels like years since it’s been here”
...
Sun, sun, sun, here it comes

George Harrison
The Beatles
Abbey Road, 1969

Learning Objectives

Upon completion of this module, the student will be able to:Upon completion of this module, the student will be able to:

- Describe the production of energy and light through the proton-proton cycle
- Be able to identify the Sun’s regions
- Be able to identify the Sun’s solar features
- Overview how we study the Sun
- Identify the potential dangers it can pose to Earth

This module overviews our star, the Sun, how light and energy are produced, the various solar regions and features, and how we observe the Sun.

2.5.5.1: What do you think?
2.5.5.2: Fission and Fusion
2.5.5.3: Our Star, the Sun
2.5.5.4: How the Sun Works
2.5.5.5: A “Quick Guide” to Solar Fusion—The Proton-Proton Cycle
2.5.5.6: Solar Regions
2.5.5.7: Major Solar Features
2.5.5.8: Studying the Sun
2.5.5.9: The Dark Side of the Sun
2.5.5.10: The National Solar Observatory
2.5.5.11: NASA’s Heliophysics Science Division

This page titled 2.5.5: The Sun is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 10: The Sun by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.1: What do you think?

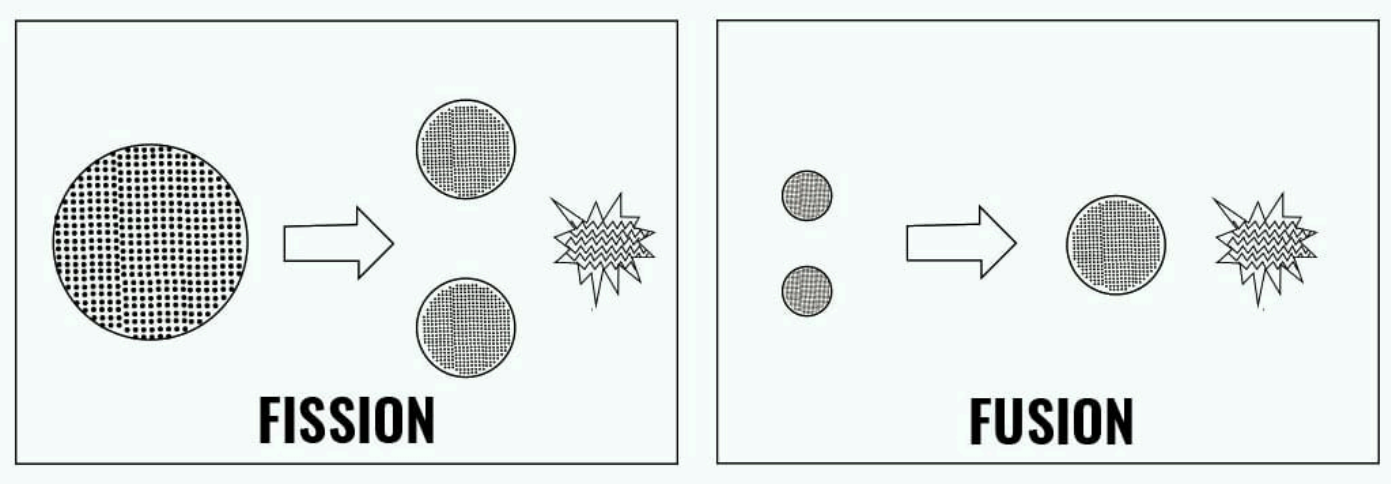
What would life be like if Earth was in the orbit of Venus or Mars? Would life be possible? Would life as we know it be possible? How much do we depend on the sun?

This page titled 2.5.5.1: What do you think? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 10.1: What do you think? by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.2: Fission and Fusion

Fission is the splitting of large nuclei of certain isotopes into smaller nuclei along with the release of energy with several free neutrons. Almost all of the fission fragments or products are radioactive. Neutrons released during the fission process can strike other nuclei and cause them to split, called a **Chain Reaction**. **Fusion** is the combining of particles to produce a new isotope and energy; **think of nuclear fusion as being the opposite of nuclear fission**. Fusion is the combining of particles into larger particles, whereas fission is the splitting of particles into smaller particles. The fusion process involves smaller nuclei than the fission process. When fusion occurs at a very high temperature — 50 million degrees Celsius, this is referred to as **thermonuclear fusion**.



The Fission versus Fusion processCC BY-SA 3. 0 | Image courtesy of Wikimedia Author: Kelvinsong.

This page titled 2.5.5.2: Fission and Fusion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.5.5.3: Our Star, the Sun

The sun, our central star in the solar system, plays a major role in life on Earth. The sun is about 300,000 times closer than the next closest star, Proxima Centauri. It takes light about 8. 3 minutes to travel from the sun to Earth; whereas light from Proxima Centauri takes over four years. Size and temperature-wise, the sun is an “average” star, at best. The sun goes through constant changes: energy output, brightness, size; all of which affect Earth.

Studying the sun and how it works allows astronomers to understand other stars. This study of how stars work is called **Astrophysics**. We specifically study the sun’s **fusion** process.

The sun’s **photosphere** vibrates like ripples on a pond. This is due to the sun producing sound waves through the gases. Sound waves and ripples emanate from within the Sun at very low frequencies deep within the Sun and higher frequencies close to the Sun’s surface. Studying these ripples and sound waves allows astronomers to examine the sun’s interior.



The sun in white light is visible through a small telescope. Note the sunspots; the large group is near the center. CC BY-SA 3. 0 | Image courtesy of Wikimedia Author: SiriusB.

This page titled 2.5.5.3: Our Star, the Sun is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 10.3: Our Star, the Sun by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.4: How the Sun Works

The sun does not burn. The belief that the sun does burn is a misconception. In stars the size of the sun or smaller, the fusion process is referred to as the **proton-proton cycle** or **proton-proton chain reaction**. The general steps to the Proton-proton cycle include:

1. It takes about 100,000 years for that energy to get from the sun's core to the photosphere.

This page titled 2.5.5.4: How the Sun Works is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 10.4: How the Sun Works by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.5: A “Quick Guide” to Solar Fusion—The Proton-Proton Cycle

Two protons fuse to yield deuterium, which fuses with a proton to form He-3. This fuses with another He-3 to form He-4, two protons, and a lot of energy is produced and released during this process. The Proton-Proton Cycle is for *less-massive* stars, like the sun.

Eighty-five percent of the sun’s energy comes from this type of reaction; other reactions include:

1. A He-3 atom and a He-4 atom combine to form a beryllium-7 ($4p^{+} + 3n^{0}$) and a γ
2. A Be-7 atom captures an e- to become lithium-7 atom ($3p^{+} + 4n^{0}$) and a neutrino
3. A Li-7 combines with a p+ to form two He-4 atoms

It appears that the sun has used about 50% of its Deuterium. So, what happens when the sun uses 100% of its Hydrogen? *The Proton-Proton Cycle will continue until the sun runs out of Hydrogen*. This is much like a car running out of gas – unless you can syphon fuel from another car (or another star).

How do we know that Solar Fusion happens? We look for the neutrinos from the fusion process by counting the actual number of neutrinos received versus the calculated number of neutrinos expected.

More-Massive Stars

For stars more massive than the sun, the Proton-Proton cycle can still occur, but another reaction sequence becomes more favorable for converting hydrogen to helium. It is called the **CNO cycle** ;**C** arbon- **N** itrogen- **O** xygen. These stars are ~1. 33 solar masses or larger with the final fusion end product being the element iron, Fe. The process goes no further because iron cannot be fused due to its binding energy.

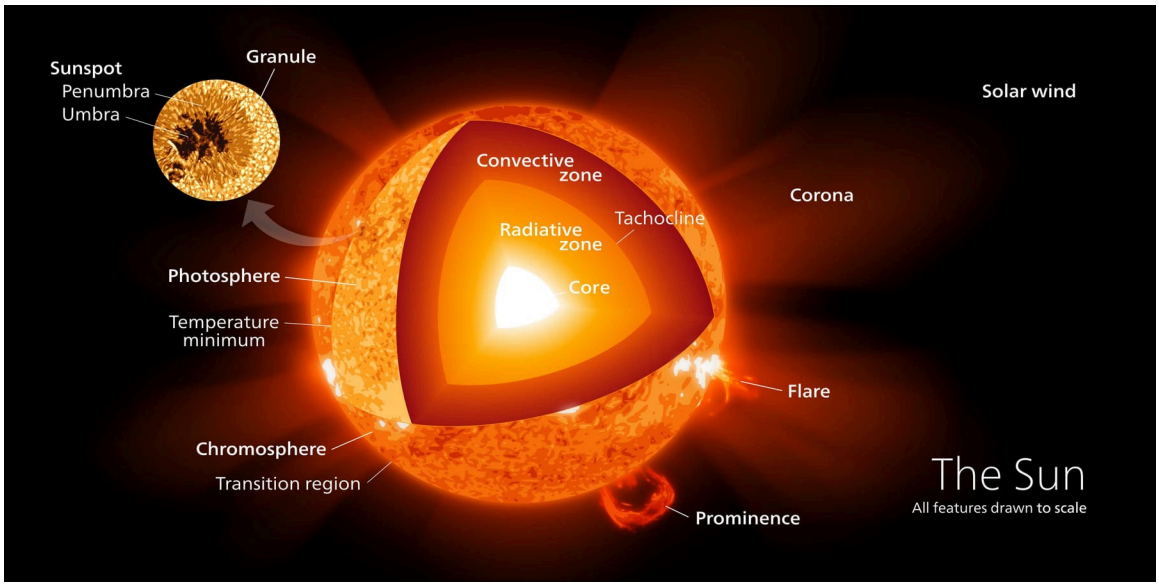
This page titled 2.5.5.5: A “Quick Guide” to Solar Fusion—The Proton-Proton Cycle is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 10.5: A “Quick Guide” to Solar Fusion—The Proton-Proton Cycle by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.6: Solar Regions

Astronomers identify four major regions within the sun.

1. First there is the sun's **Core**, the central region where nuclear fusion occurs. This is a very high temperature gaseous soup of charged particles called plasma at 15 million degrees.
2. Next is the **Radiative Zone**, where energy generated in the core is carried by light that bounces from particle to particle through the Radiative Zone.
3. Above the radiative zone is the **Interface Layer**, where it is thought the sun's magnetic field is generated. This layer or zone lies between the Radiative and Convective Zones.
4. The **Convective Zone** is the region in which energy is transported outward by convection; 70% of the sun's radius. Convection occurs because heated fluids rise and cooler fluids fall. Convection is the area of the sun which shows rising heat – recall that heat rises and cold falls.

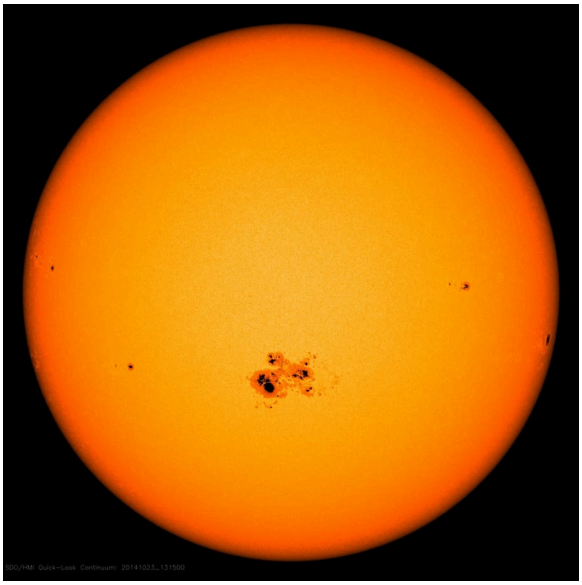


Regions of the SunPublic Domain | Image courtesy of NASA.

This page titled 2.5.5.6: Solar Regions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

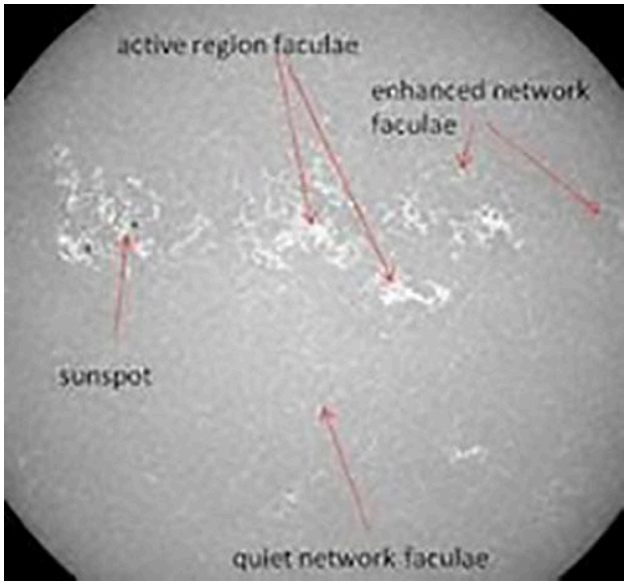
- 10.6: Solar Regions by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.7: Major Solar Features



Photosphere, with SunspotsPublic Domain | Image courtesy of NASA.

The **Photosphere** is the visible portion of the sun. It consists of the area in which gaseous layers change from completely opaque to being transparent. The photosphere is also the lowest observable layer of the solar atmosphere.



Solar *Faculae*Image courtesy of courtesy of P. Foukal, Heliophysics, Inc.

Faculae are bright areas visible on the Photosphere. The term faculae is Latin for “little torch. ”

Image courtesy of Hinode JAXA/NASA/PPARC.

Granulation is vast gas bubbles, with rising centers and sinking edges. This bubbling pattern is due to solar convection, as noted in the sun’s regions. Each granulation cell is about 1,000 miles across; compare that to the diameter of Earth at a little under 8,000 miles. Granulation is also a photosphere feature.

Sunspot’s Umbra (dark, center portions) and Penumbra (lighter, outer portions).

Image courtesy of Nasa/SDO.

We observe **limb darkening** because the sun is a sphere, near what is seen as the edge of the solar disk, the light must travel farther through the solar atmosphere. This causes the limb to be dimmer than the rest of the disk.

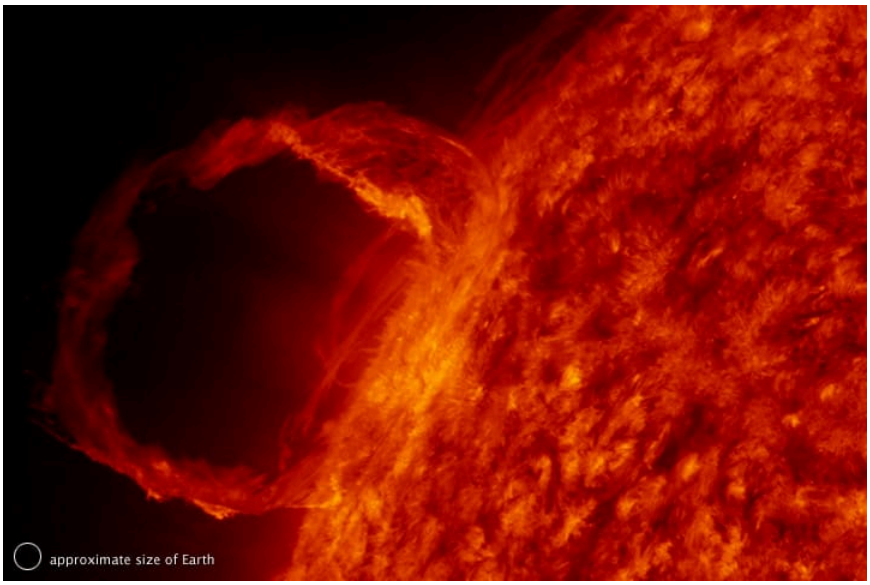
Sunspots are temporary disturbed magnetic areas in the Photosphere. Sunspots come in many shapes and sizes, according to the status of the sun's magnetic field. This magnetic field traps gas, slowing its motion, and making it cooler than the surrounding area. Sunspots look dark because it is cooler than the surrounding area. Usually, sunspots have two parts: a darker central region called the **umbra** surrounded by a lighter region called the **penumbra**. Astronomers have noted a sunspot cycle. This cycle averages in duration of slightly more than 11 years, in which **sunspot**frequencies varies from a maximum to a minimum and back to a maximum again.



Image courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

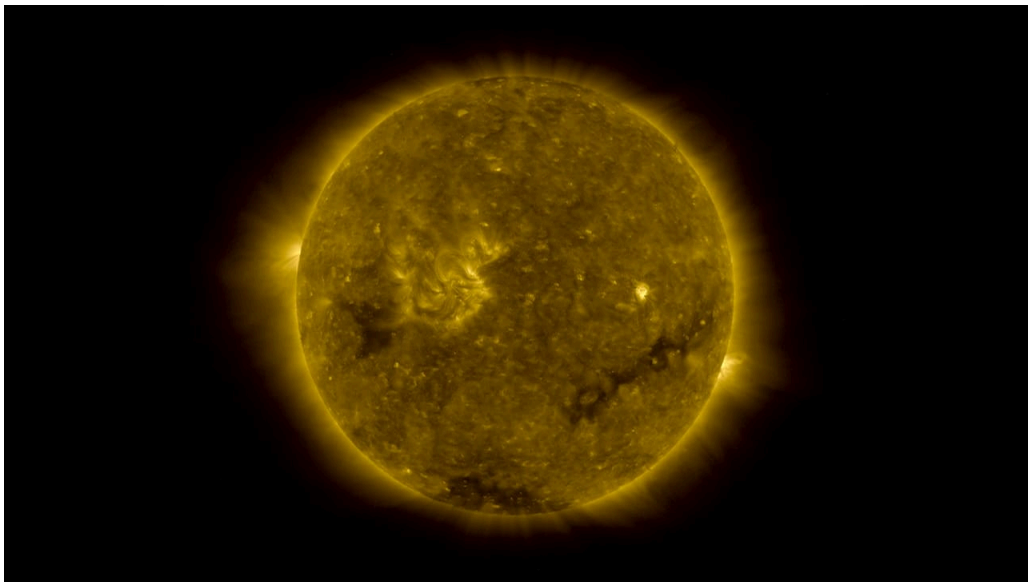
The **Chromosphere** layer of the sun's atmosphere located above the Photosphere / beneath the Corona. The term 'chromosphere' means the sphere of color. This is due to the fact that hydrogen atoms emit energy called hydrogen-alpha (H-α) radiation. The chromosphere is reddish in appearance, therefore it is hard to see.

Prominences are bright gas clouds ejected from the sun and shaped by the sun's magnetic field. They appear as spikes, loops, "trees," detached regions, and more when they occur at the sun's edge as seen from Earth. When seen straight on, however, prominences look like dark lines, known as **filaments**, silhouetted against the solar disk because they are slightly cooler than the surface beneath.



Solar Dynamics Observatory *Spacecraft*Image courtesy of NASA, image and animation from the Goddard Space Flight Center Scientific Visualization Studio and the Solar Dynamics Observatory.

Flares occur when the sun's atmosphere suddenly releases built-up magnetic energy. These emit radiation storms and are by far the Solar System's largest explosions.



NASA SDO, Solar Dynamics Observatory SpacecraftPublic Domain | Image courtesy of NASA.

The **Corona** is the sun's outer atmosphere; charged particles and dust at low density and very high temperature (10 million K).

Black and white image showing the sun's bright outer atmosphere.

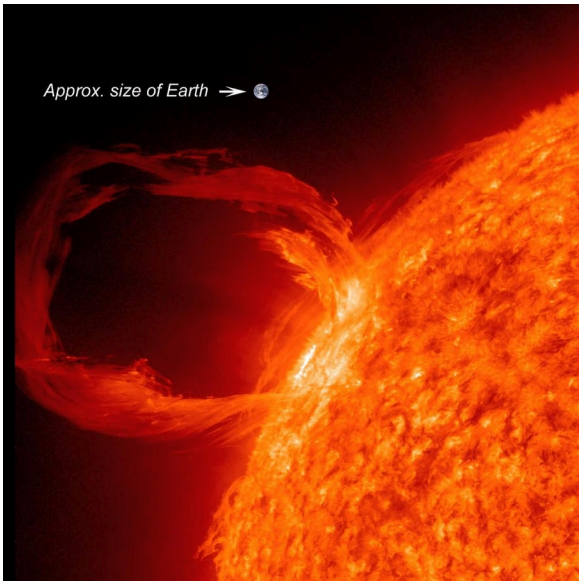
Solar *Corona*Image courtesy of Mike Reynolds, Ph. D. of FLorida State College at Jacksonville.

Solar Wind streams off of the sun in all directions at speeds of about 1 million miles/hour); the source of the solar wind is the sun's Corona.

Solar Wind – Computer Generated Image with a SOHO Solar Photograph.

Public Domain | Image courtesy of NASA / SOHO.

A **Coronal Mass Ejection** or **CME** occurs when huge gas bubbles threaded with magnetic field lines are ejected from the sun over several hours. If a CME directly hits Earth, the results can be catastrophic; from disruption of orbiting satellites to destroying electrical power stations. It takes 3 to 72 hours for a CME to strike Earth.



Public Domain | Image courtesy of NASA.

This page titled 2.5.5.7: Major Solar Features is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 10.7: Major Solar Features by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.8: Studying the Sun

We can clearly see the sun's surface features, examine its characteristics in a variety of wavelengths, and glean insight into the nuclear mechanisms of not only the sun, but also of other stars. When observing the sun in **Visible light**, you MUST use a proper solar filter — in front of your eyes, the telescope or binoculars to view the sun safely. A proper solar filter is safe because it does not transmit ultraviolet or infrared radiation, both of which are much more harmful to your eyes than light. It also drops the sun's brightness to a comfortable level. And a pinhole viewer can also be constructed to safely – and indirectly – view the sun and sunspots.

You might have heard that you should never look at an eclipse. First of all, all lunar eclipse phases are safe to view. The partial phase of a solar eclipse, or an annular solar eclipse requires a proper solar filer. It is NOT the eclipse that causes eye damage, it is the sun. During solar eclipse totality, no filter whatsoever is required.

In 1814, optician and physicist Joseph Frauenhofer (1787-1826) invented the first spectroscope. Frauenhofer was comparing the spectra of fire to that of the Sun when he noted dark lines in the sun's continuous spectrum. Named the **Frauenhofer lines**, these absorption lines are a function of the sun's photosphere.

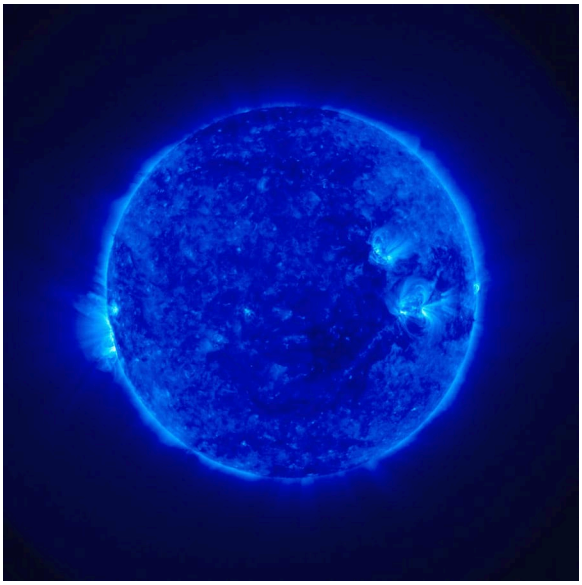
Such a spectrum is created when light passes through a gas or a liquid, or strikes a solid surface. You see the spectrum of the absorption; the wavelengths of light absorbed by the material are absent in the spectrum, leaving blank spaces behind

Perhaps the most-impressive is **Hydrogen-Alpha, H-a**. H-a filters center on a wavelength of 6562. 8 Å and allow through only a tiny part of the red light the sun produces, blocking all other colors. Solar flares and prominences are best seen through an H-a filter.

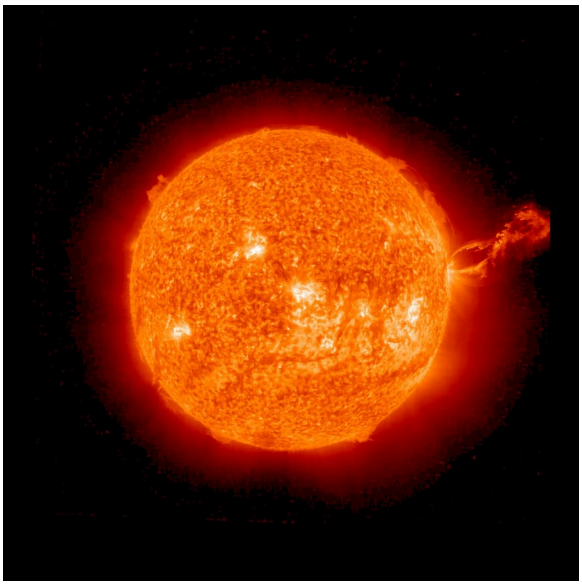
Studies of the element calcium, Ca, on the sun are done at **Calcium H** (3969Å) and **Calcium K-lines** (3933. 7Å). This is instrumental in determining the solar atmosphere's depth.

Other specific features are visible at narrow bandwidths. High granulation and supergranulation detail, and impulsive-phase flare eruption kernels are visible at the **Sodium Na-D line** (5895. 9Å). Flare and magnetic lines research is conducted at the **Helium D3 line** (5875. 61Å); this is one of the most impressive visually.

Light emitted by highly charged iron (Fe) ions at high temperatures is best studied in the **Ultraviolet, UV. And X-Rays** are released from the sun's upper atmosphere and corona.



Public Domain | Image courtesy of NASA / GSFC.



Public Domain | Image courtesy of NASA / GSFC.

Our Sun is important – *critical* – to life on Earth, yet is it not the calm star, which we all might presume...

Watch on YouTube these NASA solar videos:

"Fiery Looping Rain on the Sun," Video Link *Note the Earth to scale from 1:06 to 1:22.*

"Eruptive events on the sun can be wildly different. Some come just with a solar flare, some with an additional ejection of solar material called a Coronal Mass Ejection (CME), and some with complex moving structures in association with changes in magnetic field lines that loop up into the sun's atmosphere, the corona" (NASA, retrieved, March 29, 2015).

On July 19, 2012, an eruption occurred on the sun that produced all three events:

- The Solar Dynamics Observatory's AIA Instrument collected the footage in this video. SDO collected one frame every 12 seconds, and the movie plays at 30 frames per second, so each second in this video corresponds to 6 minutes of real time. The video covers 12:30 a. m. EDT to 10:00 p. m. EDT on July 19, 2012.

  "Magnificent Eruption," Video Link

  Consider the following questions based on the YouTube videos:

  1. What was your first reaction to seeing/hearing these videos?
  2. How would you describe the Earth when compared to the Sun in size in the first video, *"Fiery Looping Rain on the Sun?"*
  3. As we 'look' at our Sun in the sky, it seems fairly quiet. Yet what do these videos show us?
  4. Why would we want to monitor the Sun's activity? What can this teach us about other stars?
  5. Your speculation: if one of these Coronal Mass Ejection outbursts were to strike Earth directly, could it cause any problems?

This page titled 2.5.5.8: Studying the Sun is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 10.8: Studying the Sun by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.9: The Dark Side of the Sun

The sun – and every star – is not a consistent, even-running object. The sun is a variable star; not as much variation or changes as most stars classified as variable. But the sun does go through sunspot cycles and sunspot activity extremes.

Both long- and short-term differences in solar activity are believed to affect the global climate, yet it has shown to be challenging to link the variation of the sun to Earth's climate. One school of thought is that these solar variations are due to sunspot cycle extreme minimums, which can make for a colder climate period. One hypothesis is that the Little Ice Age, which occurred between 1550 and 1850, was due to low solar activity. During the 1607-1608 winter, extreme frost was reported in Jamestown, Virginia. New York Harbor froze in the winter of 1780; very cold periods were well-documented throughout Europe and worldwide.

The release and outbursts of energy through solar flares and Coronal Mass Ejections can cause devastating effects on Earth. Solar flare X-Rays travel at the speed of light, 186,000 miles per second or 3 x 10⁸ meters/second. Electrically charged particles, released from a CME, travel at around 1 to 2 million miles an hour; think of this as a cosmic tsunami. These are capable of taking down Earth's electric power grid and knocking out orbiting satellites, with catastrophic results. One such event ooccurred in Quebec, Canada due to a CME, March 1989. In 1859, enormous solar storms struck Earth. The Aurora were so bright at night people thought it was daylight. And these CMEs literally burned out that day's only electronic system: the telegraph.

This page titled 2.5.5.9: The Dark Side of the Sun is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 10.9: The Dark Side of the Sun by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.10: The National Solar Observatory



The National Solar Observatory Telescope *Tower*Image courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

The main telescope at NSO Sacramento Peak is the Richard B. Dunn Solar Telescope. This imposing solar tower, 136 feet high, produces 0. 2 arc-second resolution and over many years has produced some of the finest solar images and data. Another 220 feet of the Dunn Solar Telescope lies below ground, making the total length of this imposing instrument 356 feet. The instrument was upgraded to include adaptive optics; another advantage to reduce seeing effects while observing the sun.

This page titled 2.5.5.10: The National Solar Observatory is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 10.10: The National Solar Observatory by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.5.5.11: NASA’s Heliophysics Science Division

Nasa’s Heliophysics Science Division conducts research on the sun, its extended solar-system environment (the heliosphere), and interactions of Earth, other planets, small bodies, and interstellar gas with the heliosphere. Division research also encompasses geospace — Earth’s uppermost atmosphere, the ionosphere, and the magnetosphere — and the changing environmental conditions throughout the coupled heliosphere (Solar System weather). National Aeronautics and Space Administration, Goddard Space Flight Center

This page titled 2.5.5.11: NASA’s Heliophysics Science Division is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 2.6: STELLAR PROPERTIES

Learning Objectives

Upon completion of this module, the student will be able to:Upon completion of this module, the student will be able to:

- Identify the basic stellar characteristics
- Describe how astronomers use parallax
- Identify terms, such as light-year, parsec, stellar luminosity, and magnitude system
- Detail Stellar Types and those responsible for the development of the stellar types
- Describe how the H-R Diagram works and what it shows
- Identify the different types of binary stars and variable stars
- Differentiate between globular and open clusters

2.6.1: Stellar Objects

2.6.1.1: Module Introduction

2.6.1.2: Variable Stars

2.6.1.3: Cepheids

2.6.1.4: Cataclysmic Variables

2.6.1.5: Star Clusters

2.6.1.6: Consider this...

2.6.1.7: The Unknown in Astronomy

2.6.1.8: Star and Celestial Object Characteristics

2.6.2: Stellar Distances

2.6.2.1: Distances to the Stars and other Celestial Objects

2.6.2.2: Stellar Distances

2.6.2.3: Everything is Moving

2.6.2.4: How Close are the Close-by Stars?

2.6.2.5: Light-Years

2.6.2.6: Parsec Vs. Light-Year Measurement

2.6.3: Stellar Cartography

2.6.3.1: Stellar and Celestial Object Brightness

2.6.3.2: Magnitude System

2.6.3.3: Celestial cartography

2.6.3.4: Planispheres

2.6.3.5: Star Colors

2.6.3.6: Spectral Type

2.6.3.7: Sample Spectral Types

2.6.4: The Hertzsprung-Russell Diagram

2.6.4.1: The Hertzsprung-Russell Diagram

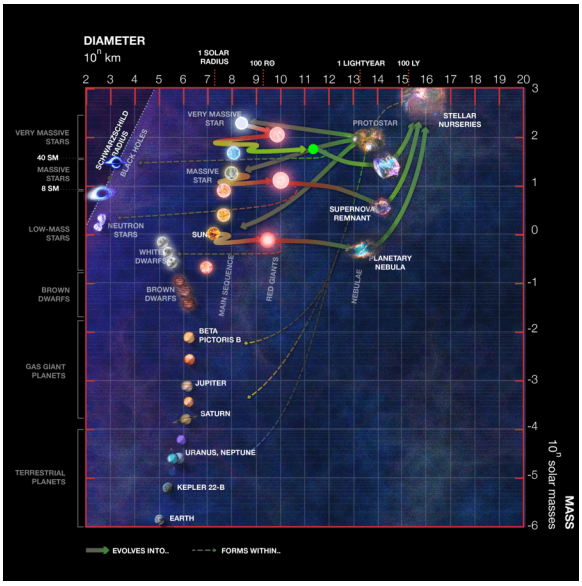
2.6.4.2: The Four Hertzsprung-Russell Diagram Stellar Groups

This page titled Chapter 2.6: Stellar Properties is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.1: Module Introduction by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.
- 11: Stellar Properties by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.1: STELLAR OBJECTS

"Stellar Objects" covers a wide variety of celestial bodies.



A logarithm plot of stars, planets and other celestial objects by diameter and mass.

Image courtesty Wikipedia. Author: Avsa. Liscense.

2.6.1.1: Module Introduction

2.6.1.2: Variable Stars

2.6.1.3: Cepheids

2.6.1.4: Cataclysmic Variables

2.6.1.5: Star Clusters

2.6.1.6: Consider this...

2.6.1.7: The Unknown in Astronomy

2.6.1.8: Star and Celestial Object Characteristics

This page titled 2.6.1: Stellar Objects is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.6.1.1: Module Introduction

Children's lullabies were quite popular in decades past. Many of us learned a number of these lullabies as children, though they do not seem to be as popular today. In Rhymes for the Nursery, published by sisters Jane and Ann Taylor in 1806 in London was a simple rhyme about a star. Many of us have heard and even sung the first verse of what we call Twinkle, twinkle little star. And some of these questions in this children's lullaby are still the same astronomers and astrophysicists are asking today.

> The Star
>
> *Twinkle, twinkle little star,*
> *How I wonder what you are,*
> *Up above the world so high,*
> *Like a diamond in the sky.*
> *When the blazing sun is gone,*
> *When he nothing shines upon,*
> *Then you show your little light,*
> *Twinkle, twinkle, all the night.*
> *Then the traveller in the dark,*
> *Thanks you for your tiny spark,*
> *He could not see which way to go,*
> *If you did not twinkle so.*
> *In the dark blue sky you keep,*
> *And often through my curtains peep,*
> *For you never shut your eye,*
> *'Till the sun is in the sky.*
> *As your bright and tiny spark,*
> *Lights the traveller in the dark.*
> *Though I know not what you are,*
> *Twinkle, twinkle, little star.*
> *Twinkle, twinkle, little star.*
> *How I wonder what you are.*
> *Up above the world so high,*
> *Like a diamond in the sky.*
> *Twinkle, twinkle, little star.*
> *How I wonder what you are.*
> *How I wonder what you are.*
>
> Jane Taylor
> The Star, *Rhymes for the Nursery*, 1806

This module looks at the characteristics of stars, the terminology astronomers use to describe the stars, and the types and classifications of stars.

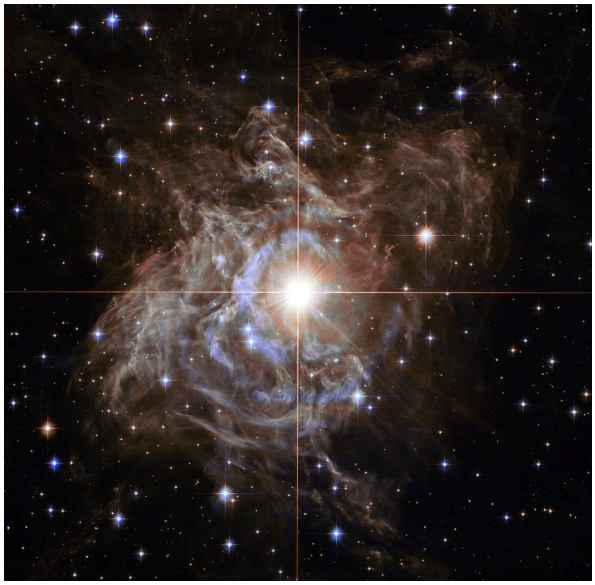
This page titled 2.6.1.1: Module Introduction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.1: Module Introduction by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.1.2: Variable Stars

Sometimes a star will change luminosity not due to an eclipse, but due to some sort of physical characteristic. These are called **variable stars** or **variables**. These changes in brightness can range from 1/1000 of a magnitude to 20 magnitudes over a period of a fraction of a second to many years. Over 150,000 variables are known, and many others are suspected to be variables.

There are two major classes of variable stars: **Pulsating Variables** and **Cataclysmic Variables**. **Pulsating variable stars** swell and shrink, which affects the star's brightness. One important class of pulsating variable stars is the **Cepheid Variables**.



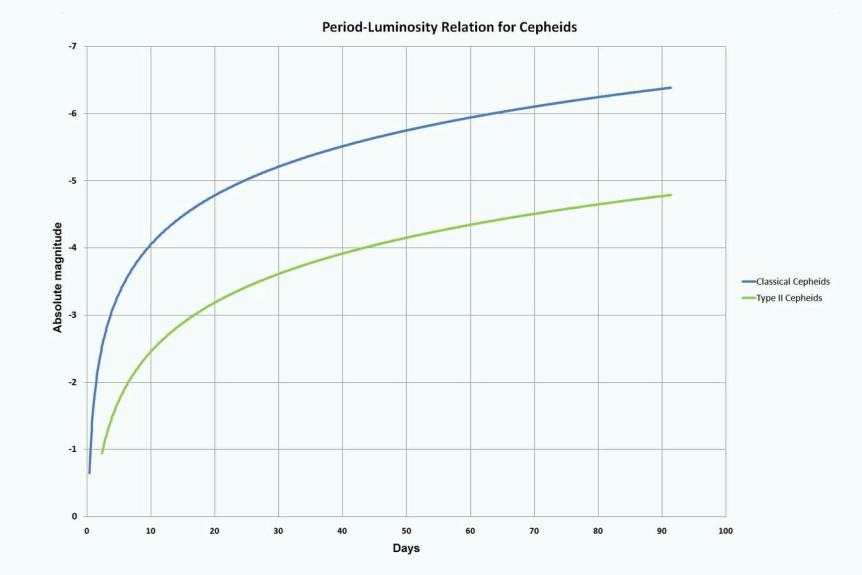
Public Domain | Image courtesy of NASA.

This page titled 2.6.1.2: Variable Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.2: Variable Stars by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.1.3: Cepheids

Often simply called **Cepheids**, these variables are massive, high-luminosity stars with short periods of one to 70 days and light variations of 0. 1 to 2 magnitudes. Astronomers use Cepheids as a “standard” for distances because of the direct relationship between a Cepheid’s luminosity and variable pulsation period. Once you determine a specific Cepheid’s luminosity and variable pulsation period, you can determine how distant it is because of the inverse square law. This makes Cepheids a major galactic distance indicator.



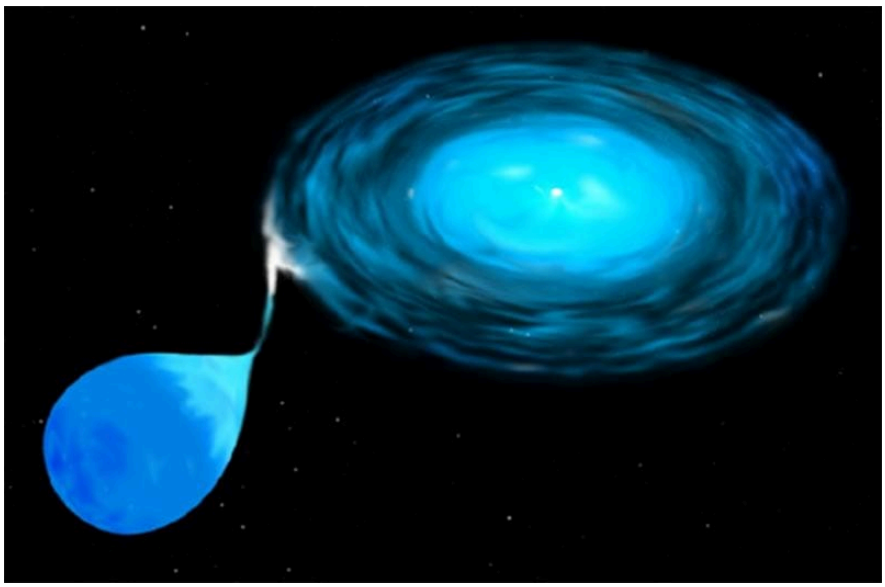
Public Domain

This page titled 2.6.1.3: Cepheids is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.6.1.4: Cataclysmic Variables

Cataclysmic variables are binary stars, which consist of a white dwarf primary and an orbiting secondary star. The secondary star is transferring matter to the primary star. This causes the primary star to irregularly outburst a significant increase in brightness. Cataclysmic variables were originally called **novae** (singular: nova) from the Latin term *new* , since the star seemingly appeared out of nowhere. After the outburst, the cataclysmic variable will eventually dim back down significantly until more stellar fuel is transferred from the secondary star to the primary star. This process will go on until there is no more stellar fuel to transfer.

The primary star (upper right) is pulling stellar material off the secondary star (lower left). Once a significant amount of mass builds up on the primary star, it will outburst and brighten significantly. Then the star will dim back down, repeating the process as long as fuel is available from its secondary star.



Public Domain | Image courtesy of NASA.

This page titled 2.6.1.4: Cataclysmic Variables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.4: Cataclysmic Variables by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.1.5: Star Clusters

Stars are often seen to group into what astronomers term as **star clusters**. Star clusters are gravitationally bound groups of stars. There are two major types of star clusters. **Open Clusters (OC)** are groups of young, hot stars that are physically related by being held together by gravity; less than 50 to 100 or more stars. **Globular Clusters (GC)**, or **globulars**, are tight groupings of very old stars; 10,000 to one million or more stars. Open clusters are irregular in shape or pattern, whereas globular clusters are spherical in shape. Open cluster stars can somewhat disassociate with other open cluster star members, since the gravitational attraction thus association with each other is loose and can be affected by other astronomical bodies passing near the OC.

Photographically, the two types of star clusters are easy to tell apart. Open clusters have fewer stars and appear to be not-well organized or symmetrical in shape. Globulars have many, many more stars than open clusters and are symmetrical and spherical in shape.

Open Clusters



Images courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

Globular Clusters



Images courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

This page titled 2.6.1.5: Star Clusters is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.5: Star Clusters by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.1.6: Consider this...



Tyson at the 40th Anniversary Apollo 11 celebration, July 20, 2009Public Domain | Image courtesy of NASA.

“So strong was that imprint [of the night sky] that I’m certain that I had no choice in the matter, that in fact, the universe called me. ”

—Dr. Neil deGrasse Tyson Astrophysicist and host of *Cosmos From an interview with Stephen Colbert, January 29, 2010*

This page titled 2.6.1.6: Consider this... is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.6: Consider this... by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.1.7: The Unknown in Astronomy

One of the best techniques astronomers employ is in understanding what is normal. For example, if we consider our Sun a normal star, how do other stars compare? When is a star atypical when compared to the Sun? And atypical in what way?

It is usually the abnormal or weird objects that are the most interesting. One example is stars that vary significantly in brightness and energy output on a regular basis. Why do these stars go through these variations? Or consider stars that also vary significantly in brightness and energy output, yet on an irregular basis. What makes them so much different than regular variable stars, or steady stars like our Sun?

This page titled 2.6.1.7: The Unknown in Astronomy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.7: The Unknown in Astronomy by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.1.8: Star and Celestial Object Characteristics

Star characteristics consider physical characteristics such as **stellar mass, size, surface temperature, and luminosity**. These basic characteristics assist in detailing specifics about a star; not all stars are the same and these variations in characteristics can be significant. The distance from the Sun and Earth to a specific star is also important as astronomers detail the star's characteristics. Of these characteristics, **stellar mass** is the most important characteristic. The mass of a star determines things, such as how quickly a star will consume its stellar fuel through the fusion process to the star's final ending when it has used all of its fuel. Yet how does one measure a star's mass? How do you weigh a star?

It turns out that Isaac Newton studied this question. He turned to Kepler's Third Law, where one measures the **period** and **average distance** of the object's orbit about a star. Yet, you need two objects – a star and an object orbiting a star – to use this solution. It turns out that over 50% of all stars have a companion star. So, astronomers use this adaptation on Kepler's Third Law developed by Newton to measure the mass of the two binary stars.

Kepler's 3rd Law — $a^3 = kP^2$

Where **a** is the orbiting object's semi major axis, **P** is the orbiting object's period to orbit, and **k** is a constant, referred to as Kepler's constant.

By examining the color of each of the stars in the binary system, you can compare two single stars with the same colors.

Stellar mass is usually related to the mass of the Sun, where the Sun equals 1 m_{sun}, 1 $m_{\odot}$, or 1 solar mass. The bright star Sirius, the Dog Star in the constellation Canis Major, is about 2. 02 m_{sun}. One of the most massive stars is Eta Carinae, with a mass somewhere between 100 to 150 times the mass of the Sun, 100-150 $m_{\odot}$.

Stellar mass units — m_{sun} :: $m_{\odot}$ :: solar mass

A star's mass will vary over its lifetime, depending if it adds, or accretes, mass from another star, loses mass to another star, or simply loses mass through the normal processes, such as through its stellar wind or pulsating outputs.

Stars are occasionally classed by their stellar masses based upon their evolutionary behavior as stars approach the end of their nuclear fusion. In the next module, we will introduce the classes of stars, based on their solar masses.

Stellar size refers to a star's diameter or radius. Stars range in diameter, from neutron stars with diameters of about 40 kilometers or 25 miles, to supergiants with diameters of approximately 900,000,000 kilometers or 540,000,000 miles — about 650 times the Sun's diameter.

A star's **surface temperature**, measured in Kelvin, K, is dependent on the star's diameter and the rate of energy production at the stellar core, and is measured at the star's photosphere. An estimate of the surface temperature is the star's **color**, often called the **color index**. Annie Jump Cannon was the first to sort spectral data and designed the stellar spectral classes.

The hotter the star, the whiter it will appear, whereas the cooler the star, the redder it will appear. Think of heating a piece of metal; the hotter it is, the whiter the metal will appear. As the metal cools, it will appear orange and then red in color. The reddish-colored metal is still hot, yet cooler than when the metal was white blue-white in color.

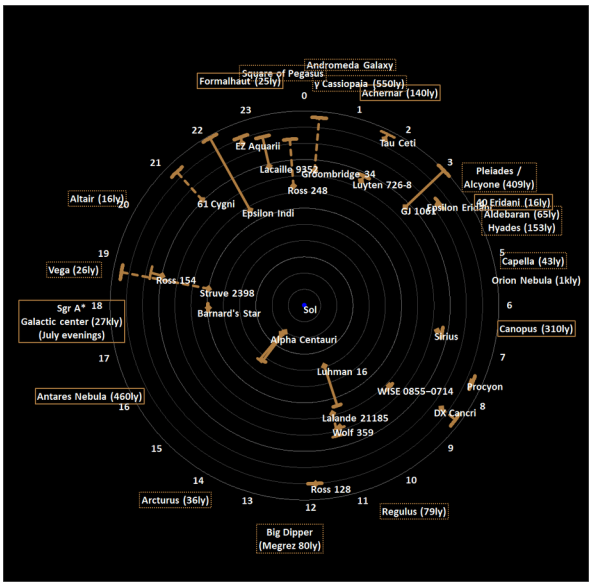
Stellar luminosity is the amount of light and other radiant energy released by a star. A star's luminosity is dependent on its diameter (sometimes noted as the star's radius, d = 2r) and its surface temperature.

This page titled 2.6.1.8: Star and Celestial Object Characteristics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.8: Star and Celestial Object Characteristics by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.2: STELLAR DISTANCES

Distances and time are linked in astronomy. Because of the way light travels, the further something is away, the further back in time we are looking.



A map of popular stellar objects within 12 light years of us.

Image courtesy Wikipedia. Author: Nsae Comp. Liscense.

2.6.2.1: Distances to the Stars and other Celestial Objects

2.6.2.2: Stellar Distances

2.6.2.3: Everything is Moving

2.6.2.4: How Close are the Close-by Stars?

2.6.2.5: Light-Years

2.6.2.6: Parsec Vs. Light-Year Measurement

This page titled 2.6.2: Stellar Distances is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.6.2.1: Distances to the Stars and other Celestial Objects

When discussing stars, astronomers consider other factors, such as the **star's distance** and **brightness**. These quantities might seem easy to determine at first, yet they are challenging. And to fully understand each star's physical characteristics, we need to determine these most basic of quantities.

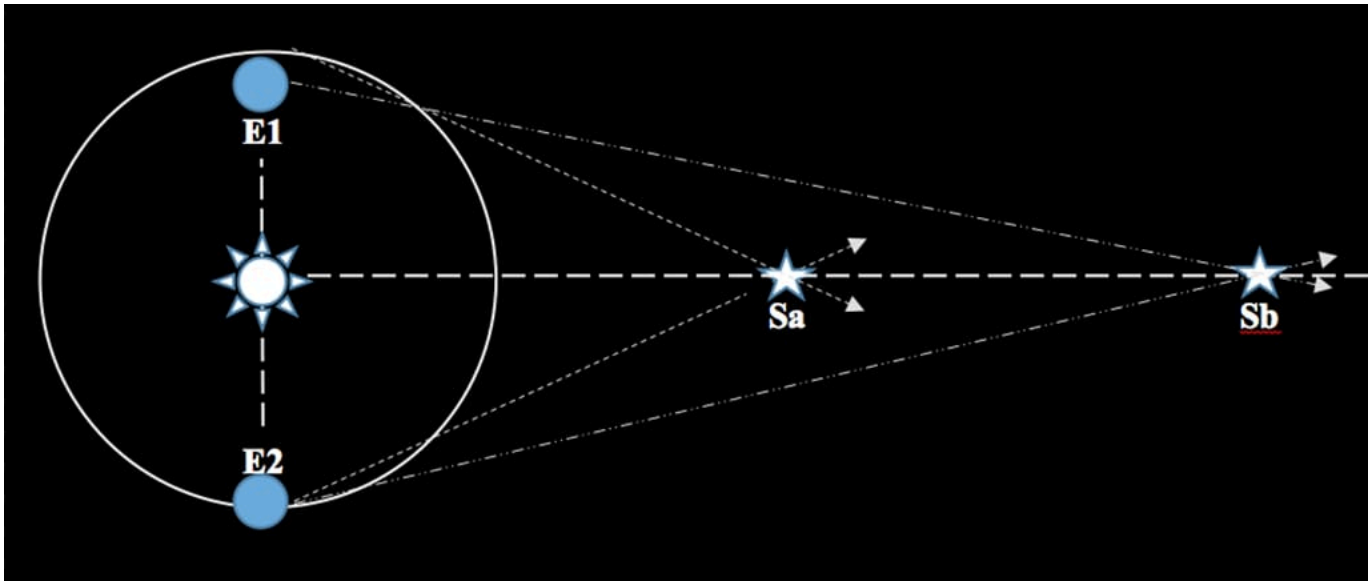
The distance to a specific star – beyond our Sun – is not an easy thing to determine. There is no way to use a measuring tape; the tool we are used to using in our own personal universes. As with measuring distances from object to object here on Earth, radar works well for many Solar System objects, but not beyond. The radar beam spreads out too much and loses energy. And it would take over eight years for that beam to travel to the closest star beyond the Sun – Proxima Centauri – and back. By the time the radar beam returned, there would be nothing to receive here on Earth because the beam would have lost all of its energy.

So to measure the distance to nearby stars, astronomers observe an object's **stellar parallax**. This is the apparent shift of an object relative to some distant background as the observer's point of view changes. As Earth and observer orbits the Sun, the specific star being observed will appear to move relative to the background stars. **Parallax is the only method for directly measuring stellar distances.** You may try this by holding a finger up at arm's length. Open one eye and close the other. Now blink; reverse the open-closed eyes. Notice the objects in the distance appear to shift back and forth compared to your closer finger.

This page titled 2.6.2.1: Distances to the Stars and other Celestial Objects is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.9: Distances to the Stars and other Celestial Objects by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

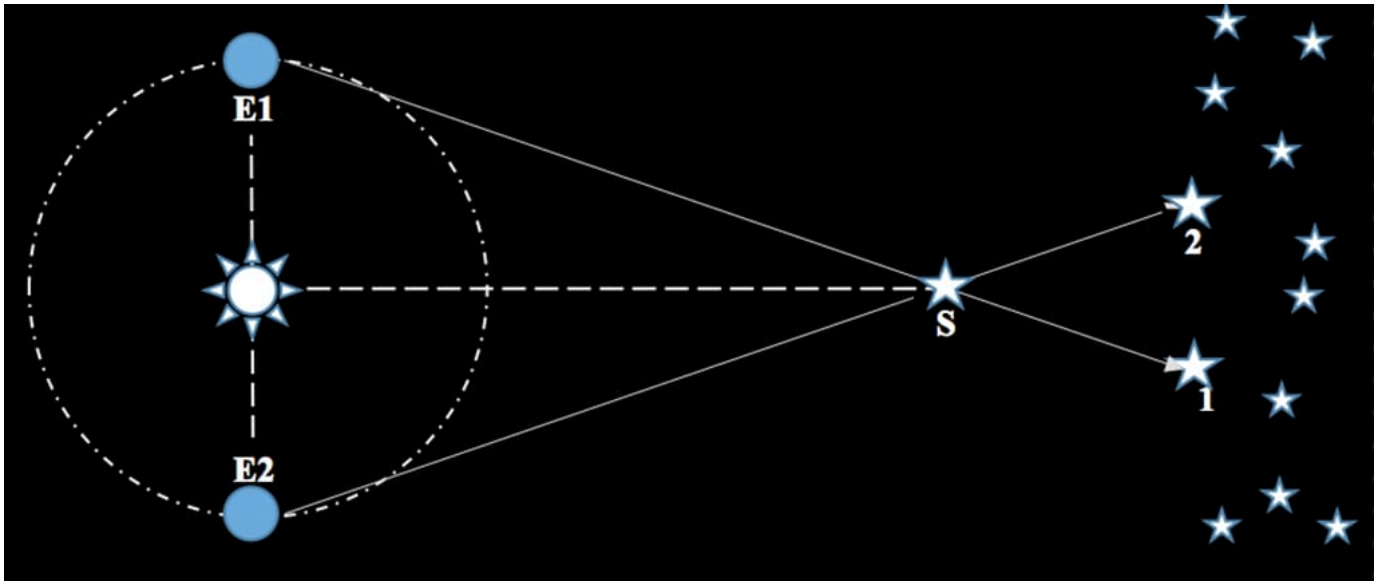
2.6.2.2: Stellar Distances



Images courtesy Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

This diagram *(not to scale)* shows Earth's orbit about the Sun. Recall that one orbit is 1 year, 12 months, or 365. 24 days. Earth is shown twice, at the 12 o'clock position, or **E1**, and at the 6 o'clock position, or **E2**. These two positions are six months apart on Earth's orbit about the Sun. The star being observed is the larger star, indicated by the **S**. The star **S** appears to shift against the background of stars from position **S1** to position **S2** over the six months. This is the parallax motion. By looking at this as a triangle and using trigonometric relationships and the angles, you can determine the distance to the star **S**.

There are issues with determining stellar parallax, and how useful it is in determining stellar distances. First, note that parallax only works with fairly close stars. Examine this diagram to see what happens when one star is at a greater distance.



Images courtesy Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

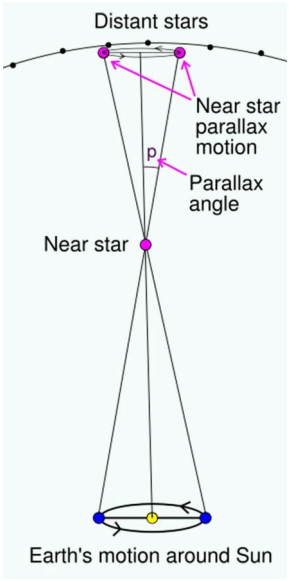
As with the first diagram, Earth is at two positions, **E1** and **E2**. In this diagram and case, there are two stars, **Sa** and **Sb**. The background stars in the first diagram have been removed to see the angles and two stars better. Note that star **Sb** is about two times as far from the Sun and Earth as star **S1**. Look at the angles; the **E1-Sa-E2** angle is much larger than the **E1-Sb-E2** angle. Think of the angles they create as pieces of pizza.

This page titled 2.6.2.2: Stellar Distances is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.10: Stellar Distances by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.2.3: Everything is Moving

Another challenge is the fact that the star's position you are measuring, the Sun and Solar System, and these background stars are also moving. The way to overcome this issue is to make more than two measurements and to take into account all of the motions – the Solar System and background reference stars.



Public Domain | Image courtesy of Wikimedia Author: Booyabazooka

This page titled 2.6.2.3: Everything is Moving is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.11: Everything is Moving by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.2.4: How Close are the Close-by Stars?

Line drawing illustrating distance from the Sun to Proxima Centaurai. Consider 4. 29 miles from the Sun if Pluto is 40 inches from the Sun.

Images courtesy Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

The closest star? **The Sun**. After the Sun is **Proxima Centauri**, part of the three-star Alpha Centauri (α Cen) star system. Proxima Centauri is about 25,219,301,250,000 miles or 40,586,533,727,000 kilometers away. So let's look at this in other ways. *A 50-mile high stack of toilet paper stretched out on 1 sheet = 1 million miles scale*. **Alpha Centauri** is often listed as the second closest star to Earth, *this is incorrect* . Proxima Centauri is the second-closest star and orbits Alpha Centauri.

This page titled 2.6.2.4: How Close are the Close-by Stars? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.12: How Close are the Close-by Stars? by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.2.5: Light-Years

Yet trillions of miles to the nearest star beyond the Sun is not a measurement or system that is user-friendly. So astronomers use light-years instead. A **light-year (ly)** is a unit of length or distance; the distance light travels in one year. It is also written as light year or lightyear. So how does light travel in one year, or what is the length of a light-year? This sounds like a time calculation, and is often misused as such in movies, television shows, and books. The light-year is a distance measurement, and is calculated by knowing the speed of light.

Using light-years instead of miles, meters, or kilometers provides one a friendly number. For example, instead of stating the Sun is 93,000,000 miles away, one can note it is 8. 33 light minutes distance. What of Proxima Centauri, 25,219,301,250,000 miles or 40,586,533,727,000 kilometers away? Using light-years, Proxima Centauri is 4. 29 ly in distance.

To give you a sense of the size of our galaxy and even the Universe, there are about 30 stars, which lie within 16. 5 ly of Earth. Only 30 stars out of the billions of stars in just our own Milky Way galaxy.

Light Years: Calculations

Light-Years calculated in miles — the English System: 186,000 miles/second x 365. 25 days/year x 24 hours/day x 60 minutes/hour x 60 seconds/minute = 5,869,713,600,000 miles or 5. 8697136 x 10^{12} miles

Light-Years calculated in meters and kilometers — the Metric System: 299,792,458 meters/second x 365. 25 days/year x 24 hours/day x 60 minutes/hour x 60 seconds/minute = 9,460,730,472,580,800 meters or 9. 4607304725808 x 10^{15} meters = 9,460,730,472,580. 8 kilometers 9. 4607304725808 x 10^{12} kilometers

Astronomical Units in 1 ly; distance from the Sun to Earth (93,000,000 miles = 1 AU): 1 ly = 63,241. 077 AU or 63241077 x 10^{4} AU * *The distance light travels in one year is 63,241. 077 times the distance from Earth to Sun.

This page titled 2.6.2.5: Light-Years is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.13: Light-Years by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.2.6: Parsec Vs. Light-Year Measurement

Astronomers also employ other types of measurements, specifically the **parsec (pc)**. The parsec is defined as the distance to an object with a parallax angle of 1 arcsecond. This takes us back to our measurements using parallax. When determining the relationship between the *pc* and *ly* , one finds that:**1 pc = 3. 261 light years (ly) = 1. 9174 x 1012 mi = 3. 0857 x 1013 km. M**

The parsec was named from the abbreviations of a **par** allax of one arc **sec** ond. The idea of the parsec was to provide astronomers a method to calculate stellar and other astronomical distances quickly and with ease from their data. The parsec is preferred by astronomers, whereas the light-year is more popular in science texts and everyday use.

This page titled 2.6.2.6: Parsec Vs. Light-Year Measurement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.14: Parsec Vs. Light-Year Measurement by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.3: STELLAR CARTOGRAPHY

2.6.3.1: Stellar and Celestial Object Brightness

2.6.3.2: Magnitude System

2.6.3.3: Celestial cartography

2.6.3.4: Planispheres

2.6.3.5: Star Colors

2.6.3.6: Spectral Type

2.6.3.7: Sample Spectral Types

This page titled 2.6.3: Stellar Cartography is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.6.3.1: Stellar and Celestial Object Brightness

There are a number of other stellar characteristics astronomers and astrophysicists consider. **Luminosity** is a star's total amount of power radiated into space, measured in watts. Think about a light bulb in your home. The output is measured in watts. Most of us use 60 watts tungsten light bulbs or the new LED bulbs with an output of around 60 watts, yet using about 10 watts of power.

How does a light bulb with a luminosity of 60 watts compare to the Sun's output? The accepted luminosity of the Sun is 3.846×10^{26} watts. With such large numbers, the Sun's luminosity is often stated as 1 $L_{\odot}$. The number 1 $L_{\odot}$ is easier to work with than 3.846×10^{26} watts, yet they mean the same quantity.

Apparent Brightness is the amount of light reaching us per unit area, measured in flux. This incoming light varies by the **Inverse Square Law**. Mathematically this is given as:

Apparent Brightness is proportional to $1/d^2$

Let's look at an example many have experienced. You see a light one mile away, and note its brightness. You move away from the light, and you again look at the light. You might think the light should be half as bright. Actually it is ¼ as bright, as given by the Inverse Square Law.

Apparent brightness at 1 mile: $1/1^2 = 1$ Apparent brightness at 2 miles: $1/2^2 = ¼$

We all experience the Inverse Square Law in real life, from driving a car at night to looking at lights on and in buildings and streetlights as we approach a light or move away from the light.

This page titled 2.6.3.1: Stellar and Celestial Object Brightness is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.15: Stellar and Celestial Object Brightness by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.3.2: Magnitude System

The **magnitude system** is a scale to show how bright stars appear. The initial magnitude system was developed by the Greek astronomer, geographer, and mathematician Hipparchus (190 BC to 120 BC). He ranked stars by their apparent brightness, with 1 being the brightest and 6 being barely visible, without the telescope or other optical aid.

He also designed his original magnitude scale, such that the star Polaris, the northern hemisphere's pole star, would have a magnitude of 2. Hipparchus' initial magnitude scale was revised by Norman Pogson in 1856. Pogson specified a 1st magnitude star is 100 times brighter than a 6th magnitude star. Based on Pogson's system, a 1st magnitude star is 2. 512 times as bright as a 2nd magnitude star. With this revision also came inclusion of brighter objects, such as the Sun and Moon, and fainter objects then visible through the telescope. What we see in our sky is called an object's **Apparent Magnitude, M_v**.

Example Apparent Magnitudes, M_v

Object

- Sun
- Moon
- Sirius (brightest star in the sky beyond the Sun)
- Polaris, the northern hemisphere's pole star
- Andromeda Galaxy, M31
- Maximum brightness of dwarf planet Ceres
- Faintest objects visible using 7×50 binocular
- Pluto at maximum brightness
- Faintest objects observable in visible light with 8m Earth telescopes
- Faintest objects observable in visible light with the Hubble Space Telescope

Apparent Magnitude – as seen from Earth

- -26. 74
- -12. 74
- -1. 4
- 1. 98
- 3. 44
- 6. 64
- 9. 50
- 13. 65
- 27. 00
- 31. 50

This page titled 2.6.3.2: Magnitude System is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.16: Magnitude System by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.3.3: Celestial cartography

Celestial cartography, also known as **uranography** or **star cartography**, is the area of astronomy that produces star maps. Most are aware of cartographers who made early and more recent maps of Earth. Both types of cartography were difficult in early history; we were not certain of the exact sizes and shapes of land masses, and there are many, many stars to chart correctly. Many early Earth and star maps were very artistic in nature and truly took not only a scientific mind, but the skills of an amazing artist.

Image of the paining The constellation of Aquarius in the 1603 star map Uranometria By Johann Bayer.

Public Domain

A star map of the Aquarius constellation.

CC BY-SA 3.0 | Image courtesy of Wikimedia Author: Rursus / kstars.

This page titled 2.6.3.3: Celestial cartography is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.6.3.4: Planispheres

Celestial objects are shown on planispheres, star charts and maps, and star atlases, as well as computer and tablet applications and programs. The printed planispheres and charts use different size “dots” to represent the object’s brightness. Many of the new star chart computer and tablet applications and programs are able to show star brightness without using a larger or smaller dot.

The charts, maps, and applications show the object’s apparent magnitude. The apparent magnitude scale is based on what we see, and does not truly compare a star’s true brightness to other stars. So astronomers devised a second scale: **Absolute Magnitude, M**. This scale defines brightness as if we moved all the stars to the same distance; 10 parsecs. These measurements, albeit hypothetical since we cannot move objects to a distance of 10 pc, are the measure of a celestial object’s intrinsic brightness. This allows astronomers and astrophysicists to directly compare star brightness’; *comparing apples to apples* .

Note:

Take our Sun, for example. It is so close, relative to the other stars and celestial objects, that it appears to be very bright, with an apparent magnitude -26. 74 Mv. Yet move the Sun to the hypothetical distance of 10 pc, and the Sun becomes a relatively faint 4. 83 M absolute magnitude object.

Comparing Some Apparent versus Absolute Magnitudes

Object

- Rigel
- Polaris (double star)
- Betelgeuse
- Vega
- Sirius
- Alpha Centauri (triple star)
- Sun
- Andromeda Galaxy, M31
- Black Eye Galaxy, M64

Apparent Magnitude, M_v

- 0. 12
- 1. 98
- 0. 42 (variable, 0. 3 to 1. 2)
- 0. 03
- -1. 4
- -0. 01
- -26. 74
- 3. 44
- 9. 36

Apparent Magnitude, Mv

- -7. 02
- -3. 6
- -2. 99
- 0. 6
- 1. 4
- 4. 3

- 4. 83
- -21. 5
- -21. 7

This page titled 2.6.3.4: Planispheres is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.18: Planispheres by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.3.5: Star Colors

As you look up at the sky at night, you might notice stars are not all white in color. If you happen to look up at a winter sky, you will be greeted by the blue-white Dog Star Sirius in the constellation Canis Major. And nearby is the reddish star Betelgeuse in Orion the Hunter. Contrast these colors to that of our yellow Sun.

Yet star color is an important indicator for astronomers; the color is an indication of a **star's surface temperature** at its **photosphere**. A star's **photosphere temperature** dictates its **color**. It turns out that the hotter stars are **whiter** and **cooler** stars are redder. This is exactly what we observe around us. If you heat a piece of metal, you will note that the hotter metal glows what we call "a white hot. " As the metal cools, it becomes orange and then red. It is still hot if it is red, it is just not as hot as the white-hot metal. Stars are the same.

We classify stars by their colors and thus their **photosphere temperatures**. This is called a star's **Spectral Type**. Stars are categorized by their spectral types. This categorization was first done by**Father Angelo Secchi** in the 1860s and 1870s. In the 1880s at Harvard University, **Edward Pickering** observed and recorded stellar spectra and **Williamina Fleming** classified Pickering's spectra with a system of letters.

In 1901, Harvard astronomer **Annie Jump Cannon** revised the various classifications to give us our more-modern Spectral Types interpretation. One of the final notes in this all-important classification system was **Cecilia Payne Gaposchkin's** 1925 doctoral dissertation, where she showed that spectral types were a sequence in stellar photosphere temperatures.

Harvard astronomer **Henry Norris Russell** was concerned about Payne Gaposchkin's dissertation conclusion that the Sun was primarily composed of hydrogen. This contradicted the theory at the time of the publication of her doctoral dissertation — that Earth could be so different than the Sun. Later, Russell came to the same conclusion, published his own paper, and barely mentioned Payne Gaposchkin's work. Russell is often given credit for this discovery, even though it was not originally his work for which to receive credit.

This page titled 2.6.3.5: Star Colors is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.19: Star Colors by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.3.6: Spectral Type

As noted, from numbers to a series of numbers and letters came the spectral type classification of Annie Jump Cannon. The basic Spectral Types are **O – B – A – F – G – K – M**. Cannon decided to drop several letters and numbers, thus the seemingly unorganized system we employ today. As can be seen in the graphic, the Spectral Type is indicative of the star's color, thus the star's photosphere temperature.

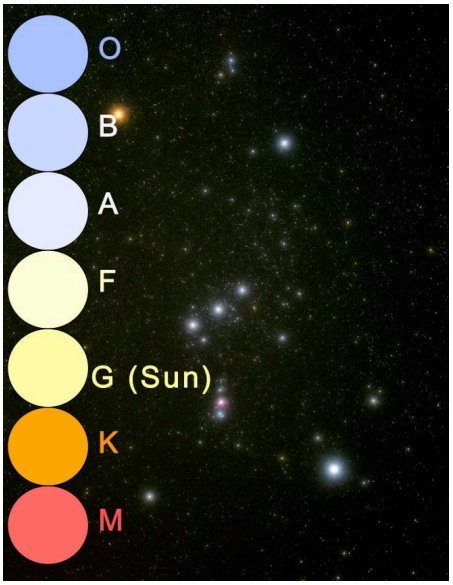


Image courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

This page titled 2.6.3.6: Spectral Type is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.6.3.7: Sample Spectral Types

Trying to remember the Spectral Types in order? A couple of excellent – and time-proven mnemonics to assist you:

O h **B** e **A** **F** ine **G** irl/ **G** uy, **K** iss **M** e **O** h **B** oy, **A** n **F** **G** rade **K** ills **M** e

There are several other Spectral Types, like the carbon RNS stars (additional mnemonic is **R** ight **N** ow **S** weetheart) or the very low temperature L, T, and Y stars (additional mnemonic is **L** ove **T**o **Y** ou).

O

Temperature Range

- >30,000 K

Color

- Blue-blue-white

Stellar Examples

- *Orion Belt Stars (3 stars)*

B

Temperature Range

- 10,000-30,000 K

Color

- Blue-white

Stellar Examples

- *Rigel, Orion's shoulder star*

A

Temperature Range

- 7,500-10,000 K

Color

- White

Stellar Examples

- *Sirius, the Dog Star in CMa*

F

Temperature Range

- 6,000-7,500 K

Color

- Yellow

Stellar Examples

- *Polaris, North Pole Star*

G

Temperature Range

- 5,000-6,000 K

Color

>

- Deep Yellow

Stellar Examples

- *Sun*

K

Temperature Range

- 3,500-5,000 K

Color

>

- Orange

Stellar Examples

>

- *Arcturus, bright star in Boötes*

M

Temperature Range

- ...3,500 K

Color

- Red

Stellar Examples

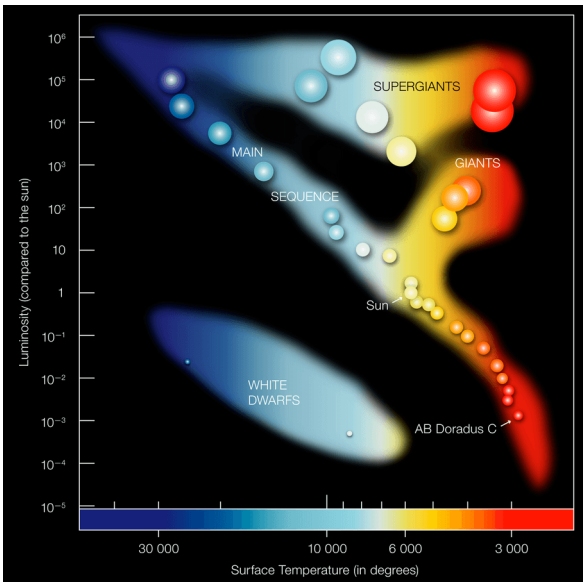
- *Betelgeuse, Orion's armpit*

This page titled 2.6.3.7: Sample Spectral Types is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.21: Sample Spectral Types by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.4: THE HERTZSPRUNG-RUSSELL DIAGRAM

In the **Hertzprung-Russell Diagram** the temperatures of stars are plotted against their luminosities (relative brightness). The position of a star in the diagram provides information about its present stage of life and its mass. Stars that burn hydrogen into helium lie on the diagonal branch, the so-called **Main Sequence**. When a star exhausts all the hydrogen, it leaves the main sequence and becomes a **Red Giant** or a **Supergiant**, depending on its mass. Stars with the mass of the Sun which have burnt all their fuel evolve finally into a **White Dwarf** (left low corner).



In the Hertzprung-Russell diagram the temperatures of stars are plotted against their luminosities.
Image courtesy Wikipedia. Author: ESO. Liscense.

2.6.4.1: The Hertzsprung-Russell Diagram

2.6.4.2: The Four Hertzsprung-Russell Diagram Stellar Groups

2.6.4.3: Types of Star Systems and Stars

This page titled 2.6.4: The Hertzsprung-Russell Diagram is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.6.4.1: The Hertzsprung-Russell Diagram

The Hertzsprung-Russell diagram, also referred to as the H-R Diagram, is a graph of stellar absolute magnitude or luminosity versus their spectral type or photosphere temperature. The H-R Diagram was created circa 1910 by Danish astronomer **Ejnar Hertzsprung** and American astronomer **Henry Norris Russell**. Plotting stellar luminosity versus color (temperature) allowed Hertzsprung Russell to see patterns in stars, from groupings of stars to how stars change over time, what is called stellar evolution. Looking at the H-R Diagram, you will first note that numerous stars are plotted. Imagine millions of stars, that is what makes up the diagram itself.

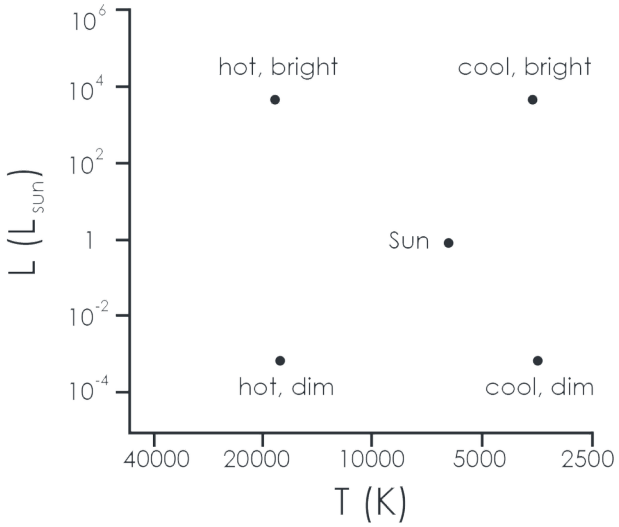


Image courtesy of Florida State College at Jacksonville.

Hertzsprung-Russell diagram. A plot of luminosity (absolute magnitude) against the colour of the stars ranging from the high-temperature blue-white stars on the left side of the diagram to the low temperature red stars on the right side. Converted to png and compressed with pngcrush.

CC BY-SA 2. 5 | Image courtesy of Richard Powell.

The graph also shows the colors of the plotted stars — note the stars on the left of the graph are more blue-white, towards the center yellowish, and on the right red to almost a purple color. That indicates the stellar temperature.

As you go higher on the graph, up the Y-axis, the stars are brighter — the stellar luminosity. This can be plotted as the L_{sun} with the $L_{sun} = 1$. It can also be plotted as Absolute Magnitude, M.

This page titled 2.6.4.1: The Hertzsprung-Russell Diagram is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.22: The Hertzsprung-Russell Diagram by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.4.2: The Four Hertzsprung-Russell Diagram Stellar Groups

The H-R Diagram plots stars such that there are four major groups.

The **Supergiants** are cool stars, which are very large and very bright. They are located towards the top right of the graph.

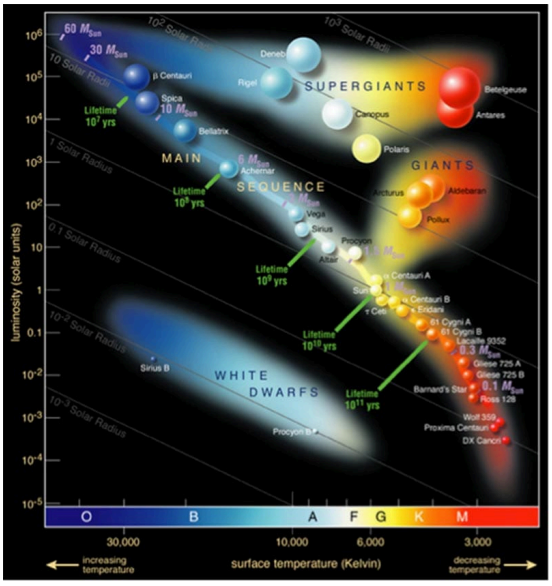
The **Giants** are cool stars, which are a little smaller and dimmer than the Supergiants.

The **White Dwarfs** are very hot stars, which are small in size and relatively dim. They are found in the lower left of the H-R Diagram.

The **Main Sequence** is a band of stars, which includes most of the stars, like our Sun. These are usually smaller stars, often dwarf stars.

The H-R Diagram can also show a star's life, referred to as Stellar Evolution. During the 1930s, it was found for stars with uniform chemical composition, there is a relationship between stellar mass and the star's luminosity and diameter.

Other important derivations of the Hertzsprung-Russell diagram can be used to show different data, for example classes of stars, protostars, brown dwarf objects, or graphs known as color index and color magnitude diagram.



CC BY 4.0 | Image courtesy of European Southern Observatory (ESO).

This page titled 2.6.4.2: The Four Hertzsprung-Russell Diagram Stellar Groups is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.23: The Four Hertzsprung-Russell Diagram Stellar Groups by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.6.4.3: Types of Star Systems and Stars

There are numerous types of stars and star systems throughout the Universe. We will touch on a few of these; some of the more common stars and star systems and some of the strangest of the stars.

Occasionally, astronomers will look through their telescopes and see two or more stars which happen to line up so that we see what appears to be a double star system, yet these stars are not physically related to each other. These are called **Optical Doubles** and simply happen by chance. Such occurrences are more infrequent than one would think. Systems in which physically associated star systems are made up of two stars are **Binary Stars**. And a **Visual Binary** is a pair of stars that we can see orbiting each other.

An **Eclipsing Binary Star**, or EBS, is a star system in which the star's orbital plane is such that it lies in the line of sight of telescopes on Earth so that the stars eclipse each other. Also called**eclipsing variables**, as one star orbits the more massive central star, the observers on Earth see eclipses occur.

First, the orbiting star will eclipse the more massive central star, eclipsing or blocking some of the light from the more massive and larger central star. Then, the smaller orbiting star will pass behind the more massive central star, usually eclipsing or blocking all of the light from the orbiting star. These eclipses can occur often if the orbiting star is close enough to the more massive central star. Or it may take weeks or months, perhaps even years, between eclipses depending on the EBS.

This diagram shows an eclipsing Binary Star Mockup. The more massive, larger yellow central star is being orbited by a second, smaller orange star. A series of eclipses will occur. If the view is more overhead or a bird's-eye view, then no eclipses will be seen from the Earth-based observer.

Image courtesy of Florida State College at Jacksonville.

This page titled 2.6.4.3: Types of Star Systems and Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 11.24: Types of Star Systems and Stars by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

CHAPTER 2.7: STELLAR EVOLUTION

I like the night. Without the dark, we'd never see the stars.

Stephenie Meyer
Twilight

"I think that we sare like stars. Something happens to burst us open; but when we burst open and think we are dying; we're actually turning into a supernova. And then when we look at ourselves again, we see that we're suddenly more beautiful than we ever were before!"

C. JoyBell C.

Learning Objectives

Upon completion of this module, the student will be able to:Upon completion of this module, the student will be able to:

- Define the biology-like terms astronomers use when describing stars.
- Identify the characteristics of low-mass stars.
- Describe how white dwarfs, novae, and Type 1a supernovae are formed in low-mass stars.
- Identify the characteristics of high-mass stars.
- Describe how neutron stars, supernovae, and black holes are formed in high-mass stars.
- Define GRBs.
- Describe GRB characteristics

This module looks at how stars form and develop over time, including the less-massive stars, like the Sun, and more-massive stars.

2.7.1: Birth

2.7.1.1: Star Life

2.7.1.2: Stellar Birth

2.7.1.3: Stellar Mass

2.7.1.4: Stellar Evolution

2.7.2: Life Cycle - Low-Mass Stars

2.7.2.1: Low-Mass Stars

2.7.2.2: White Dwarf

2.7.2.3: Planetary Nebula

2.7.2.4: White Dwarfs and Neighbors

2.7.2.5: Nova, Novae, and Supernova

2.7.3: Life Cycle - High-Mass Stars

2.7.3.1: High-Mass Stars

2.7.3.2: Neutron stars

2.7.3.3: Neutron Star and Companion Star Scenario

2.7.3.4: High-Mass Star Stellar Endings

2.7.3.5: Gamma-Ray Bursts (GRBs)

2.7.3.6: Checking Out the Theory

This page titled Chapter 2.7: Stellar Evolution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12: Stellar Evolution by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.1: BIRTH

Stars begin their lives in Stellar Nurseries such as the one imaged here.



Image courtesy Wikipedia. Author: ESO/APEX/DSS2/ SuperCosmos/ Deharveng(LAM)/ Zavagno(LAM). Liscense.

2.7.1.1: Star Life

2.7.1.2: Stellar Birth

2.7.1.3: Stellar Mass

2.7.1.4: Stellar Evolution

This page titled 2.7.1: Birth is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.7.1.1: Star Life

There are specific terms in biology, which deal with life: birth, growing up, getting old, death, and even evolution. For the sake of convenience and familiarity, astronomers use similar terminology when dealing with stars. This module will focus on astronomical concepts, such as stars being born or stellar birth, stellar nurseries, stellar life, stars growing old, star death, and stellar evolution. **Note that stars are not alive!** These are simply terms of familiarity. Stars are formed in **interstellar medium**, which is filled with hydrogen, helium, and dust. This interstellar medium is cold and dense. These areas in space are also called **Molecular Clouds, Interstellar Clouds, or Stellar Nurseries** (more than one star).

One debate that transpired in the 1800s into the early 1900s was the question of what composed the stars. A doctoral student at Harvard, **Cecelia Payne-Gaposhkin**, noted in her 1925 doctoral dissertation that stars were made of hydrogen. Not all agreed with Payne-Gaposhkin's findings, yet as more data was collected and researched, her theory was proven correct a few short years after she first proposed the theory.

So how are stars 'born'? That is, what are the steps to a stellar birth? First, hydrogen, helium, and dust – the interstellar medium – has to accumulate. The interstellar medium starts to collapse due to gravity; these various atoms and dust particles attract each other, forming larger and larger clumps, attracting more gas and dust. As the interstellar medium collapses, the density, pressure, and temperature within the interstellar medium increases. Think of it like this: as the density increases, more particles are striking each other, generating more energy and heat.

This page titled 2.7.1.1: Star Life is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.1: Star Life by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.1.2: Stellar Birth

There are relationships between pressure, temperature, and volume in a gas like interstellar medium:

Where:

- This relationship shows that as the pressure (P) increases, the temperature (T) increases and the volume decreases (V).

  As the pressure, density, and temperature increases, no thermal (heat) energy can easily escape. The rising pressure and density leads to the formation of a Protostar, where the core is not yet undergoing fusion. The star 'turns on' – *a star is born* – when the protostar's core temperature reaches 10,000,000 K; fusion begins as the proton-proton cycle.



Public Domain | Image courtesy of NASA.

This page titled 2.7.1.2: Stellar Birth is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.2: Stellar Birth by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.1.3: Stellar Mass

The type and life of the new star depends on the initial amount of material present: its stellar birth weight. Astronomers group stars as **Low-mass**, <1. 33 m $_{\odot}$; **Intermediate-mass**, 1. 33 m $_{\odot}$ to 4 m $_{\odot}$; and **High-mass** :>4 m $_{\odot}$. (Recall **m** $_{\odot}$ refers to the mass of the Sun, the Sun = 1 m $_{\odot}$) Each of these star groups has different lives and deaths. In this module, we will examine low- and high-mass stars.

Stellar Mass versus Star Life range, Spectral Class

Mass; m $_{\odot}$

60 m $_{\odot}$

30 m $_{\odot}$

10 m $_{\odot}$

3 m $_{\odot}$

1. 5 m $_{\odot}$

1 m $_{\odot}$

0. 1 m $_{\odot}$

Time; Year

3 million years

11 million years

32 million years

370 million years

3 billion years

10 billion years

1000's of billion years

Spectral Type

O3

O7

B4

A5

F5

G2 (Sun)

This page titled 2.7.1.3: Stellar Mass is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.3: Stellar Mass by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.1.4: Stellar Evolution



CC BY 4.0 | Image courtesy of European Southern Observatory (ESO).

Stellar evolution is the process by which a star experiences a sequence of drastic variations during its stellar life. Depending on the initial stellar mass, this lifetime ranges from a few million years for the most massive to trillions of years for the least massive, which is considerably longer than the age of the Universe at 13. 7 billion years.

This page titled 2.7.1.4: Stellar Evolution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.4: Stellar Evolution by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.2: LIFE CYCLE - LOW-MASS STARS

2.7.2.1: Low-Mass Stars
2.7.2.2: White Dwarf
2.7.2.3: Planetary Nebula
2.7.2.4: White Dwarfs and Neighbors
2.7.2.5: Nova, Novae, and Supernova

This page titled 2.7.2: Life Cycle - Low-Mass Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.7.2.1: Low-Mass Stars

Low-Mass Stars fuse hydrogen into helium, the proton-proton cycle. The classic low-mass star is the Sun. Low-mass stars have large convection zones when compared to intermediate- and high-mass stars. In **very low-mass stars**, the Convection Zone goes all the way to the star's core!Over time, a low-mass star consumes all of the hydrogen in its core – *what happens now?* Think about a car that runs out of gas. Unless you put more gas in the tank, the car will not run. The same is true with stars; no more fuel in the stellar core will end the stellar fusion process.

When the core hydrogen is used up and no more nuclear fusion occurs, the star's outer stellar layers expand and the core shrinks. At this point, the star becomes a **Subgiant Star**. The star's outer layers continue to expand and the star brightens. The star then becomes a **Red Giant Star**. In about one billion years, the Sun will begin its Red Giant phase.

So how does the star expand and get brighter if it has ceased to fuse hydrogen in its collapsing core? Helium is left in the star's core and gravity continues to shrink the core and the surrounding layers. These surrounding layers contain hydrogen; the surrounding shell of hydrogen begins to fuse, called **Hydrogen Shell Burning**. The now red giant star is now larger than out to the orbit of Mars.

Next, helium fusion begins, for which very high temperatures are required.

$$^{3}\mathrm{He} \rightarrow {}^{12}\mathrm{C}$$

This occurs when three helium nuclei are fused into one carbon.

The star begins to dim and shrink in size. Now the cycle *reverses* ; hydrogen and helium are fused at a tremendous rate. This process takes a few million years. Now the star's outer layers flow outward from the star, with the star's core is mostly composed of carbon (from the fusion of the remaining helium into carbon).

This page titled 2.7.2.1: Low-Mass Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.5: Low-Mass Stars by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.2.2: White Dwarf

The star can form a White Dwarf, then a planetary nebula. A **White Dwarf** is the inert stellar core that remains after a star has ended all core nuclear fusion. All nuclear fusion is over, that is, all available hydrogen in the star's core has been fused into helium and the helium into carbon. This White Dwarf stellar core is incredibly dense; in fact if you were to weigh a White Dwarf, it would weigh about 5 tons per teaspoon. It is extremely dense due to the collapse of the star's core; gravity has created an incredible density.

Optical Image (left) and a portion of the Hubble Space Telescope observation (right) of the globular cluster M4. The white dwarfs are circled in the HST image.

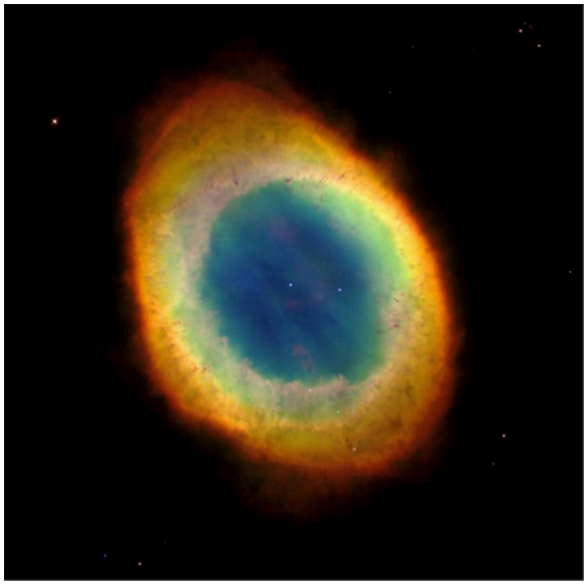
Public Domain | Image courtesy of NASA.

This page titled 2.7.2.2: White Dwarf is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.7.2.3: Planetary Nebula

The star's **planetary nebula** is created when a shell of gas is ejected by the star and illuminated by its central, White Dwarf. The term planetary nebula was coined by astronomer William Herschel in the 1780s as he was observing these objects. He brought attention to their round shapes, which reminded Herschel of planets. Planetary nebula are only visible tens of thousands of years, a short period in the life of a low-mass star. We see them due to the ultraviolet energy released by the White Dwarf. Once the White Dwarf cools and fades, neither the planetary nebula nor White Dwarf is visible. What remains is an expanding cloud of gas and dust and a carbon sphere.

Note: In the end, all that remains is a cold, dark mass composed mainly of carbon. These are occasionally referred to as **black dwarfs**, although there is debate if the Universe is old enough for any black dwarfs to exist.



Public Domain | Image courtesy of NASA / EVA.

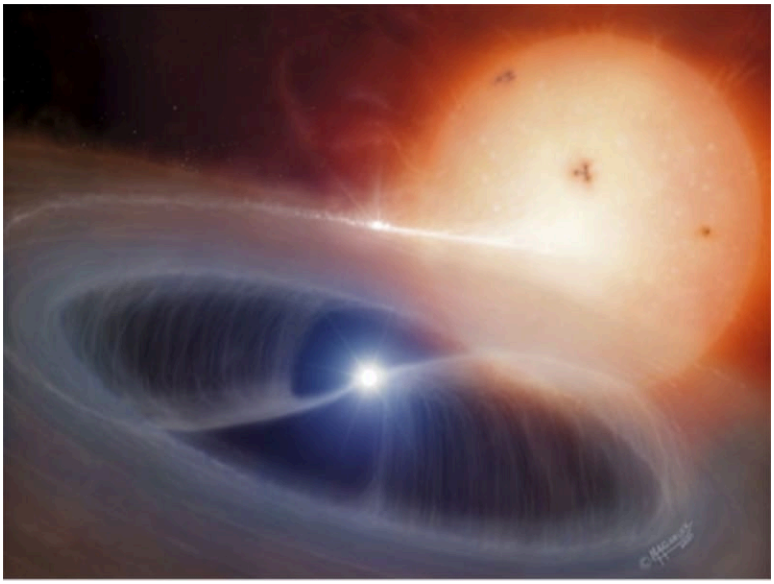


Public Domain | Image courtesy of NASA.

This page titled 2.7.2.3: Planetary Nebula is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

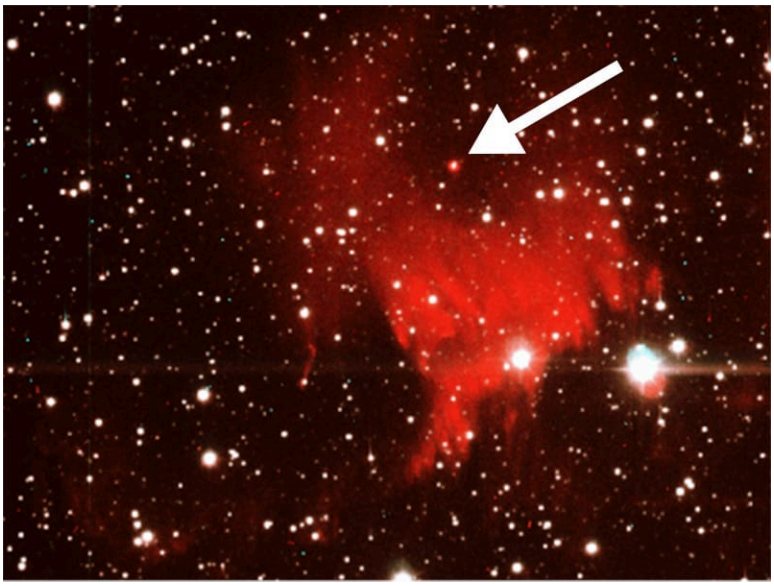
2.7.2.4: White Dwarfs and Neighbors

White Dwarf surrounded by its accretion disk. The star being cannibalized) is in the upper right. Note the differences between the White Dwarf's and the star's color and size. White Dwarf surrounded by its accretion disk. The star being cannibalized) is in the upper right. Note the differences between the White Dwarf's and the star's color and size.



Public Domain | Image courtesy of NASA.

The nova—indicated by the white arrows, provides the illumination to allow the reddish nebula to be seen. When the nova fades from view, the nebula will lose its illumination source, so it will no longer be seen.



Nova Cygni 1992 | NASA Hubble Space TelescopePublic Domain | Image courtesy of NASA.

This page titled 2.7.2.4: White Dwarfs and Neighbors is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.8: White Dwarfs and Neighbors by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.2.5: Nova, Novae, and Supernova

The word nova comes from the Latin, for ‘new. ’ Novae can brighten as much as **100,000 Suns, 100,000L $_{sun}$** . Material from the Nova fusion event radiates outward from the White Dwarf; we can view this outward radiating material as a planetary nebula. The nova process can repeat itself, *but not infinitely* . When the White Dwarf accreting mass reaches the low-mass star 1. 33M $_{sun}$ limit, a **Type 1a Supernova** is possible. First the White Dwarf begins to collapse, with fast, quick heating. Carbon fusion begins (12 in the stellar core) and then the Star “ignites,” exploding as a White Dwarf Supernova. These are usually one-time events.

Type 1a supernovae produce elements up to iron on the atomic table, and also produce elements heavier than iron, like gold, silver, and uranium. These type of supernovae have the same characteristics, such as how bright they will become and length of maximum brightness. Therefore astronomers can use them as a “standard candle,” that is a standard brightness.

Supernovae may also produce **cosmic rays**, which are composed of electrons, protons and neutrons and move at close to the speed of light. **Ultra-high energy cosmic ray particles** were discovered in 2005. These are the brightest and fastest radio blasts ever seen on the sky, seen as radio light that appear more than 1000 times brighter than the Sun and almost a million times faster than normal lightning.

The very bright star in the lower left is Supernova SN1994D in Spiral Galaxy NGC4526. Compare the brightness of the supernova to its home galaxy.



Public Domain | Image courtesy of NASA.

This page titled 2.7.2.5: Nova, Novae, and Supernova is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.9: Nova, Novae, and Supernova by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.3: LIFE CYCLE - HIGH-MASS STARS

2.7.3.1: High-Mass Stars
2.7.3.2: Neutron stars
2.7.3.3: Neutron Star and Companion Star Scenario
2.7.3.4: High-Mass Star Stellar Endings
2.7.3.5: Gamma-Ray Bursts (GRBs)
2.7.3.6: Checking Out the Theory

This page titled 2.7.3: Life Cycle - High-Mass Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

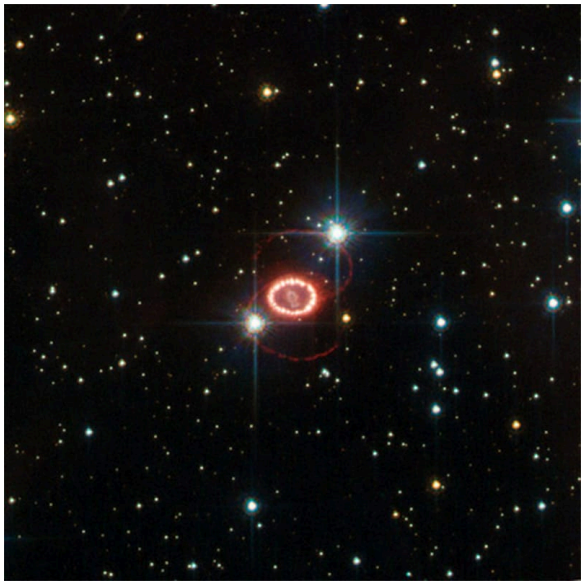
2.7.3.1: High-Mass Stars

High-Mass Stars have a very-short life when compared to low-mass and intermediate-mass stars. In high-mass stars, the fusion process includes heavy elements, the CNO Cycle; **C** arbon- **N**itrogen- **O** xygen:**H_2 → He → C → N → O → ... → Fe**

High-mass star characteristics include:

- An example of a high-mass star is the red supergiant Betelgeuse; it is 500 times the size of our Sun.

  This supernova occurred about 168,000 *ly* from Earth.

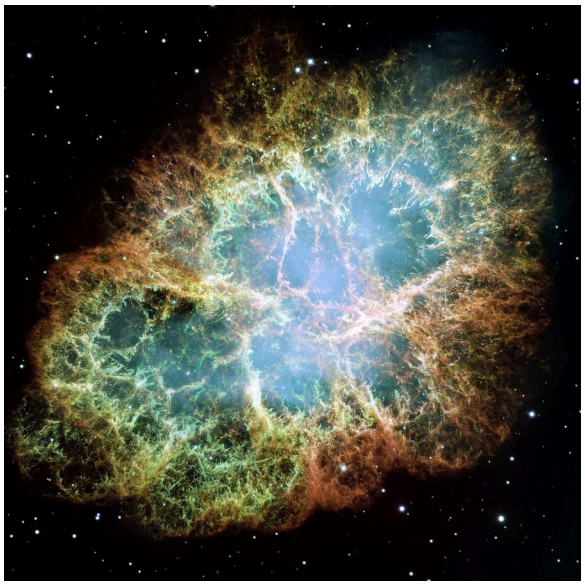


CC BY 3.0 | Image courtesy NASA and ESA

A **supernova** is an explosion of a star that briefly outshines an entire galaxy, radiating as much energy as an ordinary star like the Sun over its entire lifetime. The supernova will fade from view over several weeks to months.

The star's core collapses in an instant into a sphere; it simply cannot withstand the immense pressure and gravitational attraction. At collapse a sphere of Neutrons remains. Heavier elements are now produced due to the supernova event, and enormous amounts of energy released. An average high-mass star supernovae within 30 *ly* of Earth pose a danger to life on Earth, due to the energy and particles released in this type of supernova event. Particles released by the supernova are very harmful to life as we know it. It would be much like being close to an atomic bomb going off.

The Crab Nebula ; M1, a Supernova Remnant.



Public Domain | Image courtesy of NASA.

This page titled 2.7.3.1: High-Mass Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.10: High-Mass Stars by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

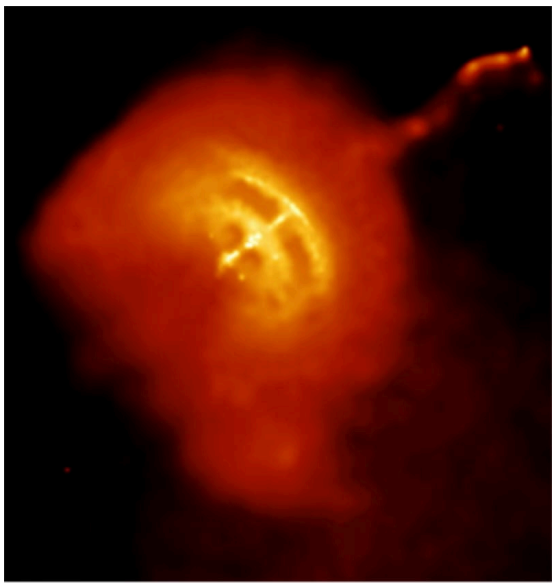
2.7.3.2: Neutron stars

Neutron stars are the spheres of neutrons created by the collapse of the iron core in a massive star supernova; roughly **10 miles** in diameter. The concept of a neutron star was first proposed in 1934 by Walter Baade and Fritz Zwicky, a year after the announcement of the discovery of neutrons. Neutron stars are composed completely of neutrons, n0; neutron stars are dim and intensely hot, and held together due to the gravitational attraction.

There are objects related to neutron stars. **Binary Neutron Stars** are two neutron stars orbiting each other. It is estimated that about 5% of all neutron stars are a part of a binary system. Pulsars – short for pulsating radio star – are rapidly rotating radio source and a type of neutron star. **Pulsars** were discovered in 1967 by Joycelin Bell and initially called LGM – ***Little Green Men*** – due to the regular electromagnetic radiation released from the pulsar. A small fraction of pulsars only emit gamma rays.

Image of Neutron star RX J185635-3754, the first direct observation of a neutron star in visible light.

Public Domain | Image courtesy of NASA.



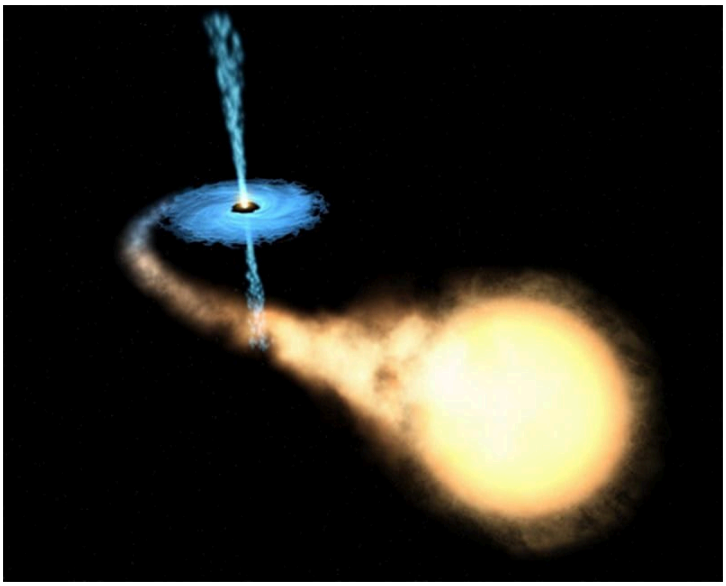
Public Domain | Image courtesy of NASA.

This page titled 2.7.3.2: Neutron stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.11: Neutron stars by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.3.3: Neutron Star and Companion Star Scenario

Artist's concept of an X-ray binary—the donor star (right) and the star being cannibalized (shown in blue on the left); the X-ray star the Accretor star.



Public Domain | Image courtesy of NASA.

Close binary systems of neutron stars and companion stars are much like a White Dwarf and Companion Binary star scenario. An accretion disk can form from material taken from the companion star by the neutron star. Neutron Stars are much hotter and more luminous than the White Dwarf scenario; they become a powerful X-ray source, called an **X-ray Binary**. These type of objects emit bursts of energy and are called **X-ray Bursters**.

This page titled 2.7.3.3: Neutron Star and Companion Star Scenario is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.12: Neutron Star and Companion Star Scenario by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.3.4: High-Mass Star Stellar Endings

There are two possible stellar ending scenarios, depending on the star's original mass. **If the star is less than three solar masses, <3 M_{sun} , this is the neutron star limit.** If it is a single star (*not* part of a binary star system), the star will eventually cool as a large cinder in space. Yet if the star is part of a **binary system**, the star will continue accreting material, jetting energy into space until all of the stellar fuel is exhausted.

If the star is greater than three solar masses, >3 M_{sun} , then the ultimate cosmic extinction occurs: a Black Hole. There is a discussion among astronomers how massive of a star is needed to form a black hole; ranges from >3 M_{sun} to 25 M_{sun} . Black holes originate as the high-mass star's iron core collapses, just prior to the star going Supernova. This is called a **Type 2 Supernova**.



Public Domain | Image courtesy of NASA / JPL.

With a black hole scenario, a catastrophic collapse of the star's core continues. This does not allow matter, or even light, to escape a certain point, called the **Event Horizon**. That is due to the fact that the escape velocity from a black hole is greater than the speed of light. The event horizon size is called the **Schwarzchild radius**. Black holes eject energy back out when overloaded with matter flowing into the black hole. Eventually, black holes will crush to a final, infinitely dense and small point called a **Singularity**.

Questions astronomers are asking about black holes include:

- The challenge with answering these questions at this time is that we can never know what happens inside a black hole because we cannot get the data out.

  We cannot see black holes directly, so how do we know they exist? There are several types of evidence of black holes. First, we see close binaries with gas flowing into nothing. The extreme Doppler shifts of stars has been observed, with apparently nothing in the area of the star to cause the extreme Doppler shift. A number of warped star fields have been observed, called **gravitational lensing**. And evidence that is somewhat direct occurs when a gas is pulled into a black hole by its strong gravitational force; the gas heats up and radiates back out into space.

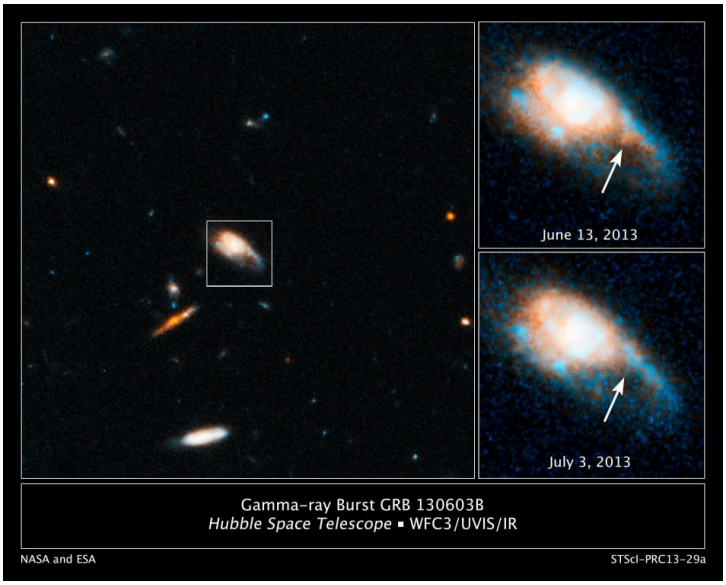
This page titled 2.7.3.4: High-Mass Star Stellar Endings is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.13: High-Mass Star Stellar Endings by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.3.5: Gamma-Ray Bursts (GRBs)

During the 1960s, there was great concern among the Superpower Countries that nuclear testing or even a nuclear attack could occur without notice. The United States developed a series of satellites to monitor gamma radiation pulses emitted by nuclear weapons tested or used in space. Gamma radiation outbursts were observed and determined to be coming from space. Further research showed these outbursts were fairly evenly distributed throughout the sky. **Gamma-ray bursts**, or **GRBs**, are flashes of gamma radiation related to extreme energetic explosions detected in very distant galaxies. GRBs are the brightest electromagnetic events known to occur in the Universe. These GRBs can range from ten milliseconds to several minutes. The initial GRB is usually followed by a longer lived afterglow, emitted at wavelengths such as X-ray, UV, visual, microwave, and radio wavelengths.

A number of satellites have been built to observe GRBs. The Compton Gamma Ray Observatory determined that GRBs were from outside our Galaxy. (There is a class of gamma-ray objects within our galaxy, but not with the extreme power of GRBs). Some have speculated that GRBs are possibly at the edge of the early Universe and the death throes of extremely massive stars, which only lasted about 1 million years. These stars appear to eject Gamma Rays after a **hypernova event**, an extreme supernova which produces gamma radiation. GRBs also appear to come from within stellar nurseries.



Public Domain | Image courtesy of NASA, ESA, N. Tanvir (University of Leicester), A. Fruchter (STScI), and A. Levan (University of Warwick).

Consider this...

"The nitrogen in our DNA, the calcium in our teeth, the iron in our blood, the carbon in our apple pies were made in the interiors of collapsing stars. We are made of starstuff. "

—Dr. Carl Sagan *Cosmos* (2002)

This page titled 2.7.3.5: Gamma-Ray Bursts (GRBs) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 12.14: Gamma-Ray Bursts (GRBs) by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.7.3.6: Checking Out the Theory

Learning Objectives

By the end of this section, you will be able to:

- Explain how the H–R diagram of a star cluster can be related to the cluster's age and the stages of evolution of its stellar members
- Describe how the main-sequence turnoff of a cluster reveals its age

In the previous section, we indicated that open clusters are younger than globular clusters, and associations are typically even younger. In this section, we will show how we determine the ages of these star clusters. The key observation is that the stars in these different types of clusters are found in different places in the H–R diagram, and we can use their locations in the diagram in combination with theoretical calculations to estimate how long they have lived.

H–R Diagrams of Young Clusters

What does theory predict for the H–R diagram of a cluster whose stars have recently condensed from an interstellar cloud? Remember that at every stage of evolution, massive stars evolve more quickly than their lower-mass counterparts. After a few million years ("recently" for astronomers), the most massive stars should have completed their contraction phase and be on the main sequence, while the less massive ones should be off to the right, still on their way to the main sequence. These ideas are illustrated in Figure 22.8, which shows the H–R diagram calculated by R. Kippenhahn and his associates at Munich University for a hypothetical cluster with an age of 3 million years.

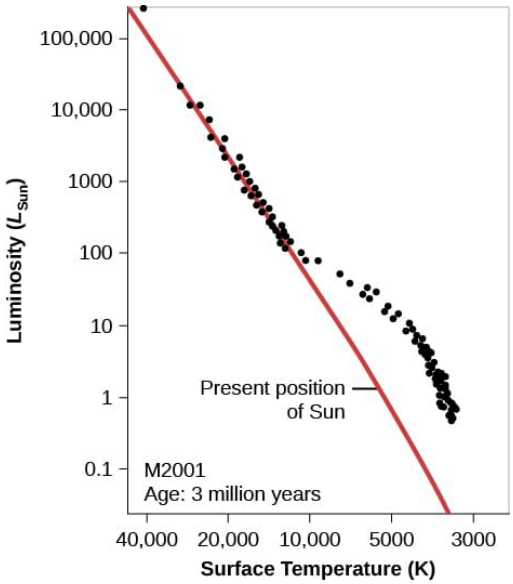


Figure 22.8 : Young Cluster H–R Diagram. We see an H–R diagram for a hypothetical young cluster with an age of 3 million years. Note that the high-mass (high-luminosity) stars have already arrived at the main-sequence stage of their lives, while the lower-mass (lower-luminosity) stars are still contracting toward the zero-age main sequence (the red line) and are not yet hot enough to derive all of their energy from the fusion of hydrogen.

There are real star clusters that fit this description. The first to be studied (in about 1950) was NGC 2264, which is still associated with the region of gas and dust from which it was born (Figure 22.9).



Figure 22.9 : Young Cluster NGC 2264. Located about 2600 light-years from us, this region of newly formed stars, known as the Christmas Tree Cluster, is a complex mixture of hydrogen gas (which is ionized by hot embedded stars and shown in red), dark obscuring dust lanes, and brilliant young stars. The image shows a scene about 30 light-years across. (credit: ESO)

The NGC 2264 cluster's H–R diagram is shown in Figure 22.10. The cluster in the middle of the Orion Nebula (shown in Figure 21.4 and Figure 21.5) is in a similar stage of evolution.

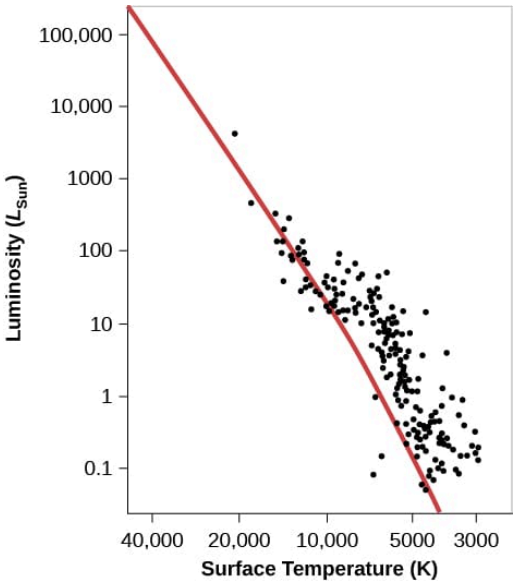


Figure 22.10 : NGC 2264 H–R Diagram. Compare this H–R diagram to that in Figure 22.8; although the points scatter a bit more here, the theoretical and observational diagrams are remarkably, and satisfyingly, similar.

As clusters get older, their H–R diagrams begin to change. After a short time (less than a million years after they reach the main sequence), the most massive stars use up the hydrogen in their cores and evolve off the main sequence to become red giants and supergiants. As more time passes, stars of lower mass begin to leave the main sequence and make their way to the upper right of the H–R diagram.

Link to Learning

To see the evolution of a star cluster in a dwarf galaxy, you can watch this brief animation of how its H–R diagram changes.

Figure 22.11 is a photograph of NGC 3293, a cluster that is about 10 million years old. The dense clouds of gas and dust are gone. One massive star has evolved to become a red giant and stands out as an especially bright orange member of the cluster.



Figure 22.11 : NGC 3293. All the stars in an open star cluster like NGC 3293 form at about the same time. The most massive stars, however, exhaust their nuclear fuel more rapidly and hence evolve more quickly than stars of low mass. As stars evolve, they become redder. The bright orange star in NGC 3293 is the member of the cluster that has evolved most rapidly. (credit: ESO/G. Beccari)

Figure 22.12 shows the H–R diagram of the open cluster M41, which is roughly 100 million years old; by this time, a significant number of stars have moved off to the right and become red giants. Note the gap that appears in this H–R diagram between the stars near the main sequence and the red giants. A gap does not necessarily imply that stars avoid a region of certain temperatures and luminosities. In this case, it simply represents a domain of temperature and luminosity through which stars evolve very quickly. We see a gap for M41 because at this particular moment, we have not caught a star in the process of scurrying across this part of the diagram.

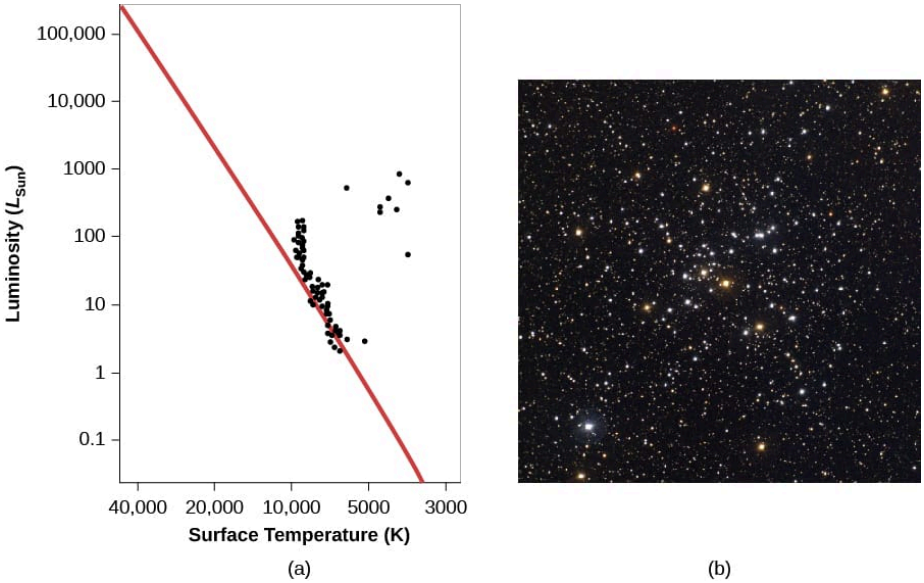


Figure 22.12 : Cluster M41. (a) Cluster M41 is older than NGC 2264 (see Figure 22.10) and contains several red giants. Some of its more massive stars are no longer close to the zero-age main sequence (red line). (b) This ground-based photograph shows the open cluster M41. Note that it contains several orange-color stars. These are stars that have exhausted hydrogen in their centers, and have swelled up to become red giants. (credit b: modification of work by NOAO/AURA/NSF)

H–R Diagrams of Older Clusters

After 4 billion years have passed, many more stars, including stars that are only a few times more massive than the Sun, have left the main sequence (Figure 22.13). This means that no stars are left near the top of the main sequence; only the low-mass stars near the bottom remain. The older the cluster, the lower the point on the main sequence (and the lower the mass of the stars) where stars begin to move toward the red giant region. The location in the H–R diagram where the stars have begun to leave the main sequence is called the main-sequence turnoff.

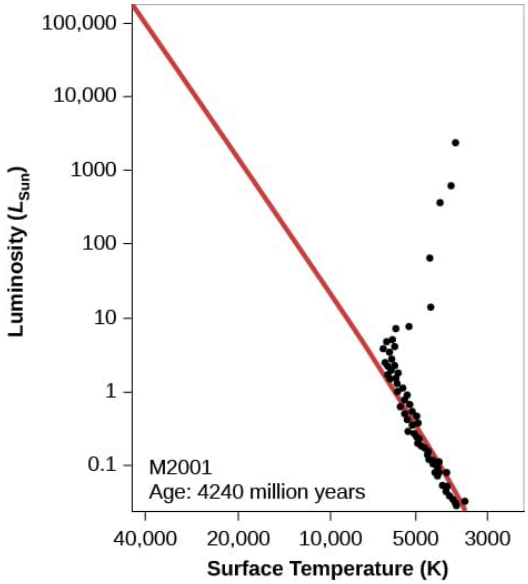


Figure 22.13 : H–R Diagram for an Older Cluster. We see the H–R diagram for a hypothetical older cluster at an age of 4.24 billion years. Note that most of the stars on the upper part of the main sequence have turned off toward the red-giant region. And the most massive stars in the cluster have already died and are no longer on the diagram.

The oldest clusters of all are the globular clusters. Figure 22.14 shows the H–R diagram of globular cluster 47 Tucanae. Notice that the luminosity and temperature scales are different from those of the other H–R diagrams in this chapter. In Figure 22.13, for example, the luminosity scale on the left side of the diagram goes from 0.1 to 100,000 times the Sun's luminosity. But in Figure 22.14, the luminosity scale has been significantly reduced in extent. So many stars in this old cluster have had time to turn off the main sequence that only the very bottom of the main sequence remains.

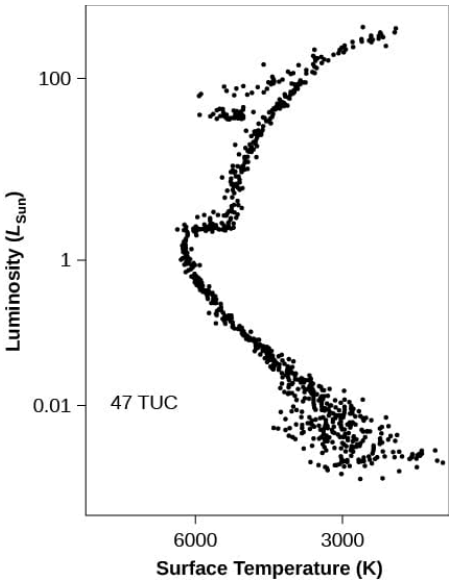


Figure 22.14 : Cluster 47 Tucanae. This H–R diagram is for the globular cluster 47. Note that the scale of luminosity differs from that of the other H–R diagrams in this chapter. We are only focusing on the lower portion of the main sequence, the only part where stars still remain in this old cluster.

Link to Learning

Check out this brief NASA video with a 3-D visualization of how an H–R diagram is created for the globular cluster Omega Centauri.

Just how old are the different clusters we have been discussing? To get their actual ages (in years), we must compare the appearances of our *calculated* H–R diagrams of different ages to *observed* H–R diagrams of real clusters. In practice, astronomers use the position at the top of the main sequence (that is, the luminosity at which stars begin to move off the main sequence to become red giants) as a measure of the age of a cluster (the main-sequence turnoff we discussed previously). For example, we can compare the luminosities of the brightest stars that are still on the main sequence in Figure 22.10 and Figure 22.13.

Using this method, some associations and open clusters turn out to be as young as 1 million years old, while others are several hundred million years old. Once all of the interstellar matter surrounding a cluster has been used to form stars or has dispersed and moved away from the cluster, star formation ceases, and stars of progressively lower mass move off the main sequence, as shown in Figure 22.10, Figure 22.12, and Figure 22.13.

To our surprise, even the youngest of the globular clusters in our Galaxy are found to be older than the oldest open cluster. All of the globular clusters have main sequences that turn off at a luminosity less than that of the Sun. Star formation in these crowded systems ceased billions of years ago, and no new stars are coming on to the main sequence to replace the ones that have turned off (see Figure 22.15).

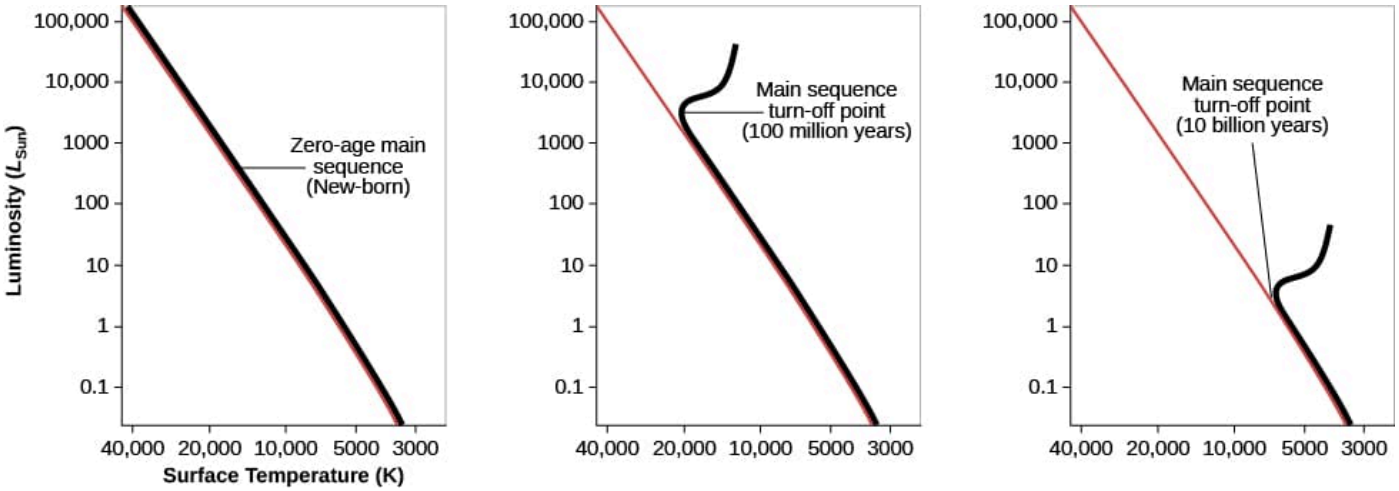


Figure 22.15 : H–R Diagrams for Clusters of Different Ages. This sketch shows how the turn-off point from the main sequence gets lower as we make H–R diagrams for clusters that are older and older.

Indeed, the globular clusters are the oldest structures in our Galaxy (and in other galaxies as well). The youngest have ages of about 11 billion years and some appear to be even older. Since these are the oldest objects we know of, this estimate is one of the best limits we have on the age of the universe itself—it must be at least 11 billion years old. We will return to the fascinating question of determining the age of the entire universe in the chapter on The Big Bang.

This page titled 2.7.3.6: Checking Out the Theory is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 22.4: Checking Out the Theory by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/astronomy.

CHAPTER 2.8: GALAXIES

"There are at least as many galaxies in our observable Universe as there are stars in our galaxy. "

Martin Rees
Lord Rees of Ludlow
Astronomer
(1942 -)

Learning Objectives

Upon completion of this module, the student will be able to:Upon completion of this module, the student will be able to:

- Define what a galaxy is and the broad characteristics of galaxies, including galaxy types.
- Explain the history of observing galaxies.
- Recognize Hubble's Law and its implications in astronomy.
- Differentiate among galaxy-related objects, including Quasars, Blazars, galactic grouping, WIMPs, gravitational lensing

This module presents an overview of the islands of stars found throughout the Universe: the galaxies. Like many objects we have discovered, galaxies vary in characteristics such as size, shape, and age.

2.8.1: The Practice of Studying Galaxies

2.8.1.1: What do you think?

2.8.1.2: Galaxies

2.8.1.3: The History of Studying Galaxies

2.8.1.4: Hubble's Major Contributions

2.8.1.5: Hubble's Law

2.8.2: Galactic Classification

2.8.2.1: Classification of Galaxies

2.8.2.2: Elliptical Galaxies

2.8.2.3: Spiral Galaxies

2.8.2.4: Irregular Galaxies

2.8.2.5: Active Galaxies

2.8.2.6: Quasar-stellar Objects

2.8.2.7: Blazars- Quasar-like Object

2.8.3: Galactic Modern Theory

2.8.3.1: Galaxies, Distance, and Age

2.8.3.2: The Ages of Galaxies and What that Reveals

2.8.3.3: Galactic Groups

2.8.3.4: Dark Matter

2.8.3.5: WIMPs and String Theory

2.8.3.6: Gravitational Lensing

2.8.4: The Milky Way Galaxy

2.8.4.1: What do you think?
2.8.4.2: The Milky Way — Our Home Galaxy
2.8.4.3: Milky Way Galaxy Satellites
2.8.4.4: Milky Way Galaxy Research
2.8.4.5: Galileo Galilei, First to See the Milky Way Galaxy
2.8.4.6: The Milky Way Galaxy's Shape
2.8.4.7: Andromeda Galaxy
2.8.4.8: X-Ray and Gamma-Ray of the Milky Way Galaxy
2.8.4.9: Earth's Position within the Milky Way Galaxy
2.8.4.10: A Spinning Spiral

This page titled Chapter 2.8: Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13: Galaxies by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.1: THE PRACTICE OF STUDYING GALAXIES

The Andromeda Galaxy (also designated Messier 31, M31, or NGC 224) is a massive barred spiral galaxy located approximately 2.5 million light-years from Earth.



Andromeda Galaxy.

Image courtesy Wikipedia. Author: Pablo Carlos Budassi. License.

2.8.1.1: What do you think?

2.8.1.2: Galaxies

2.8.1.3: The History of Studying Galaxies

2.8.1.4: Hubble's Major Contributions

2.8.1.5: Hubble's Law

This page titled 2.8.1: The Practice of Studying Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.8.1.1: What do you think?

Looking at these images, can you understand how astronomers could have misunderstood the nature of galaxies by what they saw in the night sky?



Public Domain | Image courtesy of NASA.



Public Domain | Image courtesy of NASA.



Public Domain | Image courtesy of NASA.

This page titled 2.8.1.1: What do you think? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.8.1.2: Galaxies

A **galaxy** is a massive, gravitationally bound system of stars, stellar remnants, interstellar medium, and dark matter. The term galaxy comes from the Greek galaxias, literally translated as milk or milky, referring to our Milky Way galaxy. Galaxies can range in size from the dwarf galaxies — as small as ten million stars, to giant galaxies with one hundred trillion stars. It is currently estimated that there are 170 billion galaxies in the Universe. It is believed that nearly all galaxies also contain stellar systems with planets orbiting many of their stars, like our Milky Way galaxy. However, that has not been confirmed, but is a solid theory based on what we have discovered in the Milky Way, and comparing the characteristics of the Milky Way to other galaxies.

A number of galaxies are believed to have **supermassive black holes** at their centers. The Milky Way's supermassive black hole is called **Sagittarius A** * or **SGR A** * . These supermassive black holes have been observed in a number of other galaxies; astronomers are trying to determine if this is a consistent characteristic.

These supermassive black holes are hypothesized to be the principal driver of the **active galactic nuclei** or **AGN**. An AGN is a compact region located at a galaxy's center, which has a significantly higher luminosity from one to several ranges of the electromagnetic spectrum.

This page titled 2.8.1.2: Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.2: Galaxies by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.1.3: The History of Studying Galaxies

The study and discussions of galaxies goes back millennia. **Abd al-Rahman al-Sufi**, a Muslim astronomer from Persia (modern Iran), was the first to recognize the Large Magellanic Cloud, an irregular galaxy, and in 964 AD made the earliest observation of the Andromeda Galaxy, in which al-Sufi described Andromeda as a *small cloud* . Al-Sufi was the first to observe other galaxies than the Milky Way galaxy.

In his text, al-Sufi discusses a number of astronomical observations and topics, including his observations of what he called Al Bakr, the Large Magellanic Cloud.

Thomas Wright hypothesized in 1750 that the Milky Way galaxy was a flattened disk of stars. Wright also thought that some of the 'nebulae' observed were actually objects like the Milky Way. (Nebulae was the term used to describe all of these nebula-like objects, many of which are now known to be galaxies.) Five years later, in 1755, **Immanuel Kant** coined the term *Island Universe* to describe these 'nebulae' that Wright had hypothesized.

During the search for comets from 1771 to 1784, **Charles Messier** completed a catalog listing of a number of bright nebulae. Several of the Messier Objects are now known to be galaxies. When**William Herschel** created his 1786 catalogue of deep sky objects, he used the phrase spiral nebula for a number of the objects, such as the Andromeda nebula. This indicated that Herschel was able to discern the s*piral nature* of these objects. **Lord Rosse** completed a new, large telescope in 1845 and was able to differentiate between elliptical and spiral nebulae.

Deep sky objects are those objects *other than* Solar System and stellar system objects, and individual stars. These include objects such as star clusters, nebulae, and galaxies.



A sketch of the Whirlpool Galaxy, Messier 51 by Lord Rosse in 1845. Public Domain | Image courtesy of NASA.

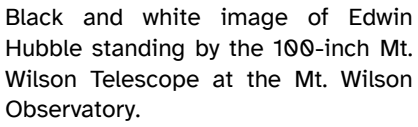


A photograph taken by the Hubble Space Telescope. Public Domain | Image courtesy of NASA.

This page titled 2.8.1.3: The History of Studying Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.3: The History of Studying Galaxies by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.1.4: Hubble's Major Contributions



Black and white image of Edwin Hubble standing by the 100-inch Mt. Wilson Telescope at the Mt. Wilson Observatory.

California Institute of Technology, Mt. Wilson ObservatoryPublic Domain | Image courtesy of NASA.

Edwin Hubble was hired to work at the new Mt. Wilson observatory in 1919, studying cloudy patches called nebulae.

American astronomer Edwin Hubble made three major contributions to the field of galactic astronomy.

First Contribution

First, using the new 100-inch telescope at Mt. Wilson Observatory in California, he demonstrated that some of these nebulae, like the Andromeda nebula, were actually objects – galaxies – far beyond our Milky Way galaxy.

This contradicted the view at the time; the Milky Way was considered the Universe. Yet Hubble's observations of Cepheid variables in these galaxies and comparing them to Cepheids in the Milky Way led him to his controversial conclusion.

Second Contribution

Second, Hubble was the first to classify galaxies based on what he observed, from 1922 to 1923. He classified these based on shapes: elliptical, spiral, and irregular, called a galaxy's visual morphology. Hubble's classification led to his Hubble Galactic Tuning Fork or Hubble Sequence — how he thought galaxies evolve.

Third Contribution

The third contribution is Hubble's formulation of the redshift distance law in 1929, better known as Hubble's Law. The law states that the more distant a galaxy, the greater the redshift. We can determine a galaxy's receding velocity by its redshift— *most of the time* . (Receding velocity is how fast the galaxy is moving away from us.)

$$\mathbf{V = H_o d}$$

Where:

- **V** is the velocity
- $\mathbf{H_o}$ is Hubble 's constant
- **d** is the distance to the object

One of the ongoing controversies surrounding Hubble's Law is determining Hubble's constant, Ho. One of the goals of the Hubble Space Telescope, named for Edwin Hubble, was to determine H_o, which it was able to do through years of observation and data.

Rearrange the equation to find the distance:

$$\mathbf{d = v / H_o}$$

This page titled 2.8.1.4: Hubble's Major Contributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.4: Hubble's Major Contributions by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.1.5: Hubble’s Law

The appearance of Hubble’s Law was not without controversy, either. Two years prior to Hubble’s formulation, **Georges Lemaître**, a Belgian Catholic priest, published a paper in the journal*Annales de la Société Scientifique de Bruxelles* , which detailed the relationship between distance and velocity, supporting an expanding Universe Model. Lemaître earned his doctorate in Astronomy at Massachusetts Institute of Technology, MIT, under the direction of Dr. Harlow Shapley.

It appeared, however, that astronomers wanted to acknowledge Hubble’s work instead of Lemaître. This might have been due to the fact that Lemaître’s doctorate advisor, Harlow Shapley, strongly opposed Hubble’s earlier extragalactic objects research and findings.

This page titled 2.8.1.5: Hubble’s Law is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.5: Hubble’s Law by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.2: GALACTIC CLASSIFICATION

Hubble's "tuning fork" classification scheme:

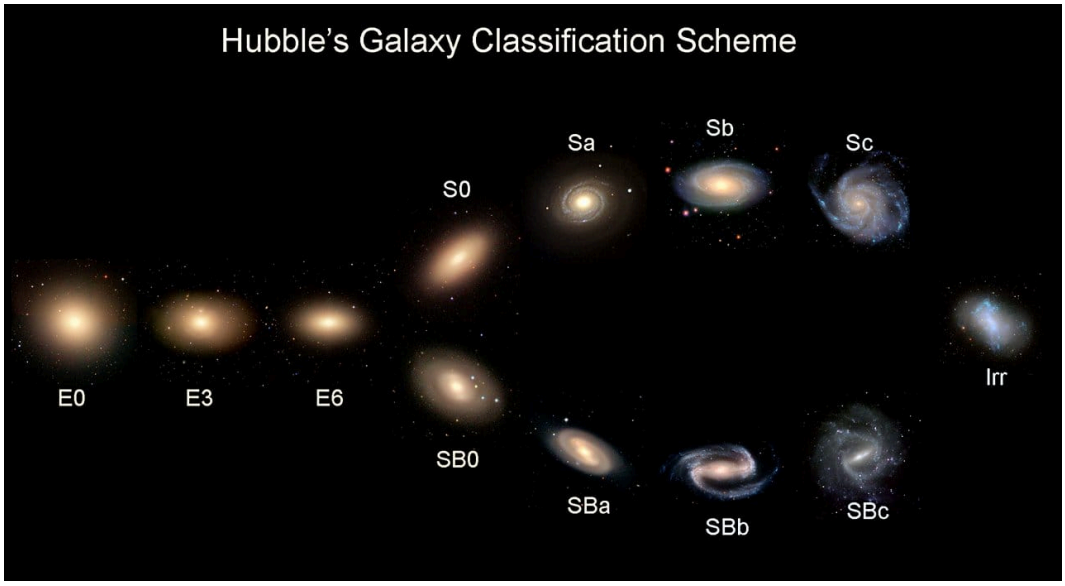


Image courtesy of ESO Supernova. Author: Zooniverse.

2.8.2.1: Classification of Galaxies

2.8.2.2: Elliptical Galaxies

2.8.2.3: Spiral Galaxies

2.8.2.4: Irregular Galaxies

2.8.2.5: Active Galaxies

2.8.2.6: Quasar-stellar Objects

2.8.2.7: Blazars- Quasar-like Object

This page titled 2.8.2: Galactic Classification is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.8.2.1: Classification of Galaxies

Edwin Hubble was the first to classify galaxies based on the shapes, which he observed. These observations included Elliptical galaxies, Spiral galaxies, and Irregular galaxies. This lead to his **Hubble Galactic Tuning Fork** (Hubble Sequence)—how Hubble thought galaxies evolved.

Astronomers now recognize four major groups of galaxies:

- Elliptical Galaxies
- Spiral Galaxies
- Irregular Galaxies
- Active Galaxies

This page titled 2.8.2.1: Classification of Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.6: Classification of Galaxies by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.2.2: Elliptical Galaxies

Elliptical galaxies are those galaxies classified according to their shape and size. The shapes go from round – soccer-ball like – to extreme elongated – football-like – and are classified E0 (soccer ball shapes) to E5 (football shapes). Elliptical galaxies have very little dust and gas, and appear to be the oldest class of galaxies. They are not the dominant type of galaxy found in the Universe; some estimates place Elliptical populations are at around 10% to 15% of all galaxies.



M87Public Domain | Image courtesy of NASA.



NGC 1316Public Domain | Image courtesy of NASA.



NGC 1132Public Domain | Image courtesy of NASA.

This page titled 2.8.2.2: Elliptical Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.8.2.3: Spiral Galaxies

Spiral galaxies are those galaxies which appear like flattened disks with a central bulge, usually with arms. There are three main types or classes of Spirals based on their shapes.

Spiral Galaxy Types

Spirals

Classes

- **Descriptions**
  - **Classes**
    - **Descriptions**
      - **Classes**
        - **Descriptions**
          - No Spiral Arms, but disk-shaped. Appear to be old when compared to other spirals



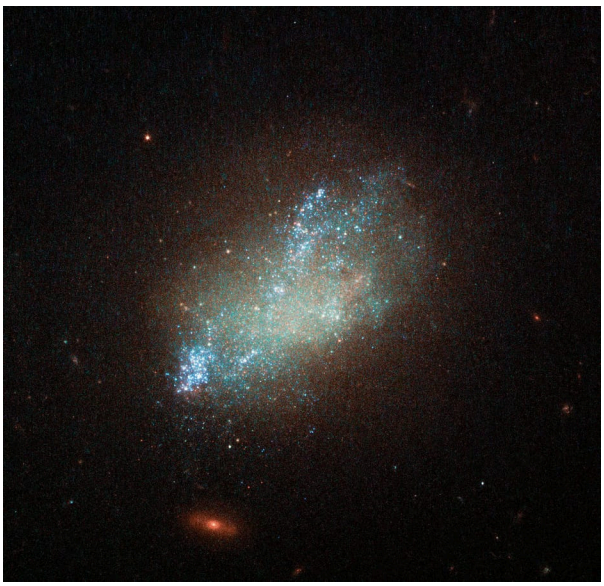
NGC 2787, Lenticular SpiralPublic Domain | Image courtesy of NASA.

This page titled 2.8.2.3: Spiral Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

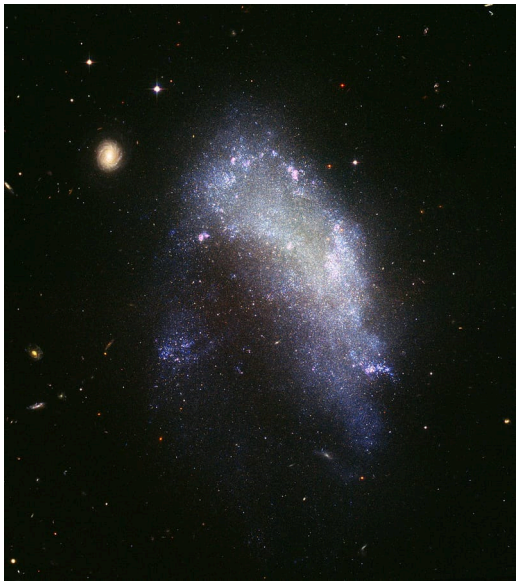
- 13.8: Spiral Galaxies by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.2.4: Irregular Galaxies

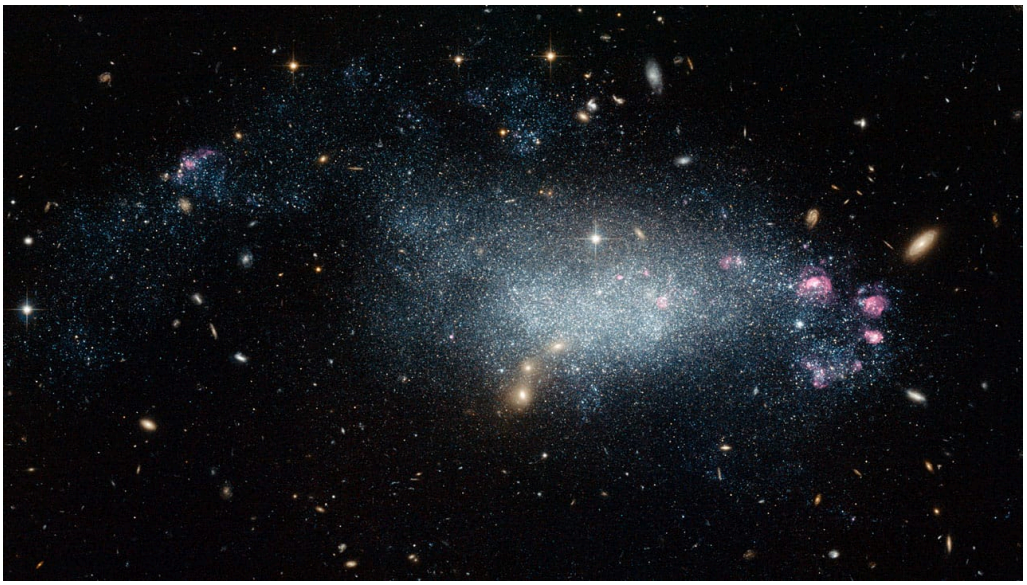
Irregular galaxies are those galaxies, which look neither Elliptical nor Spiral and have no true shape or form. Irregulars make up a smaller percentage of all galaxies, estimated to be about 25%; they were probably more prevalent in the early Universe. Irregular galaxies are dusty, but not as much as spirals, and usually small in size when compared to other galactic types. They are often satellite galaxies of larger spirals or ellipticals.



IC 559CC BY 3. 0 | Image courtesy of ESA/Hubble, NASA, D. Calzetti (UMass) and the LEGUS Team.



NGC 1427APublic Domain | Image courtesy of NASA.

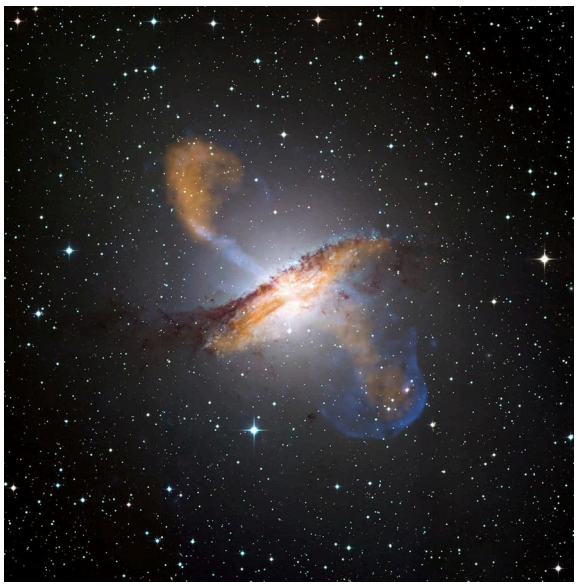


DDO 80, Dwarf GalaxyCC BY 3.0 | Image courtesy of ESA/Hubble.

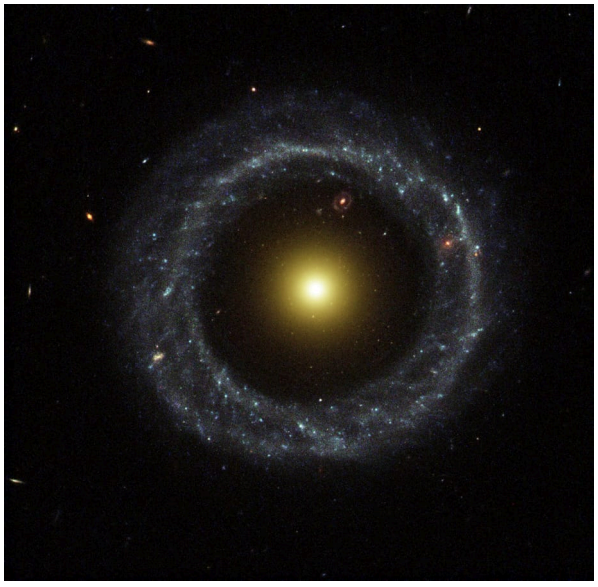
This page titled 2.8.2.4: Irregular Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.8.2.5: Active Galaxies

Active galaxies are those galaxies with extreme energy output; approximately 10% of all known galaxies fall into this group. There are a number of types and classifications of Active galaxies. The**Seyfert galaxies** look like Spirals but produce extreme energy output. **Radio galaxies** emit large amounts of radio electromagnetic radiation. There are other galaxy related objects, such as Starburst galaxies, Quasars and Blazars.



Centaurus A, Radio galaxyCC BY 4. 0 | Image courtesy of ESO/WFI (Optical); MPIfR/ESO/APEX/A. Weiss et al. (Submillimetre); NASA/CXC/CfA/R. Kraft et al. (X-ray).



Hoag's Object, Ring galaxyPublic Domain | Image courtesy of NASA.



M82, Starburst galaxyPublic Domain | Image courtesy of NASA.

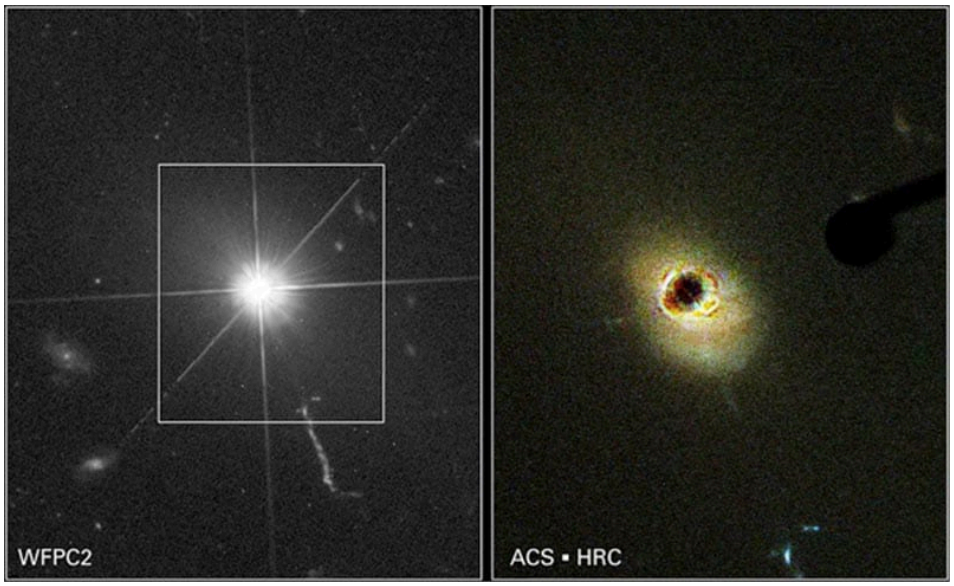
This page titled 2.8.2.5: Active Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.10: Active Galaxies by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.2.6: Quasar-stellar Objects

Quasars, or **Quasi-stellar Objects**, look like stars, yet are the most luminous, powerful, and energetic objects known in the Universe. Quasars, or QSOs, are extreme radio sources, emitting the energy of an entire galaxy or more — 1,000 times that of the Milky Way. The fuel source for QSOs appear to be very energy efficient, super-massive black holes. The left image shows the brilliant Quasar; the spikes of light also demonstrate the star-like appearance of the Quasar. This image was taken with the Hubble's Wide Field Planetary Camera, WFPC. The right image shows the bright Quasar blocked (appears black), providing a view of the Quasar's host galaxy. This image was taken with the Hubble Advanced Camera for Surveys, ACS.

Most QSOs are found at the edge of the observable Universe; representing the Universe in the distant past. They exhibit the highest-observable red shifts and the most distant of objects currently observable. A recently discovered (2010) Quasar was seen at 13.3 billion light-years distant. This is considered to be near the limit of the observable Universe. Quasars were most likely formed through early galactic collisions. These collisions formed the super-massive black holes at the centers of Quasars.



Quasar 3C 273Public Domain | Image courtesy of NASA.

This page titled 2.8.2.6: Quasar-stellar Objects is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.11: Quasar-stellar Objects by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.2.7: Blazars- Quasar-like Object

Blazars appear to be Quasar-like objects; they exhibit very focused jets of energy, which are traveling at 99%c – 99% of the speed of light. If one is in a Blazar's energy jet path, the Blazar will appear very bright. Blazars are classified as very compact Quasars. Often faint traces of the host galaxy is visible. Both Blazars and Quasars are classified as **Active Galactic Nuclei**, AGN. And some objects thought to be variable stars or extreme radio sources have been reclassified as either Quasars or Blazars.

This page titled 2.8.2.7: Blazars- Quasar-like Object is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.8.3: GALACTIC MODERN THEORY

Galaxies consist of stars, planets, and vast clouds of gas and dust, all bound together by gravity. The largest contain trillions of stars and can be more than a million light-years across. The smallest can contain a few thousand stars and span just a few hundred light-years. Most large galaxies have supermassive black holes at their centers, some with billions of times the Sun's mass.



An artist's rendition of the Milky Way Galaxy.

Image courtesy of NASA. Author: NASA/JPL-Caltech.

2.8.3.1: Galaxies, Distance, and Age

2.8.3.2: The Ages of Galaxies and What that Reveals

2.8.3.3: Galactic Groups

2.8.3.4: Dark Matter

2.8.3.5: WIMPs and String Theory

2.8.3.6: Gravitational Lensing

This page titled 2.8.3: Galactic Modern Theory is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

2.8.3.1: Galaxies, Distance, and Age

There are a number of characteristics which galactic shapes, brightness, outputs, and specifics can tell us about galaxies. **Galactic distances** are more difficult to determine for the more-distant galaxies. Methods like radar and parallax are of no use. So astronomers have developed a **standard galactic brightness** ; there is a relationship between a spiral's luminosity and how fast it rotates. Called the **Tully-Fisher Relationship**, the faster a galaxy spins, the brighter the galaxy. So once the spin rate is determined, the brightness and then distance can be determined.

Another method is to examine a **galaxy's white dwarf supernovae** luminosities. This provides another "standard" brightness with which we can compare.

Galactic Distance and Position versus Galactic Age

Most of the galaxies are moving away from each other; each with a velocity of V. This infers that the galaxies must have been closer together at one time. Recall that Hubble's Law is a relationship between velocities and distances, and is related to expansion through the Hubble Constant, H_o .

This goes back to rearranging Hubble's Law **$V = H_o$ to $d = v/H_o$** and allows us to infer not only the age of galaxies, but the Universe itself, time. We have a good idea at the value of **H_o** . So let's say we can determine how far galaxies are away at varying distances, **d**. We already know the velocity, **V**, from determining the distance; this is the velocity each galaxy is moving. Since you know the velocity and distance, use a simple formula and you can now calculate the time.

V = d/t t = d/VWhere: **V** is velocity | **d** is distance | **t** is time.

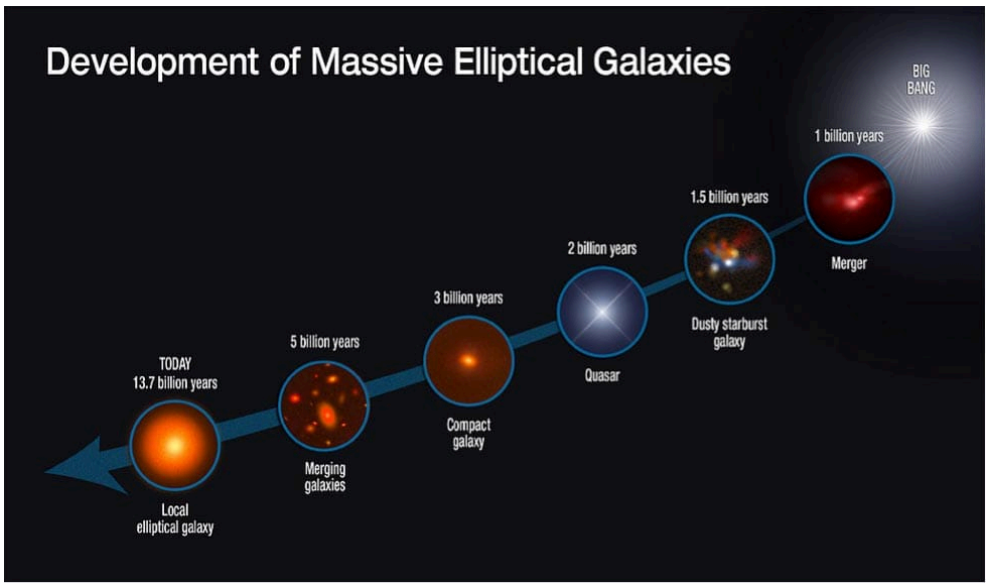
This page titled 2.8.3.1: Galaxies, Distance, and Age is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.13: Galaxies, Distance, and Age by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.3.2: The Ages of Galaxies and What that Reveals

First, how did galaxies form? That concept is not as well understood. One current hypothesis is that just after the beginning of the Universe, there was a lot of hydrogen, H2, and helium, He; the simplest and first elements formed. The hydrogen and helium was probably not a uniform distribution of H2 and He, but in clumps. There was an initial expansion of the H2 and He and eventually gravity slowed the expansion, forming **Protogalactic clouds**. Stars and thus galaxies formed out of these Protogalactic clouds. The leading galaxy-formation hypothesis holds that the Milky Way and other galaxies began small and grew bit by bit for the most part, gravitationally acquiring intergalactic gas and dust and merging with galaxies in their immediate neighborhood.

A 2010 study suggests that several large and seemingly disparate chunks of the galaxy formed at the same time from the collapse of a single blob of gas and dust.



Public Domain | Image courtesy of NASA.

What type of galaxy formed first?

What we currently see as Elliptical Galaxies with reddish stars. Think about stellar color as an indicator of age. And think back to globular clusters as a retirement home for stars – red stars.

This page titled 2.8.3.2: The Ages of Galaxies and What that Reveals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.14: The Ages of Galaxies and What that Reveals by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.3.3: Galactic Groups

Galaxies appear to assimilate in groups throughout the Universe. For example, the Milky Way and Andromeda galaxies are members of **The Local Group**, more than 54 gravitationally bound galaxies, most are dwarf galaxies. Smaller galactic groups are a part of galaxy superclusters. The Local Group is part of the Virgo Supercluster, composed of at least 100 galaxy groups and clusters.

Three-dimensional representations show the locations of galaxies in the Local Group and the Virgo Supercluster.

CC BY-SA 3.0 | Image courtesy of Wikimedia Author: Andrew Z. Colvin.

This page titled 2.8.3.3: Galactic Groups is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.8.3.4: Dark Matter

Dark Matter is matter believed to exist from its observed gravitational effects. We have never seen any light or energy from Dark Matter, yet we can “see” the effects of Dark Matter. Dark matter neither emits nor absorbs light, which is our primary means of observation. And it appears that this unseen Dark Matter is the vast majority of mass in the Universe. How do we know Dark Matter is there in the first place? We can **weight a galaxy** — look at the gravitational effects on objects in a galaxy. Does it match up with what we are seeing? We can also observe the **Mass-to-Light ratio:** the amount of mass versus the amount of light, which we observe.

It is important to understand Dark Matter if we want to ponder the fate of the Universe. Astronomers believe there are two possible scenarios: the Universe will continue to expand or the Universe will collapse, called the **Big Crunch**. The Big Crunch will occur if there is no or not enough Dark Matter.

The Milky Way’s Mass-to Light Ratio

We only see about 10% of the mass of the Milky Way; 90% of the mass is unseen, thus, the term **Dark Matter**. In the 1930s, Fritz Zwicky first proposed Dark Matter in the galaxies. At first it was not readily accepted, but as astronomers realized the need to find the source for the Milky Way’s (and other galaxies) missing matter, the concept of Dark Matter made sense. Galaxy mass physics shows Dark Matter exists. The challenge now is devising a method and instrumentation to physically observe Dark Matter.

This page titled 2.8.3.4: Dark Matter is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.16: Dark Matter by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.3.5: WIMPs and String Theory

WIMP: Weakly Interacting Massive Particles

WIMPs are hypothetical extraordinary Dark Matter. Again, they are dark by their nature (WIMPs do not emit any sort of radiation that can be observed with our current technologies. They interact through weak forces, not gravity and are slow moving enough to collect into galaxies, called **cold dark matter**.

String Theory

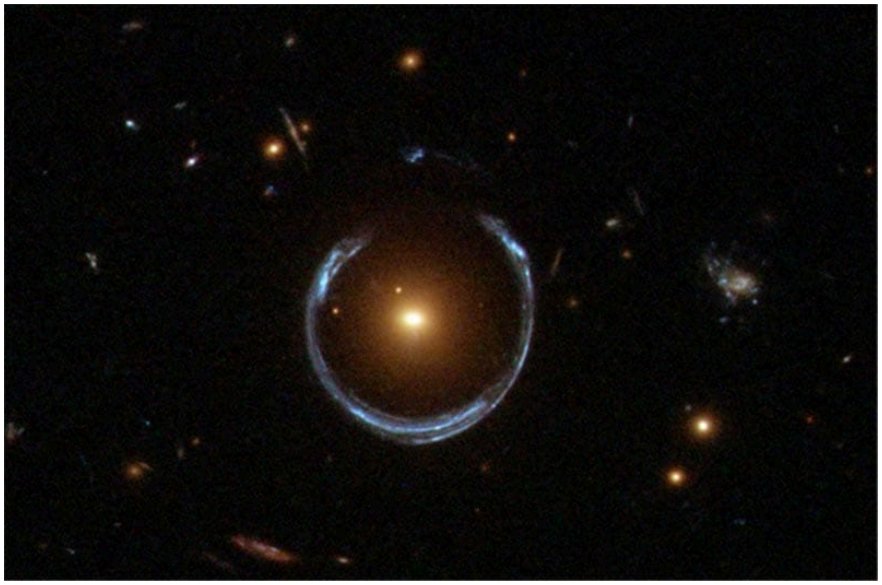
A model of physics, which interprets all particles and forces in terms of specific vibration modes of submicroscopic particle *strings* . String theory originated in the late 1960s. **Strings** – elementary one-dimensional particles – must be stretched under tension through the Universe, like a guitar string. Supporters of the String Theory state that it unifies the natural known forces: electromagnetic, gravitational, strong and weak.

This page titled 2.8.3.5: WIMPs and String Theory is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.17: WIMPs and String Theory by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

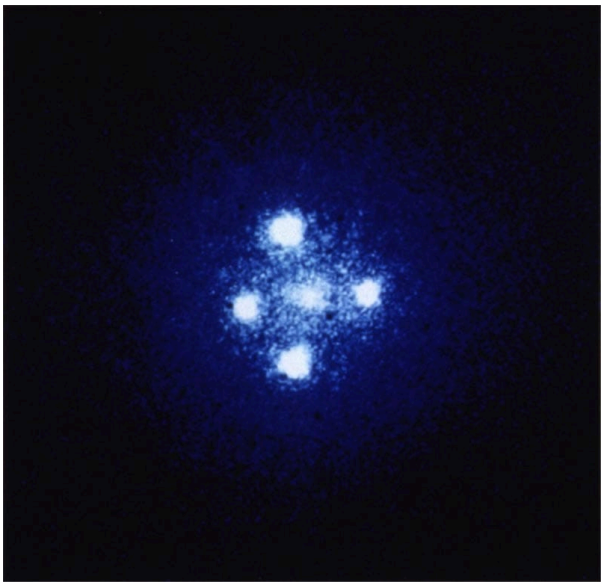
2.8.3.6: Gravitational Lensing

Massive Objects can act as **Gravitational Lenses**, which is the magnification or distortion of an image caused by light bending through a gravitational field. Light emitted from a *source* bends around intermediate mass usually called the *deflector or cluster mass distribution* . Gravitational lensing was predicted by Einstein's General Theory of Relativity.



Public Domain | Image courtesy of NASA.

Gravitational lensing can take place on several scales, from strong lensing to microlensing. Fritz Zwicky postulated in 1937 that this could allow galactic clusters to act as gravitational lenses. In 1979, Zwicky's galactic clusters lensing theory was confirmed through observations of the *Twin QSO, SBS 0957+561* .



Public Domain | Image courtesy of NASA / ESA.

Consider this ...



Public Domain

Often called a “pioneer of the distant stars,” astronomer Edwin Hubble (1889–1953) played a pivotal role in deciphering the vast and complex nature of the universe. His meticulous studies of spiral nebulae proved the existence of galaxies other than our own Milky Way. Had he not died suddenly in 1953, Hubble would have won that year’s Nobel Prize in Physics.

This page titled 2.8.3.6: Gravitational Lensing is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 13.18: Gravitational Lensing by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.4: THE MILKY WAY GALAXY

"See yonder, lo, the Galaxyë
Which men clepeth the Milky Wey,
For hit is shyt. "

Geoffrey Chaucer
The House of Fame, Circa 1380

Learning Objectives

Upon completion of this module, the student will be able to:Upon completion of this module, the student will be able to:

- Identify the characteristics of the Milky Way Galaxy
- Identify the contributors to our understanding of the Milky Way Galaxy

This module presents an overview of our home galaxy, the Milky Way Galaxy, its characteristics, satellite galaxies, and how it compares to other galaxies.

2.8.4.1: What do you think?

2.8.4.2: The Milky Way — Our Home Galaxy

2.8.4.3: Milky Way Galaxy Satellites

2.8.4.4: Milky Way Galaxy Research

2.8.4.5: Galileo Galilei, First to See the Milky Way Galaxy

2.8.4.6: The Milky Way Galaxy's Shape

2.8.4.7: Andromeda Galaxy

2.8.4.8: X-Ray and Gamma-Ray of the Milky Way Galaxy

2.8.4.9: Earth's Position within the Milky Way Galaxy

2.8.4.10: A Spinning Spiral

This page titled 2.8.4: The Milky Way Galaxy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 14: The Milky Way Galaxy by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.4.1: What do you think?

With the clear skies millenniums ago and celestial objects so brilliant like the Milky Way, it is no wonder that many ancient civilizations worshipped the heavens. If it was clear and no light pollution existed today, would people still worship the heavens? Why or why not?

This page titled 2.8.4.1: What do you think? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.8.4.2: The Milky Way — Our Home Galaxy

It is estimated that the Milky Way Galaxy has a population of 100 billion to 400 billion stars. Initially, astronomers did not associate the Milky Way Galaxy with other galaxies they saw through their telescopes. Recall, that they called these other galaxies “nebulae. ”

Characteristics of the Milky Way Galaxy

The Milky Way Galaxy is a Spiral Galaxy and most likely a barred spiral. This characteristic is difficult to confirm due to our location within the Milky Way Galaxy and all of the nebulosity we have to look through. Recall, that the Milky Way Galaxy is part of the Local Group of galaxies, and that group is a part of the Virgo supercluster of galaxies.

This large barred spiral galaxy is believed to be similar in shape to the Milky Way Galaxy. Dust lanes, spiral arms, and a bar at the galaxy’s center are visible.

Some have referred to this image as a ‘picture postcard’ of the Milky Way Galaxy.



NGC 6744, A Barred Spiral GalaxyCC BY 3. 0 | Image courtesy of ESO.

This page titled 2.8.4.2: The Milky Way — Our Home Galaxy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 14.2: The Milky Way — Our Home Galaxy by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.4.3: Milky Way Galaxy Satellites

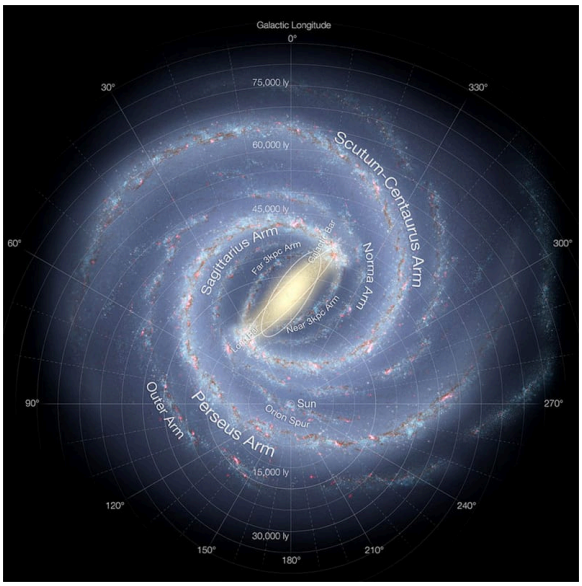
The Milky Way Galaxy has two major satellite galaxies: the Large Magellanic Cloud (LMC) and the Small Magellanic Cloud (SMC). These are visible from the Southern Hemisphere. And there are also a number of dwarf galaxy satellites – at least ten – which orbit the Milky Way Galaxy. Some astronomers hypothesize that there could be hundreds of Milky Way dwarf galaxy satellites.

Milky Way Galaxy Satellites

Our galaxy is 100,000 light years in diameter, that is, across the disk. And it is 1,000 light years wide – "thick. " Think of spiral galaxies like a Frisbee in shape — thin, flattened discs.

The Sun is 28,000 light years from the Galactic Center. From our position within the Milky Way Galaxy, it is challenging to study the Milky Way Galaxy due to the Interstellar Medium. Remember that spiral galaxies are dusty; the Milky Way Galaxy is "dusty" and we cannot see through the Interstellar Medium. We are able to look through the Milky Way Galaxy using radio astronomy and infrared observing techniques.

This detailed annotated artist's impression of the Milky Way Galaxy shows the galaxy's structure, including the location of the spiral arms, the Sun, and other galactic components, such as the central bulge.



Public Domain | Image courtesy of NASA.

This page titled 2.8.4.3: Milky Way Galaxy Satellites is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 14.3: Milky Way Galaxy Satellites by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.4.4: Milky Way Galaxy Research

The Milky Way Galaxy: A brief astronomical history

Aristotle (384-322 BC) wrote in *Meteorologica* that **Anaxagoras** (500-428 BC) and Democritus (460-370 BC) thought the Milky Way might consist of distant stars. Aristotle thought that the Milky Way was due to the ignition [that] *takes place in the upper part of [Earth's] atmosphere, in the region of the world which is continuous with heavenly motions* . Other contemporaries of Aristotle argued against his idea about the Milky Way, including the argument that the galaxy's parallax would be measurable if Aristotle's idea was correct.

Recall, Abū al-Rayhān al-Bīrūnī (973-1048), a Persian astronomer and one of the most-respected Muslim scientists of his time, proposed that the Milky Way was a collection of countless fragments of the nature of nebulous stars.

This 180 ^{o} photograph is bounded on the horizon towards the image's bottom by the trees surrounding Yosemite Valley. The dark lanes within the broad band of stars are dust lanes within the Milky Way Galaxy.



The Milky Way as seen from Earth; Yosemite Valley, CA. Image courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

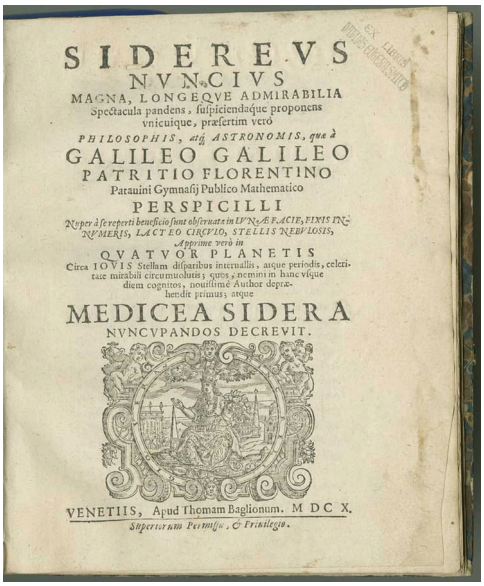
This page titled 2.8.4.4: Milky Way Galaxy Research is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 14.4: Milky Way Galaxy Research by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.4.5: Galileo Galilei, First to See the Milky Way Galaxy

Galileo was the first to see the Milky Way Galaxy in 1610 as individual stars through the telescope. Instead of seeing a cloud, he saw countless stars through his simple telescope. In this book, the first published work of observations made through a telescope, Galileo first noted the nature of the Milky Way Galaxy as seen through his telescope. His comments about the Milky Way stars included that the *"congeries of innumerable stars grouped together in clusters too small and distant to be resolved into individual stars by the naked eye. "*

Here, you see a vintage picture of title page of Sidereus nuncius (Starry Messenger) 1610, by Galileo Galilei (1564-1642).


SIDEREVS
NVNCIVS
MAGNA, LONGEQVE ADMIRABILIA
Spectacula pandens, suspiciendaque proponens
vnicuique, præsertim verò
PHILOSOPHIS, atq; ASTRONOMIS, quæ à
GALILEO GALILEO
PATRITIO FLORENTINO
Patauini Gymnasij Publico Mathematico
PERSPICILLI
Nuper à se reperti beneficio sunt obseruata in LVNÆ FACIE, FIXIS INNVMERIS, LACTEO CIRCVLO, STELLIS NEBVLOSIS,
Apprime verò in
QVATVOR PLANETIS
Circa IOVIS Stellam disparibus interuallis, atque periodis, celeritate mirabili circumuolutis; quos, nemini in hanc vsque diem cognitos, nouissimè Author deprehendit primus; atque
MEDICEA SIDERA
NVNCVPANDOS DECREVIT.
VENETIIS, Apud Thomam Baglionum. M DC X.
Superiorum Permissu, & Priuilegio.

Public Domain | Image courtesy of NASA.

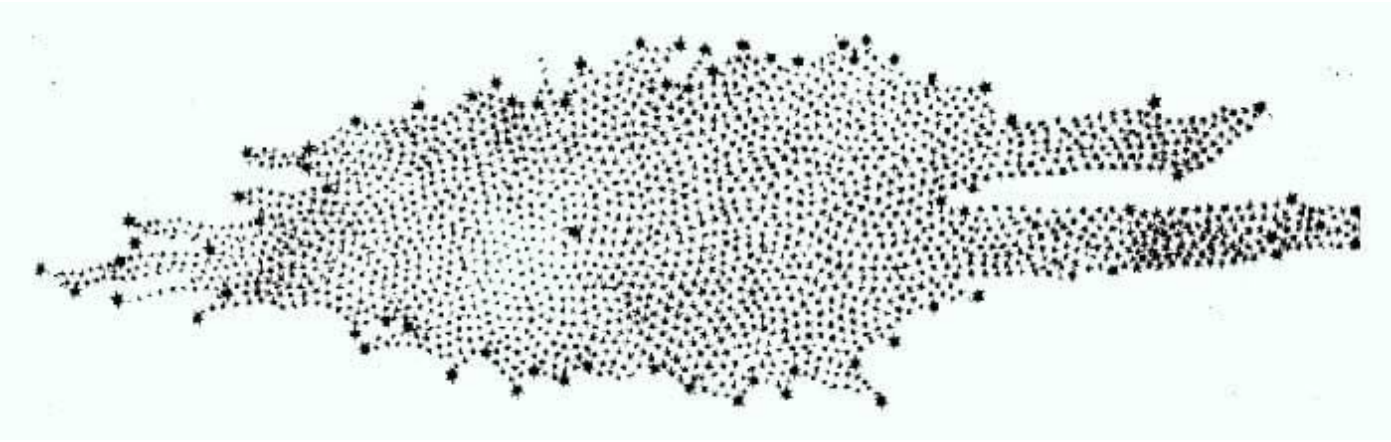
This page titled 2.8.4.5: Galileo Galilei, First to See the Milky Way Galaxy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 14.5: Galileo Galilei, First to See the Milky Way Galaxy by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.4.6: The Milky Way Galaxy's Shape

In his 1755 work *Allgemeine Naturgeschichte und Theorie des Himmels* , **Immanuel Kant** (1724-1804) correctly hypothesized that the Milky Way Galaxy might be a rotating body of an innumerous number of stars that were held together by gravitational forces, much like our Solar System yet on a more-grand scale. Kant based his work on that of Thomas Wright.

Here, you see a prepared sketch by William Herschel of the Milky Way Galaxy's shape based on his 1785 star counts. Herschel believed the Solar System was at or near the center of his sketch.



Public Domain | Image courtesy of NASA.

This page titled 2.8.4.6: The Milky Way Galaxy's Shape is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 14.6: The Milky Way Galaxy's Shape by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.4.7: Andromeda Galaxy

In 1917, **Heber Curtis** (1872-1942) was observing a nova in the Andromeda “nebula. ” Researching previous novae, including those in the Milky Way Galaxy, Curtis was convinced that Andromeda was indeed a galaxy much like the Milky Way Galaxy, and that the dust lanes seen in Andromeda were similar to those in the Milky Way Galaxy. **Edwin Hubble** (1889-1953) settled that question a couple of years later when his research and observations determined “nebula” like Andromeda were indeed actually galaxies.

One of the questions at the time was the location of our Solar System within the Milky Way Galaxy. Harlow Shapley (1885-1972) used globular clusters to show where we “live” in our galaxy. Shapley was also able to estimate the size of the Milky Way Galaxy. Yet not all of Shapley’s observations or hypotheses were correct; perhaps the biggest error was his continued argument that spiral ‘nebulae’ were actually galaxies outside the Milky Way Galaxy. Shapley actually referred to Hubble’s work and findings as “junk science. ”

Also visible in this image are two companion galaxies of the Andromeda Galaxy. Note the dark lanes within the broad band of stars; these are the dust lanes within the Milky Way Galaxy.

Most modern Milky Way Galaxy research has covered such questions as the type and size of the galaxy, the existence of a supermassive black hole at the Milky Way Galaxy’s center, and satellite galaxies associated with the Milky Way Galaxy.

Recall, **Sagittarius A * , SGR A *** , the Milky Way Galaxy’s supermassive black hole, was discovered in 1974. Its characteristics were speculative for a number of years. Continued observations of stars near Sagittarius A* indicate that the object is indeed a supermassive galactic black hole.

The **Virgo Stellar Stream** was discovered in 2006. This is a collection of stars which rises close to perpendicular to the plane of the spiral arms of the galaxy; most likely this structure is a dwarf galaxy.



The Andromeda Galaxy, *M31*Image courtesy of Mike Reynolds, Ph. D. of Florida State College at Jacksonville.

This page titled 2.8.4.7: Andromeda Galaxy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 14.7: Andromeda Galaxy by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.4.8: X-Ray and Gamma-Ray of the Milky Way Galaxy

Two **high-energy emission spherical bubbles**, X-ray and gamma-ray in nature, were detected north and south of the Milky Way Galaxy's core in 2010 by the Fermi Gamma-ray Space Telescope. It is estimated that the diameter of each bubble is about 25,000 light years. In 2014, **Ray Villard** of the Space Telescope Science Institute estimated that the Milky Way Galaxy contains at least 100 billion planets. This would result in each of the Milky Way Galaxy population of 100 billion stars to have an average of one planet per star.

Public Domain | Image courtesy of NASA.

This page titled 2.8.4.8: X-Ray and Gamma-Ray of the Milky Way Galaxy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

2.8.4.9: Earth's Position within the Milky Way Galaxy

The Solar System is not in the middle of Interstellar Medium lanes. If this was the case, we could not see out into space because the dust would block our view.

We are also not too close to the Galaxy's center. Some postulate it would always be daytime if we were near the Milky Way Galaxy's center because there are so many stars. And, also there are dangerous levels of radiation due to the Sagittarius A * supermassive black hole at the galactic center.

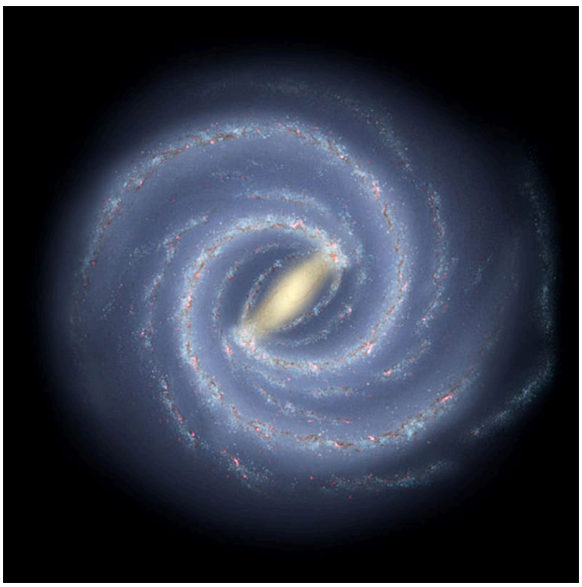
This page titled 2.8.4.9: Earth's Position within the Milky Way Galaxy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 14.9: Earth's Position within the Milky Way Galaxy by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.8.4.10: A Spinning Spiral

The Milky Way Galaxy rotates, undergoing what is known as differential rotation: everything rotates at about the same speed. This type of rotation is different than a rotating tire. Our Sun and Solar System orbits around the Milky Way Galaxy once every 230 million years. Observing the Milky Way Galaxy's rotation helps astronomers determine where the galaxy's mass is located. **Collision Alert!**

Current observations show that the Andromeda Galaxy is approaching the Milky Way Galaxy at 100 to 140 kilometers/second. It appears as if a collision may occur in 3 to 4 billion years. If Andromeda Galaxy and Milky Way Galaxy collide, individual stars within the galaxies would not collide, but instead the two galaxies will merge to form a single elliptical galaxy over the course of about a billion years.



Public Domain | Image courtesy of NASA.

Consider this...

"A Descent into the Maelstrom"

I became possessed with the keenest curiosity about the whirl itself. I positively felt a wish to explore its depths, even at the sacrifice I was going to make; and my principal grief was that I should never be able to tell my old companion on the shore about the mysteries I should see.

—*Edgar Allan Poe* 1841

This page titled 2.8.4.10: A Spinning Spiral is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 14.10: A Spinning Spiral by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

CHAPTER 2.9: COSMOLOGY

"The scientific issues that engage people most are the truly fundamental ones: is the universe infinite? Is life just a sideshow in the cosmos? What happened before the Big Bang? Everyone is flummoxed by such questions, so there is, in a sense, no gulf between experts and the rest. "

Dr. Martin Rees, The Lord Rees of Ludlow
Astrophysicist
(1942-)

Learning Objectives

Upon completion of this module, the student will be able to:Upon completion of this module, the student will be able to:

- Describe the Big Bang Theory, including the theory's strengths and weaknesses
- Describe different cosmological views of the Universe

This module presents an overview of our home galaxy, the Milky Way Galaxy, its characteristics, satellite galaxies, and how it compares to other galaxies.

2.9.1: History of Cosmology

2.9.1.1: What do you think?

2.9.1.2: In the beginning...

2.9.1.3: Physical Cosmology

2.9.1.4: The Big Bang in Eras

2.9.1.5: The Planck Era

2.9.2: "Big Bang" Cosmology

2.9.2.1: Big Bang Evidence

2.9.2.2: Recent Data on the Universe

2.9.2.3: Issues with the Big Bang

2.9.2.4: Beyond the Big Bang

2.9.2.5: What is Next for the Big Bang Theory?

2.9.3: Philosophy of Cosmology

2.9.3.1: Questions Regarding the Eventual Fate of the Universe

2.9.3.2: Religious or Philosophical Cosmology

2.9.3.3: Views on Creation

This page titled Chapter 2.9: Cosmology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15: Cosmology by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.1: HISTORY OF COSMOLOGY

An artist's rendition of the "Big Bang".



Image by Gerd Altmann from Pixabay

2.9.1.1: What do you think?

2.9.1.2: In the beginning...

2.9.1.3: Physical Cosmology

2.9.1.4: The Big Bang in Eras

2.9.1.5: The Planck Era

This page titled 2.9.1: History of Cosmology is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

2.9.1.1: What do you think?

Our Universe is spectacular – yet there is evidence of powerful events as we have seen throughout this course. The Universe appears to have started in the most powerful of events: the Big Bang. Does this surprise you? Why or why not?

This page titled 2.9.1.1: What do you think? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.1: What do you think? by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.1.2: In the beginning...

The study of the origin of the Universe is called **Cosmogony**. This is where science-philosophy-theology-religion meet. Was Creation a supernatural event; that is, did creation of the Universe take place outside of the natural realm? Or was it an event that can be naturally explained? Different philosophies have different views on these questions. Yet the heart of the question is: *can one verify creation through science & technology?* **Cosmology**, from the Greek **kosmos**, is the study of the origin, evolution, and eventual fate of the Universe. Physical cosmology is the scientific study, whereas religious cosmology is based on a body of beliefs and traditions.

Physical cosmology was studied during most of the 20th century, with large telescopes and radio astronomy. From 1989 forward, Earth-based telescopes utilized new technologies, specialized spacecraft were designed and launched, and the Hubble Space Telescope was used to study many cosmological questions. Cosmology deals with the nature of the Universe as a whole ...*including science, philosophy, and religion* .

This page titled 2.9.1.2: In the beginning... is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.2: In the beginning... by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.1.3: Physical Cosmology

The Big Bang is a theory of the Universe's earliest moments and its development over time. The Big Bang presumes that the Universe began from an incredibly small, hot, dense collection of all matter and radiation. A sudden rapid expansion and cooling of this matter and radiation occurred. The simplest of particles were first formed, then atoms, finally stars, stellar systems and galaxies.

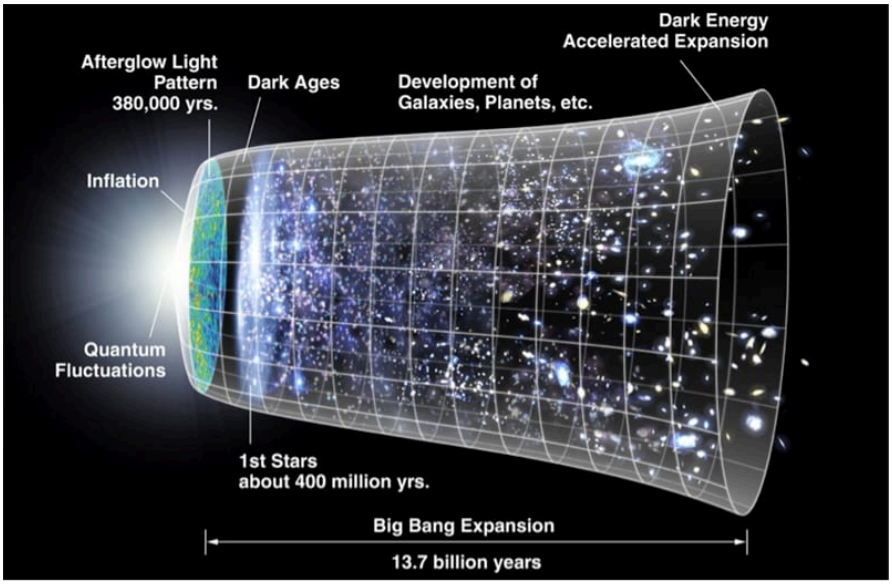
Belgian Catholic priest and scientist **Georges Lemaître** first suggested the Big Bang in 1927. English astronomer **Sir Fred Hoyle** is credited with creating the phrase Big Bang during a 1949 British Broadcasting Company (BBC) radio broadcast. It is popularly believed that Hoyle, who favored the alternative Steady State Theory, intended the phrase to be derogatory. The phrase stuck, even though it is not a good descriptor of what the Big Bang tries to detail. There was actually a contest to rename the Big Bang, yet no other suitable phrase was found.

This page titled 2.9.1.3: Physical Cosmology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.3: Physical Cosmology by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.1.4: The Big Bang in Eras

To look at the development of the Universe through the Big Bang, astronomers have divided the development into steps or **eras**. These eras are our best scientific match to the data we currently have; they seem scientifically sensible. New data is being used to revise and heighten our understanding of these eras.



Public Domain | Image courtesy of NASA.

The generally accepted Big Bang Eras include:

1. Planck
2. Grand Unification Theory (GUT)
3. Electroweak
4. Particle
5. Nucleosynthesis
6. Nuclei
7. Atoms
8. Galaxies

A summary of the Big Bang events occurred in this order:

1. Big Bang beginning – a bright and hot event →
2. **Planck Era** ; initial energy →
3. **Elementary Particles** – photons create particles and antiparticles →
4. **Inflation** of the Universe →
5. **Four Forces** become distinct (strong, weak, electromagnetic, and gravity) →
6. **Protons, Neutrons, Electrons, and Neutrinos** formed →
7. **Hydrogen ions fusion into helium ions** →
8. **Stable, Neutral Atoms** formed →
9. **Protoclouds** formed →
10. **Stars, stellar systems, galaxies** formed →
11. **Life, Humans; today**

This page titled 2.9.1.4: The Big Bang in Eras is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.4: The Big Bang in Eras by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.1.5: The Planck Era

One of the biggest puzzles is what occurred during the **Planck era**. This was an extremely short time span; the instant after the Universe began. This instant is so small we cannot even measure its length with modern technology. At the beginning, the Universe was extremely hot – so hot everything was pure energy; there was no initial matter. Over time, the Universe expanded, and just like the gas expanding as it comes out of a can of compressed air, the expanding Universe cooled. By the end of the first three minutes, atomic nuclei had solidified, and the Universe was a mix of electrons, light, and nuclei. **Inflation** was the sudden and dramatic expansion of the Universe thought to have occurred at the end of the Grand Unification Theory (GUT) era. Most likely, Inflation shaped the way the Universe looks today: structured and smooth. The Universe also reached a point of critical density.

The **four fundamental forces** we observe today became distinct after the era of inflation. It appears that these four forces were, at the beginning of time, united as one force ... the Grand Unification Theory.

These Four Fundamental Forces include:

- **Gravity** is the dominant force on a large scale; the weakest of the four forces.
- **Electromagnetic** is the force that dominates atomic and molecular interactions.
- **Strong** is the force which holds atomic nuclei together.
- **Weak** is the force that mediates nuclear reactions.

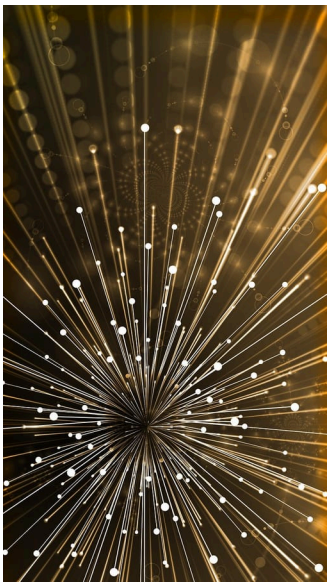
This page titled 2.9.1.5: The Planck Era is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.5: The Planck Era by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.2: "BIG BANG" COSMOLOGY

“Despite its name, the big bang theory is not really a theory of a bang at all. It is really only a theory of the aftermath of a bang.”

— Alan Guth



2.9.2.1: Big Bang Evidence

2.9.2.2: Recent Data on the Universe

2.9.2.3: Issues with the Big Bang

2.9.2.4: Beyond the Big Bang

2.9.2.5: What is Next for the Big Bang Theory?

This page titled 2.9.2: "Big Bang" Cosmology is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

2.9.2.1: Big Bang Evidence

There is respectable and solid evidence to support a Big Bang-type development of the Universe. Three of these lines of evidence are considered major.

Cosmological Redshift

The Universe is expanding – Referred to as the **Cosmological Redshift**.

Redshift analysis confirms the Universe is expanding, and the expansion rate is accelerating. The evidence for an accelerating expansion comes from observations of the brightness of distance supernovae. We observe the redshift of a supernova, which tells us by what the factor the Universe has expanded since the supernova exploded.

The redshift of an object is the amount by which the spectral lines in the source are shifted to the red. That is, the wavelengths become **l-o-n-g-e-r**. To be precise, the redshift Z is given by:

Z = [λobs-λem]/λem Z = Ho d/c (for small distances)

Where:

- Noise" found in all directions throughout the Universe; this "Noise" is radiation remaining from the initial Big Bang.

  Background "Noise" was discovered by accident by Arno Penzias and Robert Wilson, Bell Labs in 1965. They heard a hiss at their radio telescope receiver and after investigation, thought it was due to pigeon droppings on the antenna.

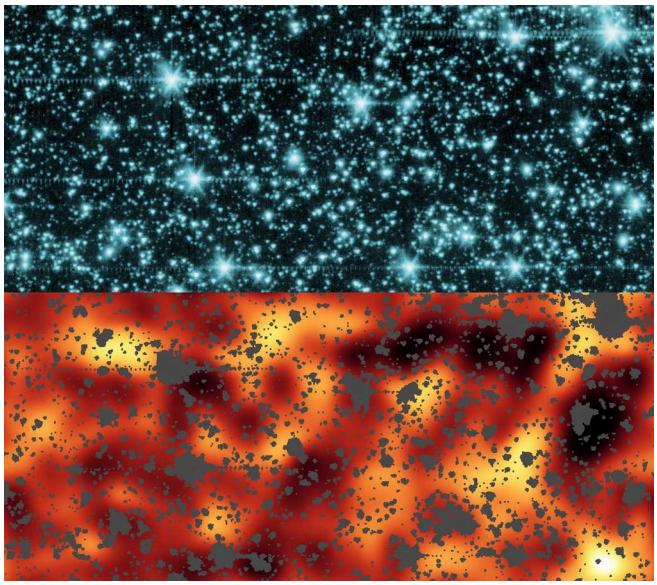
Initial Hydrogen-Helium Fusion

If the H-He early fusion theory is correct, we should find a specific amount of He in the Universe today.

Knowing the current microwave background radiation temperature of 2. 73 K allowed astronomers to mathematically predict the amount of He in the Universe; **Helium should account for approximately 25% of the mass of the Universe.** (This does not include Dark Matter/Dark Energy.) *So…what is the amount of Helium found in the Universe? 25%*

Additional lines of evidence supporting a Big Bang type of development:

- We find **large-scale homogeneity** throughout the Universe.
- There is an **abundance of light elements** : Hydrogen, Helium, Lithium, and Beryllium.
- The **Cosmic Microwave Background Radiation** itself is **uniform**.
- The temperature of the Cosmic Microwave Background Radiation is extremely uniform.
- Large-scale structure of the Universe; the **distribution of objects**.
- The **ages of stars**.



Public Domain | Image courtesy of NASA/GSFC/JPL-Caltech.

Look for that “glow” of those first stars...

- The top image was taken with the Spitzer Space Telescope. This image shows all the objects visible, as very bright lights.
- The bottom image has all objects taken out and is only in infrared. The red, orange, yellow “glow” or bright blobs are likely from the first stars: short-lived yet massive and bright

This page titled 2.9.2.1: Big Bang Evidence is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.6: Big Bang Evidence by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.2.2: Recent Data on the Universe

Wilkinson Microwave Anisotropy Probe or WMAP is a new satellite, which is providing solid data in support of cosmological research. WMAP has helped to significantly refine the age of the Universe to 13. 7 billion years old +1% (137 million years).

WMAP has also provided data in regard to the matter in the Universe:

- Dark Matter: 73%
- Cold Dark Matter: 23%
- Atoms (& Molecules): 4%

This infers that the matter we are used to – and what we are composed of – only makes up 4% of the Universe. This brings us back to the problem: where is the Dark Matter?

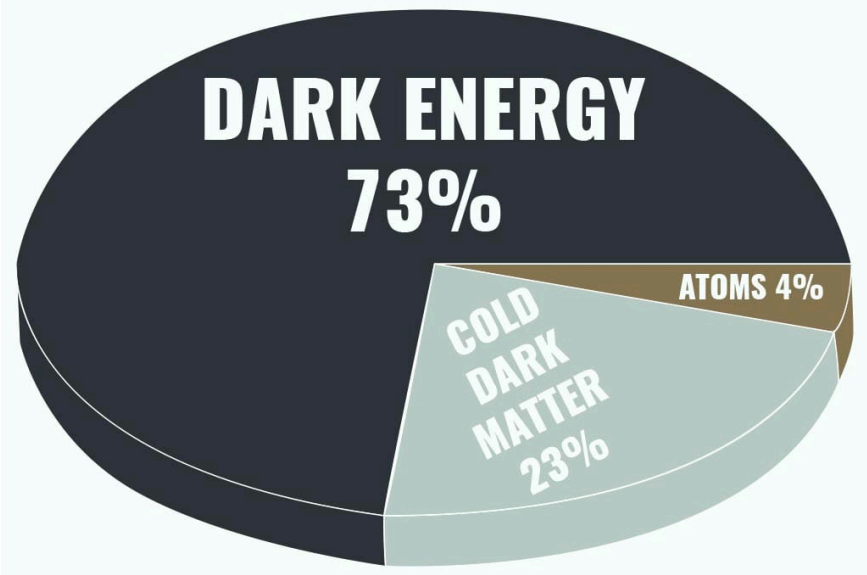


Image courtesy of Florida State College at Jacksonville.

This page titled 2.9.2.2: Recent Data on the Universe is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.7: Recent Data on the Universe by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.2.3: Issues with the Big Bang

There are several issues and questions surrounding the Big Bang Theory.

- Technically, physical cosmologists do not see any of these issues as unsolvable, even the question of where did everything come from since they are mostly concerned with what happened after the Big Bang. Cosmologists look at three technical but important problems:
  - The scope and detail of these problems is beyond this course.

This page titled 2.9.2.3: Issues with the Big Bang is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.8: Issues with the Big Bang by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.2.4: Beyond the Big Bang

Why not other theories besides the Big Bang?

There are some astronomers and scientists who do not believe the Big Bang is correct. Other theories have been proposed, including:

- **Oscillating Universe:** the Universe expands and contracts regularly on an infinite basis. This is sometimes called the *Big Bounce* in deference to the Big Bang.
- **Steady State/Quasi-Steady State:** the Universe is as it was and as it will always be... basically no changes. As matter disappears, more matter is 'created' to take its place.
- Modification Of Newtonian Dynamics **(MOND)**. Newton's laws have changed over the life of the Universe.
- **Tired Light:** Light loses energy.

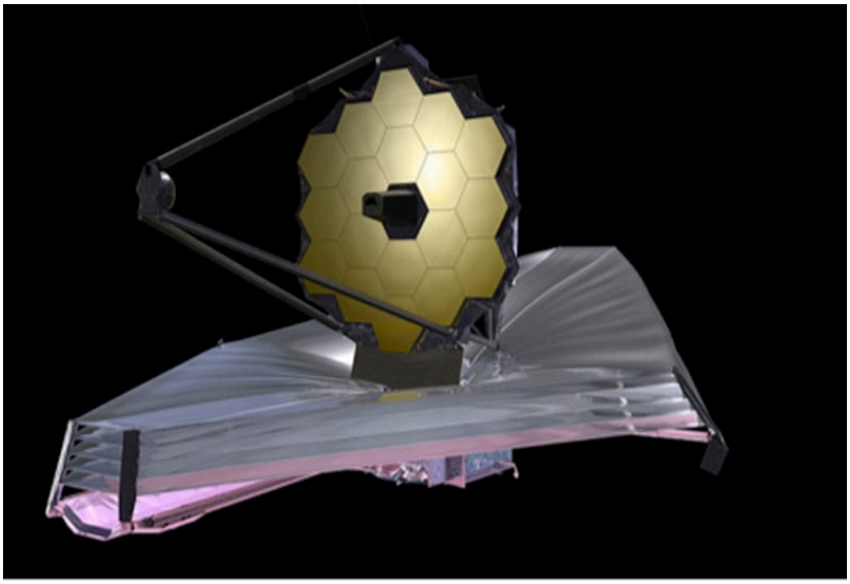
These theories and several others similar to these are supported by a very small minority of astronomers. Mostly you will see arguments against the Big Bang and in support of one of these theories.

This page titled 2.9.2.4: Beyond the Big Bang is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.9: Beyond the Big Bang by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.2.5: What is Next for the Big Bang Theory?

- Continue collecting data with the **Hubble Space Telescope** until it stops functioning.
- Improved **Earth-based telescope technology** will lead to enhanced earth-based data collecting; such as new optics, adaptive optics, and far-advanced computers and software.
- Additional data with other **orbital satellites** : infrared, ultraviolet, gamma-ray, etc., such as the **James Webb Space Telescope (JWST)** (potential launch in 2018).



Public Domain | Image courtesy of NASA.

This page titled 2.9.2.5: What is Next for the Big Bang Theory? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.10: What is Next for the Big Bang Theory? by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.3: PHILOSOPHY OF COSMOLOGY

Unlike most other empirical sciences, cosmology faces unique limitations:

- The Uniqueness of the Universe: Scientists cannot experiment on multiple parallel or benchmark universes to compare initial conditions.
- Observational Horizons: Light speed and the expansion of space mean we can only observe a finite portion of the universe (the particle and visual horizons).
- Underdetermination: Global spacetime geometries and deep past-state origins are often underdetermined by local data, forcing physicists to rely heavily on extrapolation and global uniformity principles.

2.9.3.1: Questions Regarding the Eventual Fate of the Universe

2.9.3.2: Religious or Philosophical Cosmology

2.9.3.3: Views on Creation

This page titled 2.9.3: Philosophy of Cosmology is shared under a license and was authored, remixed, and/or curated by Machele Kindle.

2.9.3.1: Questions Regarding the Eventual Fate of the Universe

What is the Universe expanding into?

There is nothing whatsoever that we have measured, or can measure, that will show us anything about this larger space. Everything that we measure is within the Universe, and we see no edge or boundary or center of expansion. Thus the Universe is not expanding into anything that we can see, and this is not a profitable thing to think about.

Will the Universe keep expanding, come to a stop, or begin to collapse?

This depends on the ratio of the density of the Universe to the critical density necessary to support continued expansion. If the density of the Universe is higher than the critical density, the Universe would recollapse in a **Big Crunch**. But current data suggest that the density is less than or equal to the critical density so the Universe would expand forever.

Even if the Universe continues to expand, what will eventually happen to the Universe?

From our current knowledge and understanding, the Universe will continue to expand. Matter will be converted to energy. This is understood from the Laws of Thermodynamics: matter cannot be created out of nothing; the Universe would undergo the conversion of its matter to energy. Finally, with matter converted into energy, eventually the Universe will become dark. This would happen in about 100 trillion years!

Astronomers believe Dark Matter and Dark Energy exist. If so, why can't we see either?

Probably because our sensors cannot register Dark Matter and/or Dark Energy. Yet every time scientists turn a higher fidelity instrument toward the Universe, they find something new. Consider this as a science history lesson: we didn't see individual cells until the microscope was invented and scientists theorized about the atomic nucleus; now we see the parts of the nucleus.

What of Parallel Universes or Multiverses?

Some current thinking points to the possibility of numerous universes, of which we are in one universe. Think of our Universe as a bubble; now, think of multiple bubbles. Is there any evidence? Yes, what appears to be "bumps" in our bubble. These bumps appear in the Cosmic Microwave Background images. It has been hypothesized that the bumps represent distinct collisions between Universes (December 2010).

This page titled 2.9.3.1: Questions Regarding the Eventual Fate of the Universe is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.11: Questions Regarding the Eventual Fate of the Universe by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.3.2: Religious or Philosophical Cosmology

Some texts and writings use the term “creation” when discussing the Big Bang Theory, as well as alternate theories.

Yet is creation an appropriate term?

Creation:

1. The act of creating or causing to exist. Specifically, the act of bringing the Universe or this world into existence.
2. Works such as art, music...

American Heritage® Dictionary of the English Language, 4th Ed. Copyright © 2000

Creation:

1. Making something. The bringing of something into existence.

Encarta® Dictionary; English Copyright © 2010

This page titled 2.9.3.2: Religious or Philosophical Cosmology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.12: Religious or Philosophical Cosmology by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

2.9.3.3: Views on Creation

It is interesting to explore various religious cosmologies, especially since astronomers cannot answer the question “where did it all come from” or “who did it? ” The laws of thermodynamics provide us with a certainty that something can be created out of nothing.

Buddihism

Nothing is really there... The Buddhist belief of emptiness; lacking reality, substance, meaning. So is there a relationship between the early, empty Universe and Zen Buddhism?

Daoist

The Way gave birth to unity; unity gave birth to duality; duality gave birth to trinity; trinity gave birth to the myriad creatures.

Daodejing, 4th century BC.

Hindu

Brahmā is the god or deva of creation. At the beginning of the process of creation, Brahmā creates the four Kumāras, the four sages. However they refuse to procreate, so Brahmā creates in his mind ten sons to carry on the creation.

Islam

The **Qur’an** contains numerous verses describing creation of the Universe. Muslims believe Allāh created the heavens and Earth in six separate eras, with the Earth being created in two eras, and in two other eras (for a total of four), Allāh furnished the creation of the Earth with mountains, rivers and fruit-gardens.

Judeo-Christian

The **Bible** has several creation passages; the most-well known is in the book of Genesis. This is also the most controversial, where the Genesis passage describes creation and development of the Universe in time periods identified as days. In other biblical text, such as the book of Job, the Universe is described as *“created out of nothing, and then being stretched. ”*

One of the religious cosmology controversies is the **age of the Universe**. Some interpret the Biblical Genesis passage as a literal six days of creation. Those who espouse this idea – called Young Earth proponents – point to other Biblical text in the Psalms and 2 Peter, which states, *“a day is to a thousand years to the Lord. ”* They reason that there were six days of creation, thus 6,000 years plus another 3,000 to 4,000 years of documented civilization, therefore a young earth and Universe of about 10,000 years. An age of the Universe at 10,000 years is extremely young in comparison to a 13. 7 billion year old Universe. This is a ratio of 1 to 1,370,000.

Yet other Biblical and religious scholars contend that the passage has been incorrectly translated from the original old Hebrew text. If one examines the original Old Testament Hebrew text, the term “day” was not used; instead the Hebrew text uses the term **Yom**.

In Hebrew, yom takes on three meanings, depending on the terms preceding or following its use:

- The Genesis use of yom appears to be a “period of time”, which means it could be a long period of time, or a short period of time. This yom period of time is not specified. Old Universe proponents contend that this interpretation of yom is the correct one, in addition to the fact the Genesis days cannot be 1,000 years because the seventh day of creation as recorded in Genesis is still ongoing.

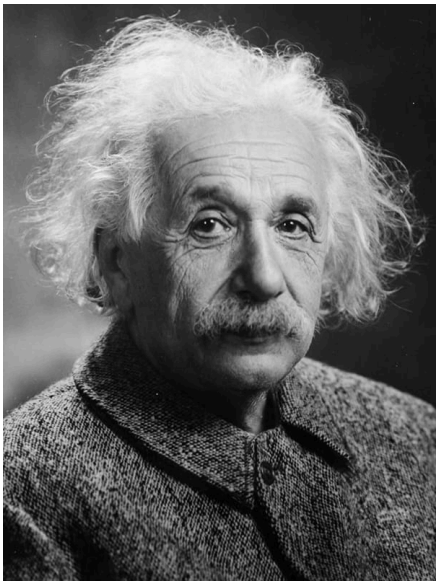
Consider this...

—Michio Kaku (1947-) Physicist

The most beautiful thing we can experience is the mysterious. It is the source of all true art and science.

God does not play dice with the Universe...

—Albert Einstein (1879-1955)



Public Domain

This page titled 2.9.3.3: Views on Creation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

- 15.13: Views on Creation by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/atd-fscj-introastronomy/.

3: CHEMISTRY

Chapter 3.1: Matter energy and their measurements

3.1.1: Matter and energy

3.1.2: What is chemistry?

3.1.3: Measurements

3.1.4: Significant Figures

3.1.5: Unit conversions

3.1.6: Equations and graphs

3.1.7: Density and specific gravity measurements

3.1.8: Heat and its measurements

3.1.9: Heat and changes in physical states of matter

Chapter 3.2: Elements

3.2.1: Dalton's atomic theory

3.2.2: Subatomic particles and a modern view of an atom

3.2.3: Atoms of elements

3.2.4: The periodic table

3.2.5: Electrons in atoms

3.2.6: The periodic trends in properties of the elements

Chapter 3.3: Compounds

3.3.1: Bonding in compounds

3.3.2: Naming binary ionic compounds

3.3.3: Polyatomic ions and their compounds

3.3.4: Naming acids

3.3.5: Naming binary covalent compounds

3.3.6: Lewis structures of molecules

3.3.7: Molecular shapes –Valence shell electron pair repulsion (VSEPR) theory

3.3.8: Polarity of molecules

3.3.9: Intramolecular forces and intermolecular forces

Chapter 3.4: Stoichiometry –the quantification of chemical reactions

3.4.1: Stoichiometry

3.4.2: The mole

3.4.3: Chemical reaction

3.4.4: Patterns of chemical reactions

3.4.5: Oxidation-reduction reactions

3.4.6: Energetics of chemical reactions

3.4.7: Stoichiometric calculations

Chapter 3.5: Solutions

3.5.1: Introduction to solution
3.5.2: Solubility
3.5.3: Electrolytes
3.5.4: Concentration of solutions
3.5.5: Osmosis

Chapter 3.6: Acids and bases

3.6.1: What is an Acid and a Base?
3.6.2: Brønsted–Lowry acids and bases
3.6.3: Strength of acids and bases
3.6.4: Acid-base equilibrium
3.6.5: Dissociation of water
3.6.6: The pH
3.6.7: Acid-base reactions
3.6.8: pH Buffers

Chapter 3.7: Gases

3.7.1: Characteristics of gases
3.7.2: The pressure-volume relationship
3.7.3: The temperature-volume relationship
3.7.4: The pressure-temperature relationship
3.7.5: The combined gas law
3.7.6: The volume-amount relationship
3.7.7: Ideal gas law
3.7.8: Dalton's law of partial pressure

Chapter 3.8: Organic Chemistry

3.8.1: Introduction
3.8.2: Hydrocarbons
3.8.3: Alcohols and Ethers
3.8.4: Aldehydes, Ketones, Carboxylic Acids, and Esters
3.8.5: Amines and Amides
3.8.6: Key Terms
3.8.7: Summary
3.8.8: Exercises

Chapter 3.9: Nuclear chemistry

3.9.1: Introduction to nuclear chemistry
3.9.2: Radioactivity
3.9.3: Half-life of radioisotopes
3.9.4: Radiation measurements
3.9.5: Ionizing radiation exposures
3.9.6: Medical uses of radioisotopes

3.9.7: Making radioisotopes for medical uses
3.9.8: Nuclear fusion and fission

This page titled 3: Chemistry is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 3.1: MATTER ENERGY AND THEIR MEASUREMENTS

3.1.1: Matter and energy

3.1.2: What is chemistry?

3.1.3: Measurements

3.1.4: Significant Figures

3.1.5: Unit conversions

3.1.6: Equations and graphs

3.1.7: Density and specific gravity measurements

3.1.8: Heat and its measurements

3.1.9: Heat and changes in physical states of matter

This page titled Chapter 3.1: Matter energy and their measurements is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.1.1: Matter and energy

Learning Objectives

- Define matter, element, molecule, compound, molecule, mixture, homogeneous and heterogeneous mixtures, the law of constant composition, states of matter, i.e., solid, liquid, gas, and plasma, energy and its types kinetic and potential energy, endothermic and exothermic and exothermic processes, and the law of conservation of energy.
- Be able to write and interpret symbols of elements, molecules, and compounds.

What is the matter?

> Matter is anything that has mass and occupies space.

Matter is a natural material that makes up the universe. The matter comprises tiny particles called atoms held together by forces called bonds. The matter is classified as **pure** if it has a constant and a non-variable composition of the type of atoms. Pure matter is either an element or a compound.

Element

> An element is pure matter that is composed of only one type of atom

For example, carbon, shown in Fig. 1.1.1, is an element. Elements can not convert to simpler matter by physical or chemical methods. There are around a hundred different elements known at this time. For example, hydrogen, oxygen, carbon, nitrogen, sodium, chlorine, iron, cobalt, gold, and silver are a few elements.

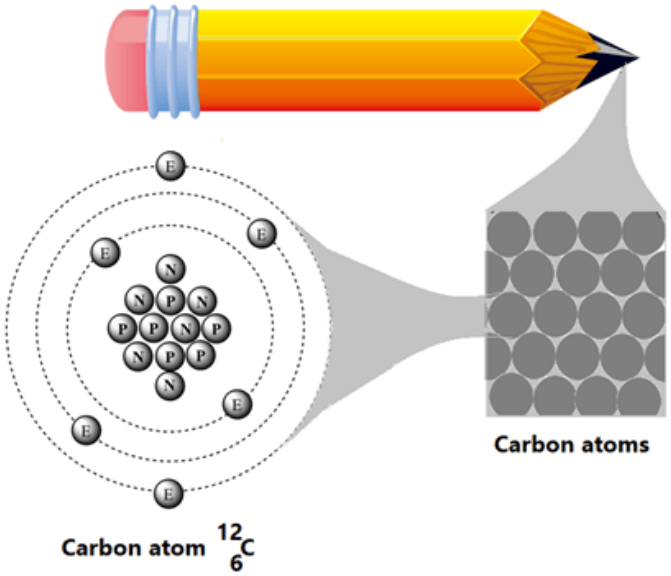


Figure 3.1.1.1: The pencil tip is carbon -an element composed of one type of atom, as shown in the model.

Symbol of an element

Symbols represent elements, the first alphabet of their English or non-English name, written in capital letters. For example, C for carbon, O for oxygen, and H for hydrogen. Usually, another alphabet is also chosen from the element's name and written as a

small letter, e.g., He for helium, Co for cobalt. Some element symbols are derived from non-English names, e.g., Fe for iron is from its Latin name Ferrum, and Au for gold is from its Latin name Aurum.

> Two consecutive capital letters do not represent an element; they may combine two elements.

CO is not a symbol of an element; it is a pure matter, a combination of carbon and oxygen bonded together in a 1:1 atom ratio.

Molecule

A molecule is a group of two or more atoms held together by forces called chemical bonds. The molecule is the smallest particle of matter that can exist freely. A single atom of some elements can exist freely and is also considered a molecule. For example, He, O_2, P_4, and S_8 are examples of elements having molecules composed of one, two, four, and eight atoms of the same element, respectively, as illustrated in Fig. 1.1.2. Although metal elements exist as a vast number of atoms bonded together by a special type of bond called metallic bonds, their symbol is that of a single atom, e.g., Fe for iron and Au for gold.

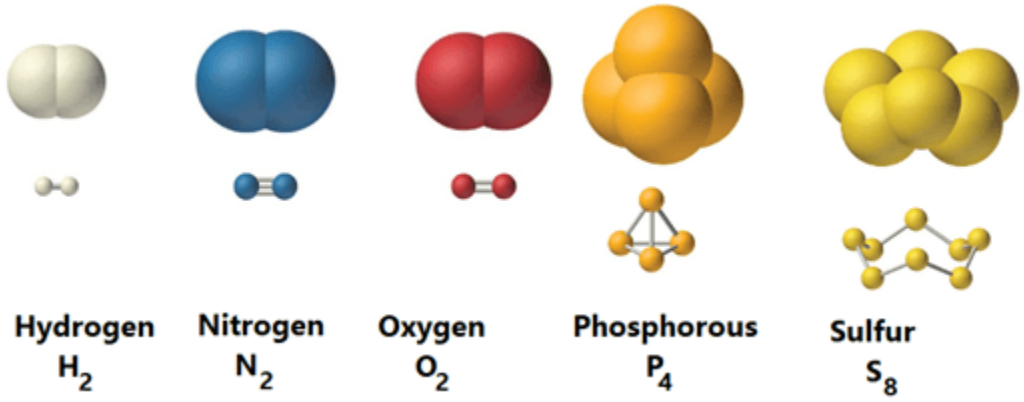


Figure 3.1.1.2: Models of molecules of some elements.

Compound

> A compound is a pure matter composed of atoms of two or more different elements in a constant whole-number ratio held together by chemical bonds.

The symbol of a compound is a combination of its constituent elements with a subscript to the right of the element symbol representing the whole number ratio of the atoms of the element in the compound. For example, H_2O symbolizes a compound called water composed of hydrogen (H) and oxygen (O) atoms in a 2:1 ratio. Similarly, NaCl symbolizes a table salt compound composed of sodium (Na) and chlorine (Cl) atoms in a 1:1 ratio. If the symbol of a compound also represents a molecule of the compound, it is called a **molecular formula**. For example, H_2O is a molecule formula of water. On the other hand, table salt is another class of compound composed of a vast number of atoms of its constituent elements arranged in a specific arrangement in 3D space called a **crystallattice**, as illustrated in Fig. 1.1.3. When the compound symbol does not represent a molecule, it only means the simple whole-number ratio of the constituent elements; it is called the **formula unit**. For example, NaCl is a chemical formula of a compound called table salt.



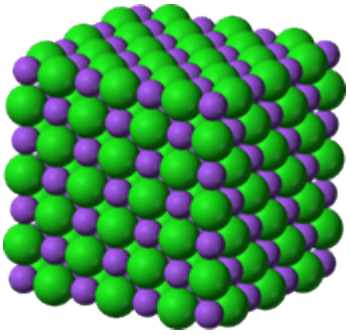


Figure 3.1.1.3: Sodium chloride (NaCl) –a compound, the crystal (left) and the model showing crystal lattice (right). Source: Halite(Salt)USGOV.jpg and Benjah-bmm27 (talk · contribs) / Public domain.

Law of definite proportion or law of constant composition

A chemical compound always contains the number of atoms of its component elements in a constant ratio, or the ratio of masses of the constituent elements is a constant in a given compound.

Pure matter, i.e., elements or compounds, has a fixed proportion of atoms of element/s independent of the source or method of their preparation. It is also called a **chemical** or a **substance**. The atoms in a compound are held together by attractive forces called chemical bonds. Constituent elements in a compound can be separated only by breaking chemical bonds and making new chemical bonds which is a **chemical reaction**.

Mixture

Two or more pure substances mixed such that ratio of the constituents is variable is called a mixture.

For example, table salt can be mixed with water, but the salt (NaCl) to water (H_2O) can be varied to give less salty or more salty water in a mixture. A mixture in which the components are thoroughly mixed, and the composition is constant throughout a given sample, is a **homogeneous mixture** or a **solution**. For example, table salt dissolved in water is an example of a homogeneous mixture or a solution. Other homogeneous mixtures include sugar dissolved in water; the air, a mixture of nitrogen, oxygen, carbon dioxide, and other gases; and metal alloys like brass, a mix of copper and zinc metals, etc.



Figure 3.1.1.4: A heterogeneous mixture of iron and sulfur (left), separated by a magnet (right). Source: Asoult / CC BY (https://creativecommons.org/licenses/by/4.0)

It is a **heterogeneous** mixture if the miitsnents are not thoroughly mixed and the composition varies within different regions of a given sample. For example, sulfur (S) powder mixed with iron (Fe) filling is a heterogeneous mixture shown in Fig. 1.1.4. Other examples of heterogeneous mixtures include smoke which is a mixture of air and carbon particles; smog which is a mixture of liquid water droplets suspended in air; and orange juice which is a mixture of sugar, water, fiber particles, etc. Fig. 1.1.5 illustrates the classification of the matter described above.

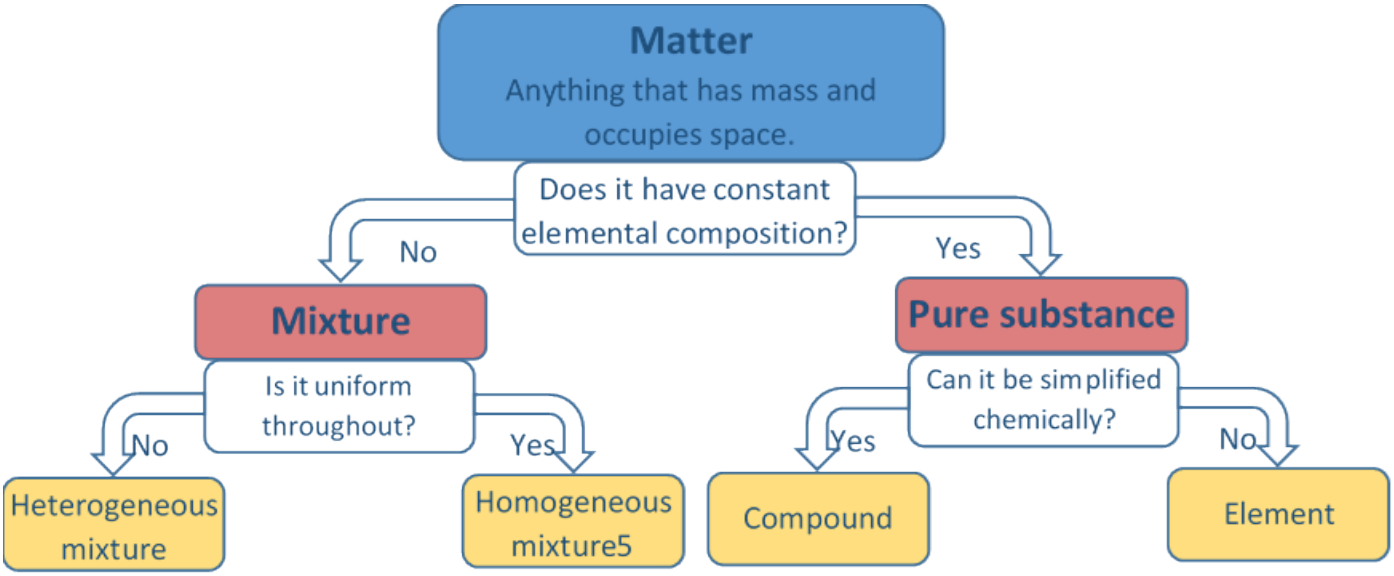


Figure 3.1.1.5: Classification of matter.

States of matter

Matter exists in one of the four physical states or phases, i.e., solid (s), liquid (l), gas (g), or plasma. Fig. 1.1.6 illustrates the four states of matter at the molecular level, and Fig. 1.1.7 shows examples of the four states of matter.

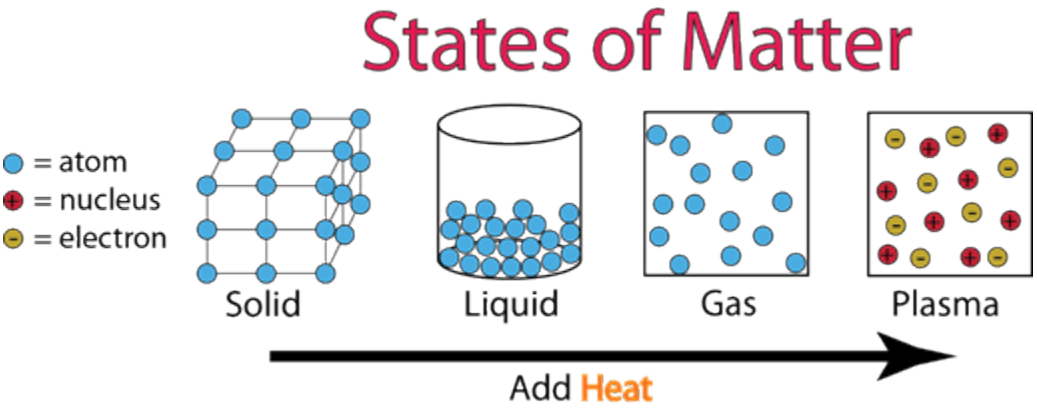


Figure 3.1.1.6: Four states of matter at the molecular level. Source: https://www.hiclipart.com/free-trans...mwpjm/download



Figure 3.1.1.7: Examples of the four states of matter. Clockwise from the top left are solid, liquid, plasma, and gas, represented by an ice sculpture, a drop of water, electrical arcing from a Tesla coil, and the air around clouds, respectively. Source: Spirit469 / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Solid-state

In the solid state**,** the particles, i.e., atoms or molecules, are very close to each other and held strongly by intermolecular forces. The particles can vibrate around their mean positions, but they cannot slide past each other. Expansion and contraction in a solid state are negligible. The solid has a fixed shape and a fixed volume.

Liquid state

In the liquid state, the particles are close enough to experience strong intermolecular interactions that usually do not let the particles cross the liquid boundary. Still, the particles can move around within the liquid. Consequently, the particles in a liquid can flow and acquire the shape of the container but have a fixed volume. Expansion and contraction are negligible in the liquid state.

Gas state

In the gas state, the particles are far apart. The intermolecular interactions are negligible in the gas phase due to the large distances between the particles. The gas molecules move in straight lines in random directions until they collide with other molecules or the walls of the container. The collisions are elastic; that is, the molecules bounce off like elastic balls, and the total kinetic energy of the system is conserved. If exposed to space, the particles keep moving into space. In other words, the particles in a gas can flow, acquire the shape of the container, and expand or contract to fill up the available space. The gases do not have a fixed shape and do not have a fixed volume.

Plasma state

In the plasma state, the particles are far apart like gases, and a portion of the negative charge of the particles, i.e., electrons, are separated from the positive charge potion, i.e., the nucleus. In other words, the atoms in the plasma state are ionized. The plasma state is not common on Earth but is the universe's most common state of matter. For example, the matter in the sun and stars is in the plasma state. Examples of the plasma state on Earth include the matter in the lightning bolts and electrical sparks.

What is energy

> Energy is a quantitative property transferred to an object and recognizable as performing work or as heat or light. In simple words: energy is the ability to do work.

What we commonly encounter other than matter is **energy**. There are two basic types of energies, i.e., the kinetic energy of moving objects and potential energy stored by the position of an object in a force field, e.g., gravitation potential energy under the gravity force and chemical potential energy under the electrical fields in the bonds, as illustrated in Figure $3.1.1.8$.

Figure $3.1.1.8$: Potential energy stored in the fuel is transformed into kinetic energy of the Harpoon missile launched on USS Thorn. (Copyright; Darantares, Public domain, via Wikimedia Commons)

Kinetic energy

The energy of moving objects is kinetic energy. The mathematical form of kinetic energy (KE) is $KE = \frac{1}{2}mv^2$, where m is the mass and v is the velocity of the moving object. Examples of kinetic energy include the energy of all moving things that we see around, like moving vehicles or a moving turbine that generates electricity. Thermal energy is also the kinetic energy of the atoms and molecules in matter.

Potential energy

Potential energy is due to the position of an object in a force field. Examples of force fields responsible for the potential energy include electric, magnetic, gravitational, and elastic forces. Examples of potential energies are electrical energy and gravitational energy. Light is potential energy due to moving electric and magnetic fields. Chemical energy is potential energy stored in chemical bonds in electrostatic potential energy.

Gravitation potential energy is due to position or height relative to the earth. The earth attracts other objects with force $F = mg$, where F is the force, *m* is the object's mass, and g is the acceleration due to gravity. When an object falls, the potential energy changes to several forms of energy, including kinetic energy, work against friction from air, sound, and work done in deformations when it hits the ground.

Chemical potential energy is due to chemical bonds that hold the atoms together in a molecule or compound through electrical forces between negative charge electrons and positive charge nuclei. Bond-forming always releases energy, and bond-breaking absorbs the same energy. Each bond has different bond energy. In chemical reactions, some bonds break, and some bonds form.

A combination of kinetic and potential energy is also possible, e.g., in mechanical waves. A sound wave is a mechanical wave that combines kinetic and potential energy –kinetic because particles move and potential due to the elasticity of the material in which the deformation (sound) is propagating.

Exothermic and endothermic reactions

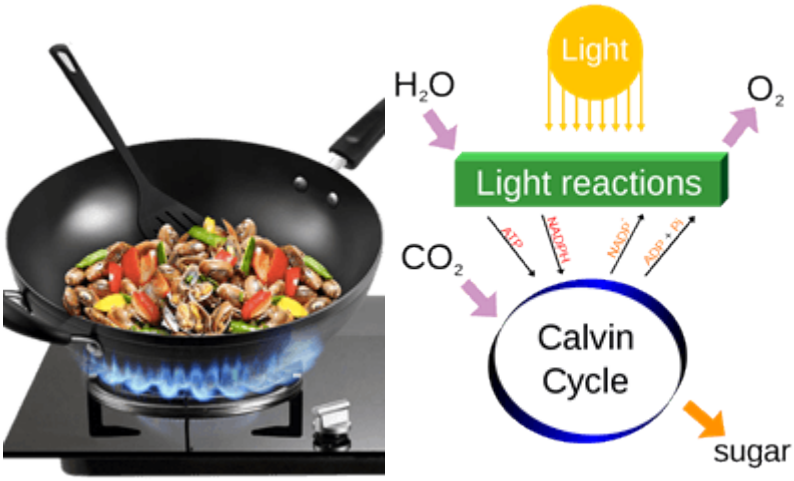


Figure 3.1.1.9: Combustion of methane –an example of an exothermic reaction that releases energy as the heat used for cooking (left), and photosynthesis - an example of an endothermic reaction that absorbs energy (right). Source: https://www.hiclipart.com/free-trans...imhhh/download and https://www.hiclipart.com/free-trans...-clipart-mslhq

The balance between the energy needed to break the bonds and the energy released from forming new bonds determines whether the chemical reaction releases or absorbs energy. Chemical reactions that release energy are called **exothermic** reactions. For example, the combustion of methane gas in kitchen burners releases energy. A reaction that absorbs energy is called an **endothermic** reaction. For example, photosynthesis converts carbon dioxide and water to glucose by absorbing sunlight, as illustrated in Fig. 1.1.9.

Law of conservation of energy

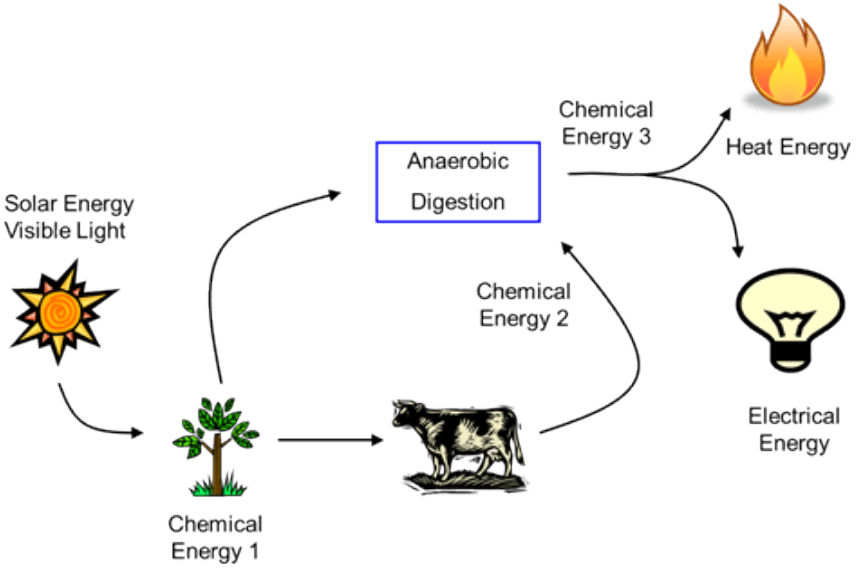


Figure 3.1.1.10: Energy can transform from one form to another, but energy can neither be created nor destroyed. Source: https://www.hiclipart.com/free-trans...muhlp/download

> The energy conservation law states that energy can transform from one form to another, but the total energy of an isolated system remains the same.

Energy can transform from one form to another through work or heat. An oscillating pendulum is an example of kinetic and gravitational potential energy periodically transforming into each other through work. Radiation energy in sunlight transfers to chemical energy during photosynthesis, heat, and work during glucose metabolism, but the total energy remains the same, as

illustrated in Figure 3.1.1.10.

This page titled 3.1.1: Matter and energy is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.1.2: What is chemistry?

Learning Objectives

- Define chemistry, physical properties, physical process, intensive and extensive properties, chemical properties, chemical process, and the law of multiple proportions.
- Be able to write and interpret chemical equations.
- Understand the scientific method and its steps, including observation, hypothesis, experiment, conclusion, law, theory, and how theory evolves knowledge.

Chemistry is the study of matter. More specifically, chemistry studies matter's composition, properties, and transformations.

The properties of matter are of two types, physical properties, and chemical properties, as illustrated in Fig. 1.3.1.

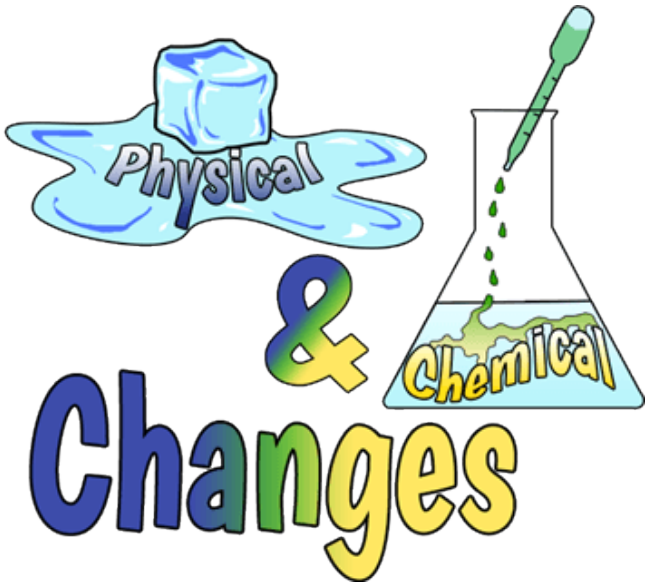


Figure 3.1.2.1: Chemical property relates to elemental composition and its changes. Physical property relates to appearance and its change. Source: https://www.hiclipart.com/free-trans...mnxsm/download

Physical property and physical process

Physical property

Physical property is the property that, when observed, does not change the elemental composition of the matter.

Examples include color, mass, volume, electrical conductivity, and heat conductivities.

Intensive properties

Physical properties that do not depend on the amount of matter are called intensive properties, e.g., color, density, and heat conductivity.

Extensive properties

Physical properties that depend on the amount of the substance, like, mass and volume, are called extensive properties.

Physical process

Any process that changes the matter somehow but does not change the elemental composition is called a physical process.

For example, melting solid to liquid or boiling liquid to a gas state are physical processes.

Mixtures can be separated using physical processes based on the differences in the physical properties of the constituents. Fig. 1.3.2 demonstrates that a magnetic material like iron can be separated from a nonmagnetic material like sulfur using a magnet.

The filtration process can separate a heterogeneous mixture of liquid and solid, like sand in water. Water passes through the filter paper leaving behind the sand particles on the filter.

The distillation process can separate homogeneous mixtures of solids in liquids or liquids in liquids based on the difference in the boiling points of the components. For example, salt dissolved in water separates by distilling off the water, leaving behind the solid salt. Distillation can also separate a mixture of two or more liquids if their boiling points differ; e.g., a distillation of crude oil separates the components based on their boiling points.

Chromatography is another technique often used to separate mixtures. For example, a mixture of inks is adsorbed on a porous paper and separated by ascending through the capillaries in the paper. The component of the ink mixture separate because some components have more ability to stay adsorbed in the solid phase and less ability to solubilize in the liquid phase than the other components. Fig. 1.3.2 Illustrates the physical separation processes described.

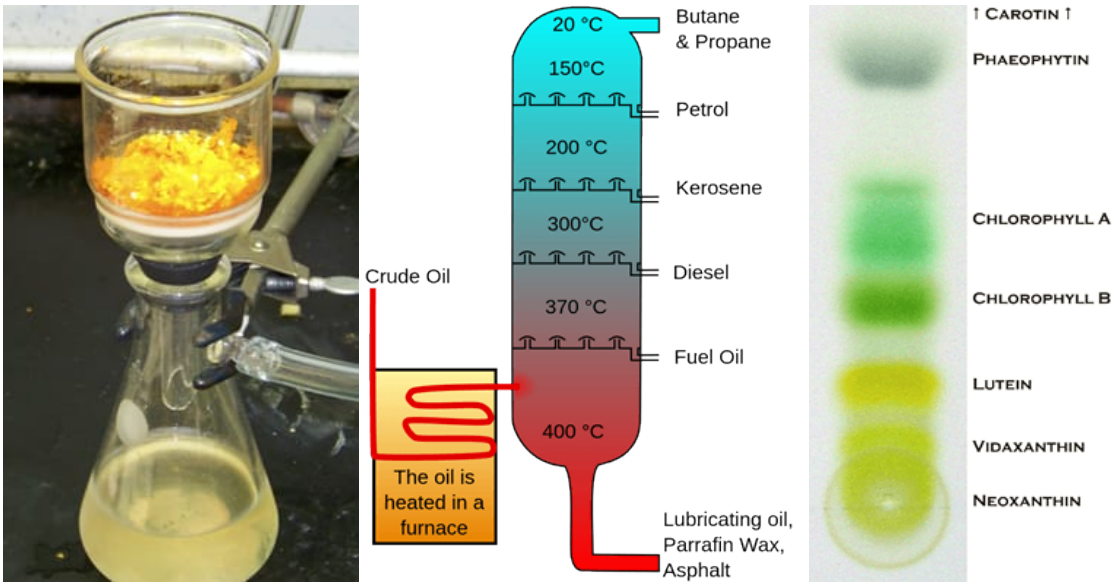


Figure 3.1.2.1: Filtration (left), a distillation of crude oil (middle), and chromatography of chlorophyll and other pigments of a leaf extract (right). Source: a) Smokefoot / Public domain, b) Crude_Oil_Distillation-fr.svg: Image originale:Psarianos, Theresa knott ; image vectorielle:Rogilbertderivative work: Utain () / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/), c) No machine-readable author provided. Flo~commonswiki assumed (based on copyright claims). / CC BY-SA (https://creativecommons.org/licenses/by-sa/2.5)

Chemical property and chemical process

Chemical property

> Chemical properties relate to the change in the elemental composition of the matter.

For example, methane (CH_4) in natural gas is combustible -this is a chemical property. It means methane (\ce{CH4}\)) and oxygen (\ce{O2}\)) change their elemental composition to become carbon dioxide (\ce{CO2}\))and water (\ce{H2O}\)) and release heat after ignition.

Chemical process

> A process that changes the elemental composition is called a chemical processor chemical reaction.

For example, photosynthesis is a chemical process that converts carbon dioxide and water to glucose using energy from sunlight.

Chemical equation

> A chemical equation represents a chemical reaction in the form of symbols of elements and compounds involved.

Substances consumed in a chemical reaction are **reactants;** the substances formed are **products**. The reactants are written on the left side, separated by a plus sign, followed by an arrow, and products are on the right side of the arrow, as illustrated in Fig. 1.3.3. For example, the following chemical equation represents the combustion of methane.

$$CH_4 + 2O_2 \longrightarrow CO_2 + 2H_2O$$

Note that the chemical formula without any preceding number, e.g., O_2 and CO_2, represents one molecule or a unit amount of the chemical. The number preceding the formula is called the **coefficient**, and it represents the number of particles or the number of units involved. For example, the coefficient of 2 in $2H_2O$ in the above chemical equation represents two molecules of water formed or two moles of water formed, where the mole is the unit amount. Note that the chemical composition has changed in the chemical reaction. Before, one substance was carbon and hydrogen atoms in a 1:4 ratio, and the other substance was oxygen atoms. After the reaction, one substance is carbon and oxygen atoms in a 1:2 ratio, and the other is hydrogen and oxygen in a 2:1 proportion.

The physical state of matter is sometimes shown in a chemical equation by the following symbols: (s) for solid, (l) for liquid, (g) for gas, and (aq) for a substance dissolved in water, as illustrated in Fig. 1.3.3.

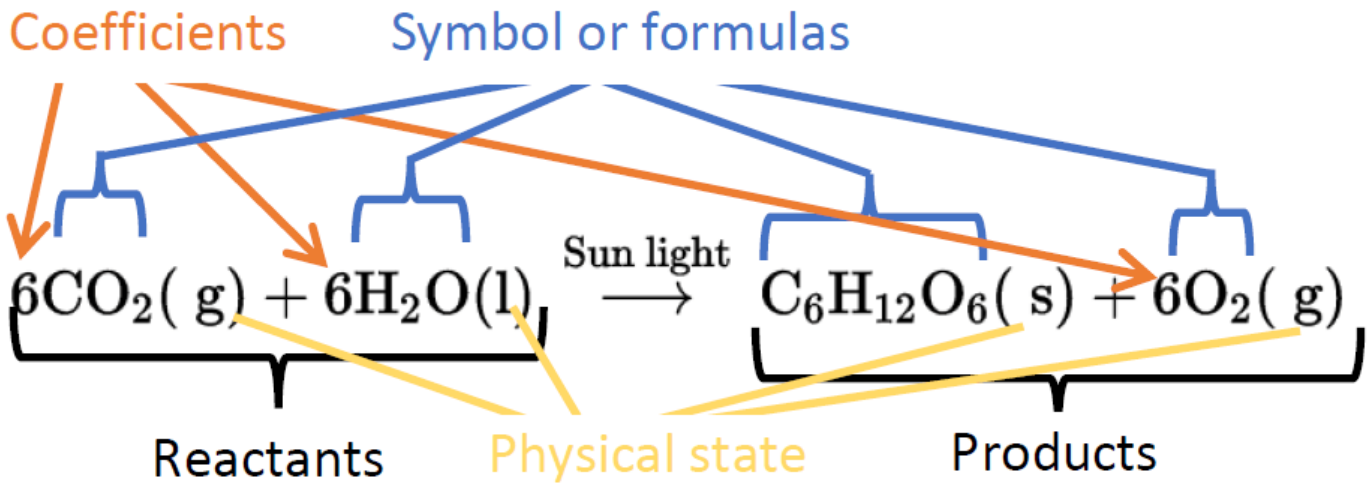


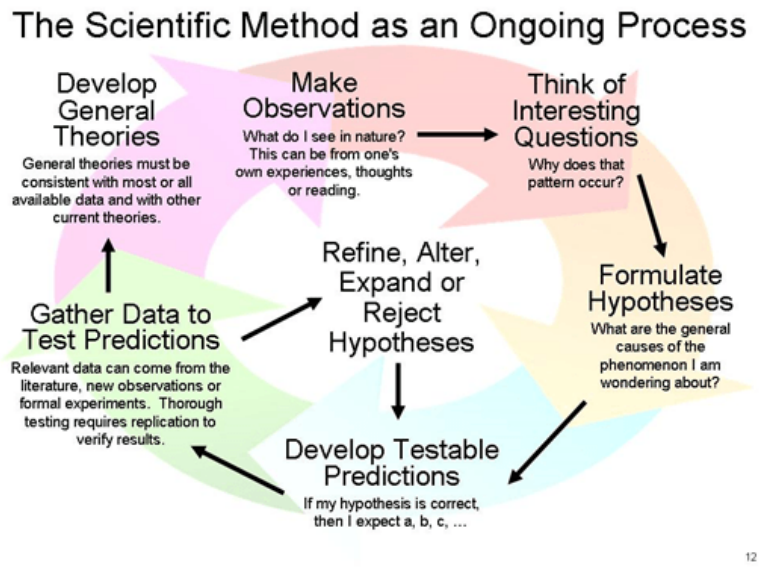
Figure 3.1.2.3: Components of a chemical equation (Copyright: Public domain).

Scientific method -how does science make progress?

Chemistry is one branch of science. Science knowledge is gathered systematically from generation to generation through the scientific method.

> The scientific method starts with making observations, giving a tentative explanation, i.e., hypotheses, testing the hypothesis, i.e., experiment, and deducing a conclusion from the investigation. A truth found through repeated experiments becomes a law, and a comprehensive explanation of related findings gathered over time becomes a theory.

Figure 3.1.2.4 illustrates these scientific method steps and is described below.



[12] Figure 3.1.2.4: Illustration of the scientific method. Source: Whatiguana / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Observation

Observation is the active acquisition of information from a primary source. For example, you fill the air in a car tire and notice that the pressure reading on the gauge increases. This is an observation.

Hypothesis

A hypothesis is a tentative explanation of the observation or a law based on available scientific knowledge. For example, John visits a friend and starts sneezing. The friend says I have a cat, and you might be allergic to cats. This tentative explanation of John's sneezing is a hypothesis.

Experiment

Experiments test the hypothesis. For example, John visits another friend with a cat to determine whether he is allergic to cats. If he sneezes in this experiment, it supports the hypothesis. If he does not sneeze, the experiment disproves the hypothesis.

Conclusion

A hypothesis proven true becomes a conclusion. The hypothesis is rejected or revised if the experiment results do not support it. For example, scientists worldwide and in different periods attempted to convert other metals to gold and failed every time. It concluded that elements do not transform into the more simple matter by any physical or chemical reaction.

Law

It becomes law if an observation is universally true in repeated experiments. Examples of law are the following.

1. The pressure of any gas is directly proportional to the amount of gas if temperature and volume are kept constant, is Avogadro's law.
2. The proportion of atoms of different elements in a compound is always the same, a **law of constant proportion**.
3. Mass before any chemical reaction is the same as after the chemical reaction, i.e., mass is conserved in any chemical reaction or process, a **law of conservation of mass.**

Theory

A theory is a comprehensive explanation based on scientific principles to explain several laws and conclusions on a related topic. For example, the knowledge gathered over time on the properties of matter led Dalton to put forward Dalton's atomic theory.

Dalton's atomic theory

Postulates of Dalton's atomic theory are:

1. elements are composed of tiny indivisible particles called atoms;
2. atoms of any one element are identical to each other but different from atoms of any other element;
3. Atoms of different elements react with each other in a constant whole-number ratio to produce a compound;
4. atoms in a compound separate and recombine to give new material. Still, the atoms are neither created nor destroyed in the reaction.

These postulates explain the properties of the matter described in the previous sections. For example, elements can not convert to simpler substances by any physical or chemical process because they are composed of one type of atom, and atoms are indivisible according to the first postulate. Compounds can convert to elements by the chemical reaction because the atoms in the compounds can separate and recombine according to the fourth postulate.

What happens after a theory is accepted?

The theory goes through the test of time. If it keeps explaining the results of future experiments, it remains valid. It is either rejected or revised if it is disproved or cannot explain some observations of future investigations. For example, "atoms are not divisible" is no more considered valid. According to current knowledge, atoms can divide into **subatomic particles** like electrons, protons, and neutrons. However, the subatomic particles do not represent the element anymore.

"Atoms of the same element are the same" has been revised because isotopes are atoms of the same element that are different in some respects. The statement "atoms are neither created nor destroyed" is still valid for chemical reactions but does not hold in nuclear reactions where atoms of one element can convert to atoms of other elements. Similarly, matter can be converted into energy and vice versa in nuclear reactions, following Einstein's famous equation: $E = mc^2$, where E is energy, m is mass, and c is the speed of light. It means the law of conservation of mass and the law of conservation of energy are not valid individually in a nuclear reaction. Still, the mass and energy together are conserved in nuclear reactions. These are examples of revisions made in theory over time.

The theory is a basic knowledge that allows the prediction of new laws and leads to new ideas on related concepts. For example, Dalton predicted the law of multiple proportions, also known as Dalton's law, i.e., if atoms of two elements can combine in one whole number ratio to give a specific compound, they may also mix in another whole number ratio to give another compound.

Law of multiple proportions

When two elements form more than one compound, the proportions of the atoms of elements in those chemical compounds can be expressed in small whole-number ratios, or the ratio of the masses of the second element in the two compounds that combines with a fixed mass of the first, is a small whole number ratio.

For example, hydrogen and oxygen atoms can mix in a 2:1 ratio to provide water (H_2O). Still, they can also combine in a 2:2 ratio to give hydrogen peroxide (H_2O_2), a different compound. Similarly, carbon and hydrogen combine in a 1:4 ratio to make methane (CH_4); they can combine in a 2:6 ratio to make another ethane compound (C_2H_6).

This page titled 3.1.2: What is chemistry? is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.1.3: Measurements

Measurements are an essential part of making observations needed to develop science. Several measurements are commonly done in everyday life, as illustrated in Fig. 1.3.1. The measured values have two components: a number and a unit.



Figure 3.1.3.1: Measurements of mass, length, time, and temperature are part of daily life. Source: https://www.hiclipart.com/free-trans...xnlrr/download

Numbers

The numbers are composed of digits. The digits are 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. The digits are written in a row in the number, e.g., 123, which means one hundred and twenty-three. The numbers include a decimal point. If the decimal point is not marked, it is assumed to be present at the right of the number. For example, 123 is 123. by conversion, with the decimal point shown in red font.

Place value of a number

The digits have place values. The place values are relative to the decimal point, i.e., the 1st digit to the left of the decimal point is ones, the 2nd is tens, 3rd is hundreds, and so on. The digit 1st to the right of the decimal point is tenth, 2nd is a hundredth, 3rd is thousandth, and so, as shown in Fig. 1.3.2. For example, in 231.45, the digit 2 is hundreds, i.e., 200, 3 is tens, i.e., thirty, 1 is ones, that is one, 4 is tenth, i.e., four divided by ten, and 5 is hundredth, that is five divided by hundred.

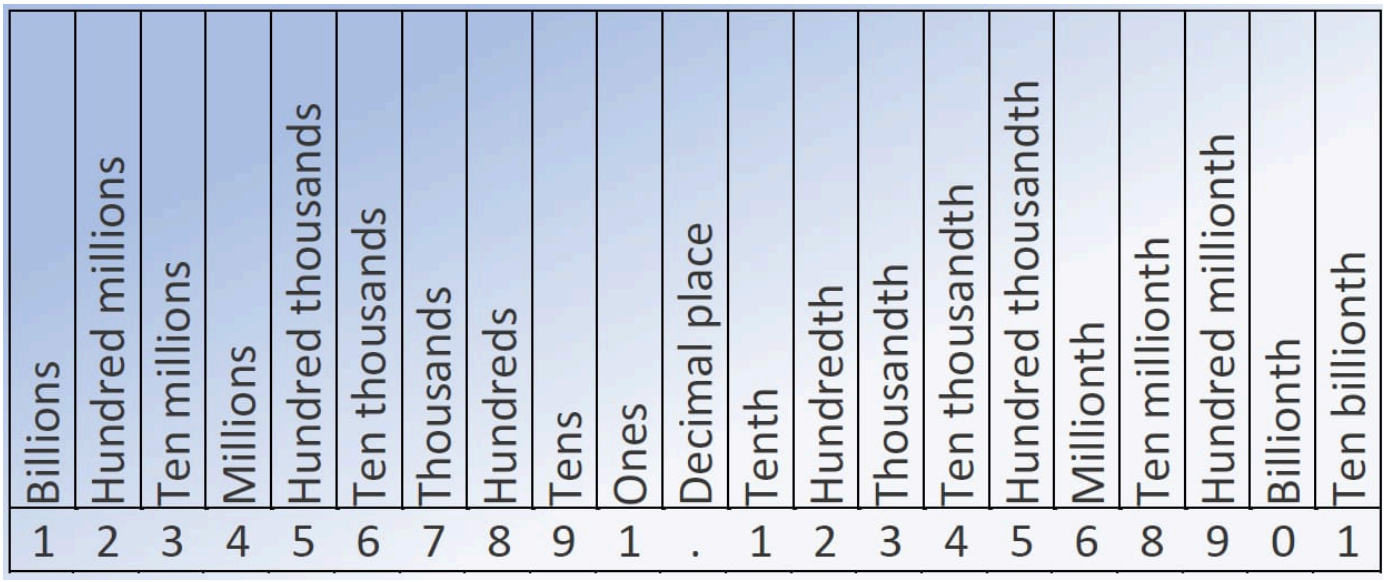


Figure 3.1.3.2: Place values of numbers.

Sign of a number

Numbers have signs, either +ve or -ve, to the left of a number, e.g., -23.4 and +430. By conversion, no sign means +ve. The signs are relative to zero; the -ve sign means the number is less than zero, and the +ve sign means the number is more than zero.

The number and the sign in calculations

The rules for the sign in a calculated answer are the following.

1. When two positive numbers add, the answer has a +ve sign, e.g., 3+2 = 5.
2. When two negative numbers add, the answer has –ve sign, e.g., -4 + (-2) = -6.
3. When two numbers having opposite signs add, subtract the smaller number from the larger number, and the answer has the sign of the larger number. For example, -5 +3 = -2.
4. In subtraction, change the sign of the subtracted number and then follow the addition rules. For example, subtract 3 from 5: 5-(+3) = 5-3 = 2. Note that 3 is subtracted, and its sign changed before operating addition. Another example: subtract -6 from 2: 2-(-6) = 2+6 = 8
5. When two positive numbers multiply, the answer has a +ve sign, e.g., 2x3 = 6.
6. When two negative numbers multiply, the answer has a +ve sign, e.g., (-4) x (-3) = 12.
7. When the two numbers multiplied have opposite signs, the answer has a –ve sign, e.g., (-3) x 2 = -6 and 4 x (-4) = -16.
8. When a number is divided by another number, it follows multiplication rules for the sign. For example, $\frac{-4}{-2} = 2, \quad \frac{4}{2} = 2, \quad \frac{-4}{2} = -2, \text{ and } \frac{9}{-3} = -3.$



Figure 3.1.3.3: Battery % remaining.

Percentage calculations

The percentage (%) is the part out of a hundred, as illustrated in Fig. 1.3.3. The percentage is calculated as part divided by the total and then multiplied by a hundred, i.e.:

$$\text{percentage } \% = \frac{\text{Part amount}}{\text{Total amount}} \times 100$$

Example 3.1.3.1

Calculate the percentage of aspirin if there is 81 mg aspirin in a 325 mg tablet?

Solution

part = 81 mg aspirin, Total = 325 mg tablet

Formula:

$$\text{percentage } = \frac{\text{Part amount}}{\text{Total amount}} \times 100$$

Plug in values in the formula and calculate:

$$\text{Percentage } = \frac{81\text{mg aspirin}}{325\text{mg tablet}} \times 100 = 25\% \text{ aspirin}$$

Example 3.1.3.2

A piece of 18K green color gold jewelry has 7.5 g gold, 2.0 g silver, and 0.5 g copper. Calculate the percentage of gold in the jewelry?

Solution

Part = 7.5 g gold, Total = 7.5 g gold + 2.0 g silver + 0.5 g copper = 10 g jewelry.

Plug in values in the formula and calculate:

$$\text{percentage } = \frac{\text{Part amount}}{\text{Total amount}} \times 100 = \frac{7.5\text{ggold}}{10\text{ g}} \times 100 = 75\%\text{gold}$$

Writing numbers in scientific notation

Sometimes the given number is too large or too small to be easily written, read, and grasped. Scientific notation is one approach to changing a too large or a too-small number into an easily readable and writable number. The following steps convert a given number to scientific notation.

1. Move the decimal point to the right or the left side, one digit at a time, till the largest non-zero digit becomes one's place. For example, move the decimal in 12,700,000 seven times to the right 

to obtain 1.27, and move the decimal in 0.000,006 six-time to the left 

to get 6. The numbers 1.27 and 6 obtained are the coefficients of the scientific notation.
2. The coefficient is multiplied by 10^x, where x is a power of ten. The power of ten equals the number of times the decimal moved. The sign of the power is +ve if the decimal moved to the left and -ve if the decimal moved to the right. For example, 12,700,000 in scientific notation is 1.27 x 10^7, and 0.000,006 is 6 x 10^{-6}.

Units

Physical properties like mass, length, and temperature are measured. The measured value is a combination of a number and a unit, as illustrated in Fig. 1.3.4. For example, a person's height is 1.83 meters, where 1.83 is a number and a meter is a unit.

Units are quantities defined by the standard that peoples agree to use as a reference.

For example, the meter is defined as the distance light travels in a vacuum in $\frac{1}{299,792,458}$ of a second.

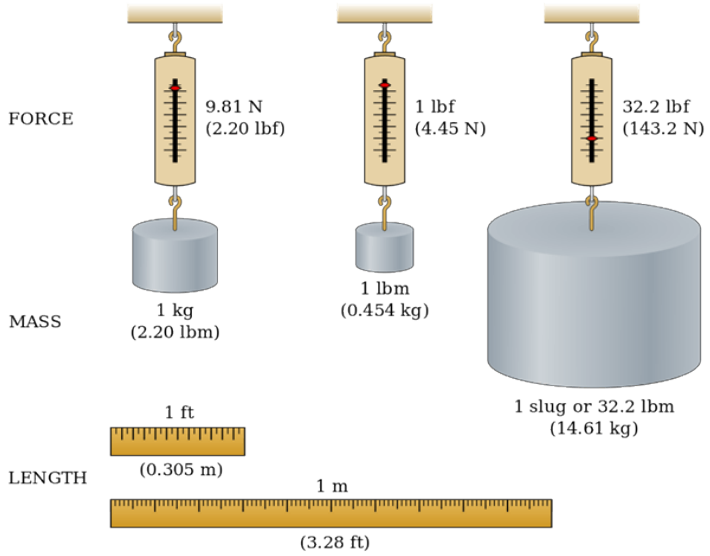


Figure 3.1.3.4: Measurement is a combination of a number and a unit. Source: https://www.hiclipart.com/free-trans...lzufr/download

Systems of units

There are different sets of units used in different systems of units. For example, 1.83 meters and 6.00 feet show the same length value but using a unit of 'meter' from the international system of units (SI) and 'foot' from the English system of units. The international system of units (SI) is universally used in scientific work. There are seven base units in SI, as listed in Table 1.

Table 1: Base units in the International System of Units (SI).

Measurement	Unit	Abbreviation
Time	second	s
Length	meter	m
Mass	kilogram	kg
Temperature	kelvin	K
Amount of substance	mole	mol
Electric current	ampere	A

Note

The following section is based on the 2019 redefinition of the SI base units: https://en.Wikipedia.org/wiki/2019_redefinition_of_the_SI_base_units#Kilogram, accessed on May 2nd, 2020

Time

> Time is the progress of existence and events that occur in succession from the past through the present to the future.

In old times, the time measuring device was the hour sandglass shown in Fig. 1.3.5. The basic unit of time is second (s), a standard unit of time in all the measurements systems. Other units of time are minute (min) which is equal to 60 s, and hour which is equal to 60 min or 3600 s.

Definition: Second

The duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium 133 atoms at a temperature of 0 K.



Figure 3.1.3.5: German; Half-hour sandglass; Horology. Source: Metropolitan Museum of Art / CC0

Length

Length is a measure of distance, i.e., a numerical measurement of how far apart the objects or points are.

Fig. 1.3.6. illustrated the concept of length. The SI unit of length is a meter (m).

Meter

The meter is defined as the distance that light travels in a vacuum in $\frac{1}{299,792,458}$ of a second.



Figure 3.1.3.6: Measure of length. Source: https://www.hiclipart.com/free-trans...nhlxx/download

Mass

The mass of an object is a measure of its inertia.

Inertia is the resistance of any physical object to any change in its velocity. Mass determines the strength of the gravitational attraction of an object to another object -a property commonly used in modern balances for mass measurements, as shown in Fig. 1.3.7. SI unit of mass is the kilogram (kg).

Kilogram

Earlier definition: The mass of one cubic decimeter of water at the melting point of ice.

Current definition: Kilogram (kg) is defined by taking the fixed numerical value of the Planck constant h to be $6.626070 15 \times 10^{-34}$ when expressed in the unit $\mathrm{kg \cdot m^2 \cdot s^{-1}}$.



Figure 3.1.3.7: Laboratory Analytical balance. Source: https://www.hiclipart.com/free-trans...bbigu/download

Temperature

Temperature is a physical property of matter that expresses hotness or coldness, as illustrated in Fig. 1.3.8.

> Temperature is a manifestation of the thermal energy of the matter, which is a source of the flow of energy in the form of heat from a hot object to a cold object when they are in contact with each other.

The SI unit of temperature is Kelvin (K).

Kelvin

Kelvin (K) is defined by taking the fixed numerical value of the Boltzmann constant k to be 1.380649×10^{-23} when expressed in the unit $kg \cdot m^2 \cdot s^{-2} \cdot K^{-1}$.

A 0 K, also called absolute zero, is the temperature of a matter at which no energy can be removed as heat from the matter. The freezing point of water is 273.15 K, and the boiling point of water is 373.15 K.

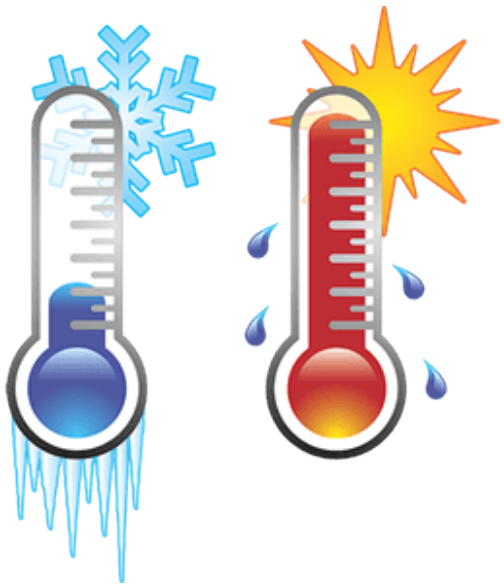


Figure 3.1.3.8: An illustration of temperature. Source: https://www.hiclipart.com/free-trans...txxww/download

Amount of substance

In chemistry, the amount of substance (n) measures the number of specified elementary entities. The elementary particles in chemistry are usually atoms in the case of elements and molecules or formula units in the case of compounds. SI unit of the amount of a substance is a mole (mol).

Mole (mol)

Mole is exactly $6.02214076 \times 10^{23}$ elementary entities.

Fig. 1.3.9 illustrates one mole of aluminum, copper, and carbon. The molar mass is the mass in grams of one mole of that substance, i.e., the mass of $6.02214076 \times 10^{23}$ atoms or molecules. Usually, the number of particles is shown with four significant figures, i.e., 6.022 x 10^{23} atoms or molecules in one mole of the substance.

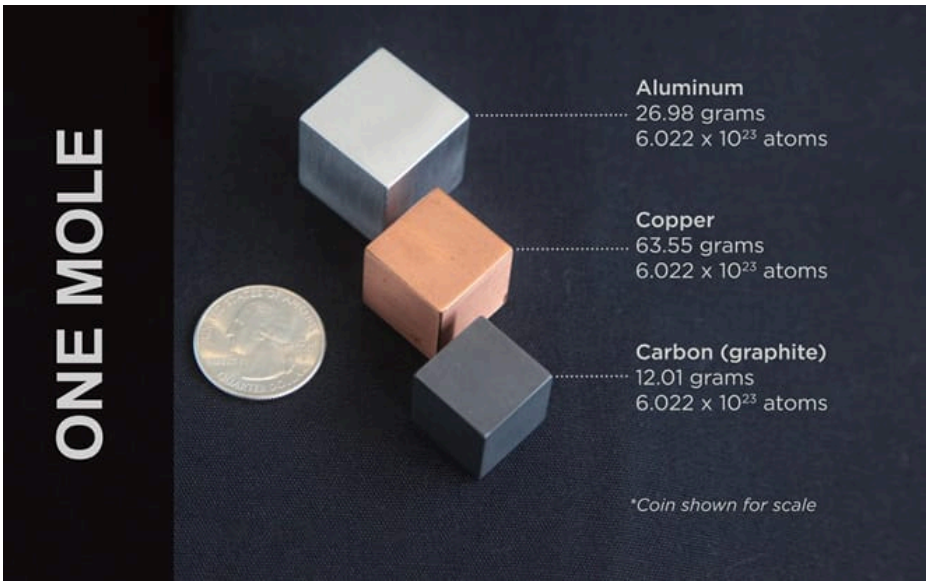


Figure 3.1.3.9: Amount in one mole of aluminum, copper, and carbon shown relative to US Quarter coin.

Source: Public information, photo: R. Press/NIST; graphic design: N. Hanacek/NIST

Electric current

Electric current is the flow rate of electric charge past a point or a region.

It could be the flow of electrons in electric wires or the flow of cations and anions in opposite directions as in electrolytes, as illustrated in Fig. 1.3.10. The SI unit of current is ampere (A).

Definition: Ampere (A)

Ampere (A) is defined by taking charge of an electron (e) to be $1.602176634\times10^{-19}$ coulomb (C), where C is equivalent to ampere-second (A.s).

Figure 3.1.3.10: Illustration of the definition and direction of electric current, i.e, the current (I) is the rate of flow charges (Q) per unit time (t). Source: And1mu/ CC-BY-SA-4.0, https://commons.wikimedia.org/wiki/F...ricCurrent.gif

Luminous intensity

Luminous intensity measures the wavelength-weighted power emitted by a light source in a particular direction per unit solid angle.

The solid angle is measured in steradian (sr), analogous to the radian. The radian is a planar angel that gives the length of the circumference of a circle, and the steradian is a 3D angle, like a cone, that gives an area on the surface of a sphere, as shown in Fig. 1.3.11.

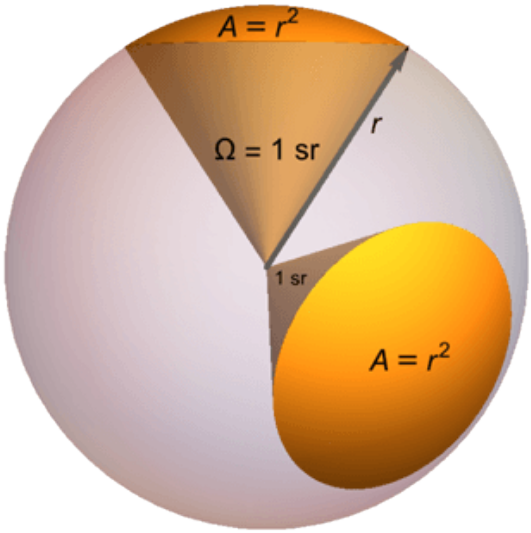


Figure 3.1.3.11: A solid angle is a three-dimensional analog of a circular angle that relates a portion of the volume of a sphere to the surface area it subtends. If that area equals the sphere's radius squared, the solid angle is one steradian. This diagram displays two solid angles of one steradian, viewed from different directions. Source: Andy Anderson / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

SI unit of luminous intensity in a given direction is Candela (cd).

Candela (cd)

Candela (cd) is defined by taking the fixed numerical value of the luminous efficacy of monochromatic radiation of frequency $540{\times}10^{12}$ Hz, K_{cd}, to be 683 $cd \cdot sr \cdot W^{-1}$, or $cd \cdot sr \cdot kg^{-1} \cdot m^{-2} \cdot s^{3}$, where **W is watt** –a SI unit of power described by $kg.m^{2}.s^{-3}$.

Prefixes in SI

In several situations, the measured number with the base unit is either too large or too small. For example, a person's height is comfortable to represent in the meter as the height is usually in a 1 m to 2 m range. However, the diameter of the earth, i.e., 12,700,000 m, and the diameter of red blood cells, i.e., 0.000,006 m, are too large and too small, respectively. The unit needs to be revised so that the number with it is easy to read and write.

> Prefixes are used in SI to increase or decrease the base unit by order of tens.

For example, kilo (k) means a thousand times, i.e., 1 km means 1000 m and 1 kg means 1000 g. Similarly, micro (μ) means one-millionth time, i.e., 1 μm is 10^{-6} m, and 1 μg is 10^{-6} g. Table 2 lists commonly used prefixes in SI.

Table 2: Commonly used prefixes in SI (note: m means "meter" in SI units, but as a prefix, it means "mili")

Prefix	Means	Abbreviation
Gega	10^{9}	G
Mega	10^{6}	M
Kilo	10^{3}	k
Deci	10^{-1}	d
Centi	10^{-2}	c

Prefix	Means	Abbreviation
Mili	10^{-3}	m
Micro	10^{-6}	μ
Nano	10^{-9}	n
Pico	10^{-12}	p

A new unit may be defined and used if there is no appropriate prefix available in SI for some specific type of measurement. For example, the diameter of atoms varies in the range of 1 x 10^{-10} m to 5 x 10^{-10} m, where the prefix pico (p, 10^{-12}) is too small, and the prefix nano (n, 10^{-9}) is large. A new unit called **angstrom** (Å) is defined as 1Å = 10^{-10} m for reporting atomic diameter and inter-atomic distances.

Derived units

> The units in SI other than the seven base units are Derived units obtained by combining the base units.

For example, The SI unit of volume is meter-cube (m^3), equal to the space occupied by a cube of 1m on each edge, as illustrated in Fig. 1.3.12.

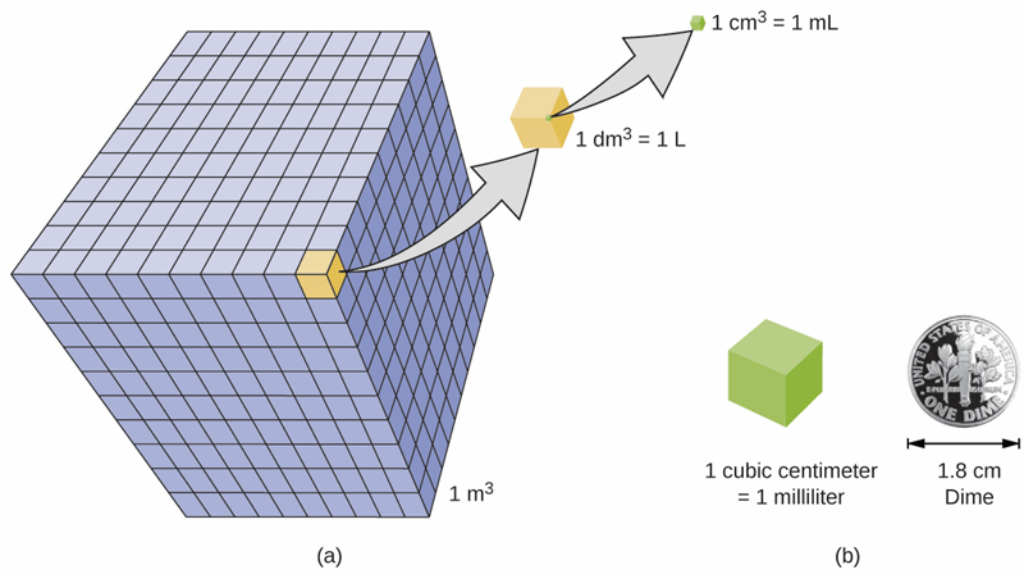


Figure 3.1.3.12: Relative volume in m^3, dm^3 and cm^3. Diagram of a 10x10x10 cubes. The small cubes represent one decimeter-cube (dm^3); the big cube represents one meter-cube (1 m^3). The purpose is to show that even though there is 10 dm in 1 m, there are 1000 dm^3 in 1 m^3. More generally, it shows that the conversion factor between units of volume is the cube of the conversion factor between corresponding length units. Source: Download for free at https://openstax.org/details/books/chemistry

Usually, the volume is reported in decimeter-cube (dm^3), commonly known as liter (L). One liter is a volume occupied by a cube that is one dm on each edge. Another commonly used unit of volume is the centimeter-cube (cm^3), which is also called cc or mL. One mL is the volume occupied by a cube that is one cm on each edge. The dm^3 is a thousandth of m^3, and cm^3 is the thousandth of dm^3, i.e., 1000 dm^3 = 1 m^3, and 1000 cm^3 = 1 dm^3, as illustrated in Fig. 1.3.11.

Relationship of SI units with metric and English system of units

SI was developed from the metric system. Some basic units are different, but both systems have much in common, using the

same prefixes. The English system of units uses a different set of units except for the common unit of time. Table 3 compares the standard measuring units in the three systems of measurement.

Table 3: Common measurement units in three conventional systems of measurements

Quantity	English unit	Metric unit	SI unit	Relationships
Mass	Pound (lb)	Gram (g)	Kilogram (kg)	1 kg = 2.205 lb 1 kg = 1000 g
Length	Foot (ft)	Meter (m)	Meter (m)	1 m = 3.281 ft
Volume	Quart (qt)	Liter (L)	Cubic meter (m^3)	0.946 L = 1 qt 1 m^3 = 1000 L
Energy	calorie (cal)	calorie (cal)	Joule (J)	4.184 J = 1 cal
Temperature	Degree Fahrenheit (°F)	Degree Celsius (°C)	Kelvin (K)	°F = (1.8 × °C) + 32 K = °C + 273.15

This page titled 3.1.3: Measurements is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.1.4: Significant Figures

Significant figures are related to errors associated with the measured numbers. It is important to understand the significant figures because when calculations are made using numbers with errors, the answer cannot have less error than the error in any original number. The answer needs correction for the significant figures.

Precision and accuracy

The measured numbers have two types of errors, systematic and random errors, that determine the accuracy and precision of the measured number, as illustrated in Fig. 1.4.1.

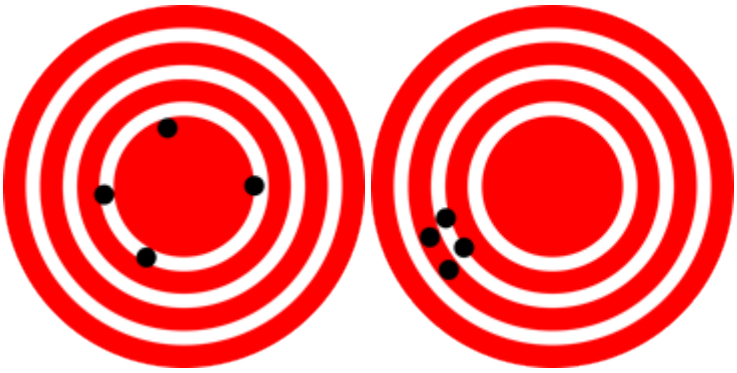


Figure 3.1.4.1: Representation of high accuracy with less precision (left) and low accuracy with high precision (right). Source: DarkEvil / Public domain

Systematic errors

Systematic errors are constant, i.e., they have the same value in every measurement. For example, meter rod is a little short or a little long than a meter, it will introduce a systematic error. Systematic errors usually happen due to inaccurate calibration of the measuring instrument. The systematic errors determine how much the measured value differs from the actual value.

Random errors

Random errors are the statistical variability of the measured number. Random errors vary from one observation to another. Random errors cancel out if many measurements are taken and averaged. Scientific measurements are usually taken at least in triplicate and averaged to minimize random errors. The random errors determine how close the repeat measured numbers are to each other.

Accuracy

> Accuracy or trueness of the measurement is defined as how close the average value is to the actual value.

The closer the average is to the actual value, the more accurate or true it is, as illustrated in Fig. 1.4.1. The trueness depends on systematic errors, i.e., less systematic error, more accurate the average.

Precision

> Precision is defined as how close the individual measurements are to each other.

The closer the individual values are to each other, the more precise the measurement is, irrespective of whether it is accurate or

not, as illustrated in Fig. 1.4.1. Precision depends on random errors, i.e., more substantial random errors mean less precision.

Exact and inexact number

There are two types of numbers, count numbers that are exact and measured numbers that are inexact.

> If the value is a counted number, it is an **exact number**.

That is, there is no error in it. For example, a purchase of one dozen oranges contains exactly 12 nos of oranges; it can not be 11.5 or 12.5.

Inexact numbers and error range

When a value is measured, it comes with an error of measurement.

> A measured number with an error is called an **inexact number**.

For example, when the same one dozen oranges are purchased by mass, the balance may read it 1572.6 g, or 1573 g, or 1570 g, depending on whether the smallest digit that the balance displays is 0.1g, 1 g, or 10 g. Suppose the balance is accurate to 1 g and reports the mass 1573 g; the actual mass may be anywhere in the range of 1572.5g-to-1573.4g. The smallest measured digit, i.e., the number in one's place, in this case, is an **estimated number** associated with an error. By convention, the estimated digit has ±1 errors associated with it. For example, the above-mentioned measured numbers are reported in science as 1572.**6** g ± 0.1 g, 157**3** g ± 1 g, or 15**7**0 g ± 10 g, respectively. The estimated digits are shown in bold fonts in the examples.

The smallest digit in the display of digital instruments is an estimated number. In measurement using instruments that do not have a digital display, the smallest digit marked on the instrument plus one digit less than the minimum marked digit is added to the reported value. The smallest reported digit is an estimated digit. For example, the length of the pencil in Fig 1.4.2 is reported as 17.**7** cm using the ruler on the bottom, where 17 includes the smallest digit marked on the ruler, and the last digit, i.e., 0.7 is an estimated digit. By convention, the error range in this value is shown as 17.**7** ± 0.1. The same length is 17.7**0** cm using the ruler on the top in Fig. 1.4.2l, where 17.7 includes the smallest digit marked on the ruler, and the last reported digit, i.e., 0, is an estimated digit. By convention, the error range in this value is shown as 17.7**0** ± 0.01. The estimated digits are marked in bold fonts.

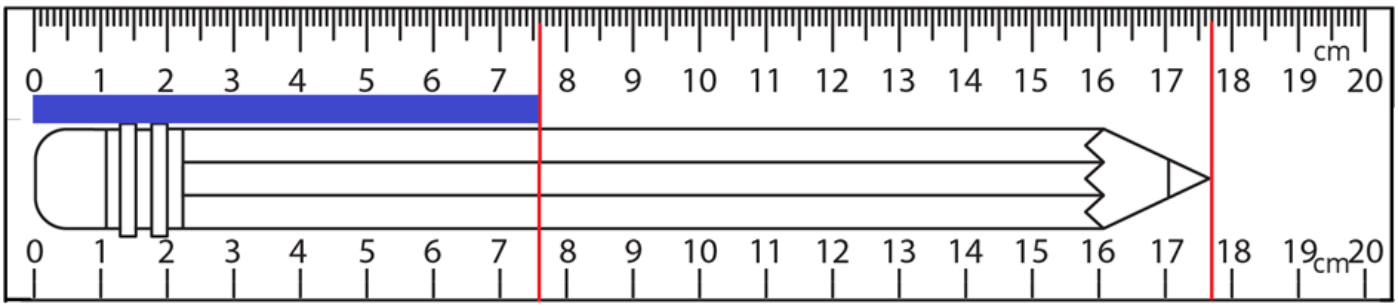


Figure 3.1.4.1: Illustration of measurements that yield inexact numbers. The length of a blue rectangle is 7.6**0** ± 0.01 cm by using a more accurate ruler on the top and 7.**6** ±0.1 cm by using a less accurate ruler on the bottom. The estimated digit is marked in bold fonts.

Rules to determine the significant figures in measured numbers

Significant figures

All the digits reported in the measured value, including the estimated digit, are significant figures (SF).

For example, 1572.6 g, 1573 g, and 1570 g have significant figures of 5, 4, and 3, respectively.

Caution

Note that zero in the last reading 157**0**g is not significant; it is a placeholder zero that is needed to place the estimated digit 7 at tens place.

It is crucial to find significant figures in measured numbers because, when they are used in calculations, the answer cannot have less error than the maximum error in any measured number used in the calculation. The rules to determine the significant numbers in a measured number are the following.

1. All non-zero digits are significant, e.g., 1572 has 4 SFs. The zeros may or may not be significant. In the following examples, the zeros in bold fonts are nonsignificant.
2. Zeros between non-zero digits are significant, e.g., 13**0**5.6 has 5 SFs.
3. Leading zeros are not significant, e.g., **0.0**134 has 3 SFs.
4. Trailing zeros are not significant if there is no decimal point present, e.g., 157**0** has 3 SFs. Trailing zeros are significant if the decimal point is present, e.g., 157**0**. has 4 SFs because the decimal point is present. Similarly, **0.0**124 has 3 SFs, but **0.0**124**0** has 4 SFs because the decimal point is present.
5. Confusion arises when more than one trailing zeros and the decimal point is absent. For example, 15**0**0 g has 2 SFs by convention, but if the balance was accurate to 1**0** g, one of the zero was an estimated digit and was significant. Converting the number to a scientific notation resolves this issue. The coefficient part of the scientific notation shows all the significant figures in the measurement. For example, the number 15**0**0 g, if shown in scientific notation as 1.5 x $1\mathbf{0}^3$ has 2 SFs, but the same number shown as 1.5**0** x $1\mathbf{0}^3$ has 3 SFs.

Rounding the calculated answer involving inexact numbers

When inexact numbers are used in calculations, the answer needs to be rounded to an appropriate number of significant figures, determined by the following rules.

Rules of rounding

1. A number is rounded by keeping the larger digits equal to significant figures and dropping or replacing the remaining smaller digits with placeholder zeros. The placeholder zeros are in bold fonts in the following examples. For example, 13543 becomes 135**00** when rounded to three significant.
2. If the largest digit dropped is 4 or less than 4, it is simply dropped. For example, 23145 becomes 231**00**, when rounded to three significant figures.
3. If the largest digit dropped is 5 or more than 5, then the smallest digit retained is increased by one. For example, 13543 becomes 14**000** when rounded into two significant figures.

Rules for determining the significant figures in a calculated answer

In the following rules, the track of significant figures that dictate the significant figures in the answer is kept by using bold fonts.

1. In addition and subtraction, the answer has the same number of decimal places as the number with the smallest number of decimal places in the original numbers. For example, 13.**2** + 12.252 = 25.**4**52 is rounded to 25.**5** to keep one decimal place.
2. In multiplication and division, the answer has the same number of significant figures as the original number with the smallest number of significant figures. For example, 1.35 x **2.1** = **2.8**35 is rounded to 2.8.

Note

If mathematical operations are performed in a series of steps, keep track of the significant figures but do not round off intermediate answers. Carry as many digits as possible from the intermediate answers to the next calculation step. Round off the final answer following the above rules. For example, (13.**2** + 12.252) x (1.35 x **2.1**) = **25.4**52 x **2.8**35 = **72**.15642 is round to 72 in agreement with **2.8**35 that should have been rounded to 2.8 in a one-step calculation. Rounding the intermediate answers will lead to incorrect final answer of 71 instead of more correct 72, i.e. (13.2 + 12.252) x (1.35 x 2.1) = 25.5 x 2.8 = 71

Caution

Exact numbers have an unlimited number of significant figures, which means they do not restrict the significant figures in the calculated answer.

Example 3.1.4.1

If 12 oranges weigh 1572.6 g calculate the mass of 1 orange in grams?

Solution

$$1572.6\ \text{g}/12 = 131.05\ \text{g}$$

Explanation: The answer has 5 SFs because 12 is a counted number and exact. The only inexact number in the calculation that dictates the significant figures in the solution is 1572.6, which has 5 SFs.

Example 3.1.4.2

One dozen oranges were were sold 11 times. Calculate the total oranges sold?

Solution

$$12 \times 11 = 132\ \text{oranges}$$

Explanation: The answer is not rounded because both the numbers in the calculation are exact, so the answer is also exact with unlimited significant figures.

This page titled 3.1.4: Significant Figures is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.1.5: Unit conversions

Conversion of values in one unit to the same value in another unit, as illustrated in Figure 3.1.5.1 is often needed in scientific calculations. The unit conversion includes the following.

- Conversion of the same type of measurement in the same system of measurement, e.g., conversion of a measured value of the length in meters to kilometers in SI;
- conversion of the same type of measurement in different systems of measurements, e.g., conversion of a measured value of the length in kilometers from SI to miles in the English system; and
- conversion of one type of measurement to another type of measurement, e.g., conversion of a measured value of the mass in g to volume in mL of a substance using the density of the substance.



Figure 3.1.5.1: Illustration of unit conversions. Source: https://www.hiclipart.com/free-trans...fksbi/download

Conversion factors

The conversion factors are derived from equality between the given unit and the desired unit. For example, 1 cm = 10^{-2} m is equality between centimeter and meter. The conversion factors are derived from the equality by the following steps.

Both sides of the equality are divided by one side to get one conversion factor. For example, $1 = \frac{10^{-2}\ m}{1\ cm}$, which is a conversion factor for cm to m.

Then both sides are divided by the other side of the equality to get the second conversion factor. For example, $1 = \frac{1\ cm}{10^{-2}\ m}$, which is a conversion factor for m to cm.

Since both the conversion factors are equal to one, multiplying a value with a conversion factor changes the number and the unit, such that the new number and the new unit together represent the same value. The unit of the given number should be opposite to the same unit in the conversion factor, i.e., numerator versus denominator or denominator versus numerator, to cancel them out and leave the desired unit in the answer. For example, 1.83m is converted to cm unit using the conversion factor $\frac{1\ cm}{10^{-2}\ m}$ as:

$$1.83\ \cancel{\text{m}} \times \frac{1\ \text{cm}}{10^{-2}\ \cancel{\text{m}}} = 183\ \text{cm}$$

Keep track of the units that cancel out and the unit left in the answer. If all the units cancel out, leaving only the desired unit means the chosen conversion factor is correct. An incorrect conversion factor leads to an unwanted unit in the answer, e.g., in the above calculation if incorrect conversion factor is chose, it will lead to:

$$1.83\ m \times \frac{10^{-2}\ m}{1\ \text{cm}} = 0.0183 \frac{\text{m}^2}{\text{cm}}$$

, where no unit is canceled and the answer has units that are not the desired. It means an incorrect conversion factor was employed.

Converting the same type of measurement in the same system

The prefixes in the SI, listed in section 1.3 Table 2, establish equality between the base and the prefixed units. For example, centi (c) means 10^{-2}. Therefore

$$1 \text{ cm} = 10^{-2} \text{ m}$$

1cm = 10^{-2} m is an equality that gives two conversion factors:

$$\frac{1\ cm}{10^{-2}\ m} \quad \text{and} \quad \frac{10^{-2}\ m}{1\ cm}$$

Some of the common qualities in SI are listed in Table 1 and the English system in Table 2.

Note

The prefixes are exact numbers. The equalities within the same system of measurement are exact numbers. Therefore the equalities and the conversion factors derived from them are exact numbers. Significant figures in the answers involving exact and inexact numbers are dictated by inexact numbers only.

Example 3.1.5.1

Convert 325 cm to m units?

Solution

$$325\ \cancel{cm} \times \frac{10^{-2}\ m}{1\ \cancel{cm}} = 3.25\ m$$

Note that in this example, the answer has the significant figures the same as in the given number because the conversion factor within the same system of measurement is numbers.

Table 1: Some of the common equalities in SI

Length	Mass	Volume
1 km = 1,000 m	1 kg = 1,000 g	1 L = 10 dL
1 m = 100 cm	1 g = 1,000 mg	1 L = 1,000 mL
1 m = 1,000 mm	1 mg = 1,000 µg	1 L = 1,000,000 µL
1 cm = 10 mm	10-1	1 dL = 100 mL
		1 mL = 1 cm^3 = 1 cc

Table 2: Some of the common equalities in the English System

Length	Mass	Volume
ft = 12 in.	1 lb = 16 oz	1 qt = 4 cups
1 yd = 3 ft	1 ton = 2,000 lb	1 qt = 2 pints
1 mi = 5,280 ft		1 qt = 32 fl oz
		1 gal = 4 qt

Conversion of the same type of measurement in different systems

Table 3 lists some common qualities between SI and the English systems.

> These equalities between different systems usually have one side in the equality, which is the number 1, as exact, while the other side is considered an inexact number.

For example, 1 kg = 2.205 pounds (lb) has an exact number (1 kg) on the left side but an inexact number (2.205 lb) with 4 SFs on the right. Remember that only the inexact numbers dictate the significant figures in the answer.

> There are some exceptions to the above general rule. Some equalities between units in different systems are defined and considered exact. They are stated to be exact in the reference tables.

For example, 1 inch = 2.54 cm is defined, which means both sides are the exact number.

Table 3: some of the common SI-to-English system equalities

Length	Mass	Volume
2.54 cm = 1 in. **defined and exact**	1 kg = 2.20 lb	946 mL = 1 qt
1 m = 39.4 in.	454 g = 1 lb	1 L = 1.06 qt
1 km = 0.621 mi	28.4 g = 1 oz	29.6 mL = 1 fl oz

Conversion of one type of measurement to another

Sometimes, equality between two different units is known under specific conditions. For example, **density (d)** which is mass (m) per unit volume (V), is a relationship between mass and volume of a given substance, i.e., $d = \frac{m}{v}$. The density is a conversion factor used to convert the volume to the mass of a substance. Reciprocal density, i.e., $\frac{1}{d} = \frac{v}{m}$, is the second conversion factor that converts the mass to the volume of the substance.

Example 3.1.5.2

The density of ethanol at 20 ^{o}C is 0.7893 g/mL; what is the mass of 10.0 mL of ethanol?

Solution

Multiply the given volume with the conversion factor that has volume in the denumerator, i.e., \(frac{m}{v} to get the mass desired.

$$10.0 \cancel{\text{mL}} \times \frac{0.7893 \text{ g}}{1 \cancel{\text{mL}}} = 7.89 \text{ g}$$

Note that 1mL is exact, and 10.0 and 0.7893 are inexact numbers with 3SF and 4 SF, respectively. The answer has 3 SFs.

Example 3.1.5.3

The density of gold is 19.30 g/mL; what is the volume of 10.123 g of gold?

Solution

Multiply the given mass with the conversion factor that has a mass in the denumerator, i.e., $fracvm$ to get the mass desired.

$$10.123 \cancel{\text{ g}} \times \frac{1 \text{ mL}}{19.30 \cancel{\text{ g}}} = 0.5245 \text{ mL} \tag{3.1.5.1}$$

Conversion factors derived from chemical equations

Chemical equations show relationships or equalities between reactants and products measured in **moles**. Mole is a SI unit for the amount of substance. Mole equals 6.02 x 10^{23} particles (atoms or molecules) of the substance. Like dozen means 12 of something, mole means 6.02 x 10^{23} atoms of element or molecules/formula units for compounds in a chemical equation.

> For calculations in chemistry, the number of moles of a substance is considered equal to its coefficient in a balanced chemical equation.

For example:

$$2\,H_2 + O_2 \longrightarrow 2\,H_2O$$

gives the following equalities and their corresponding conversion factors:

1. equality: 1 mole O_2 = 2 moles H_2, conversion factors: $\frac{1 \text{ mol } O_2}{2 \text{ mol } H_2}$ and $\frac{2 \text{ mol } H_2}{1 \text{ mol } O_2}$
2. equality: 1 mole O_2 = 2 mole H_2O, conversion factors: $\frac{1 \text{ mol } O_2}{2 \text{ mol } H_2O}$ and $\frac{2 \text{ mol } H_2O}{1 \text{ mol } O_2}$ and
3. equality: 2 moles H_2 = 2 moles H_2O, conversion factors: $\frac{2 \text{ mol } H_2}{2 \text{ mol } H_2O}$ and $\frac{2 \text{ mol } H_2O}{2 \text{ mol } H_2}$.

The use of these conversion factors in the calculation is explained in the following example.

Example 3.1.5.4

If 5 moles of oxygen (O_2) is consumed, how many moles of water are produced by the chemical equation mentioned above?**Solution**

Given: 5 mol O_2, Desired: ? mol H_2O

Multiply the given quantity with the conversion factor that has the given unit in denumerator and the desired unit in the numerator:

$$5.0 \cancel{\text{ mole } O_2} \times \frac{2 \text{ mole } H_2O}{1 \cancel{\text{ mole } O_2}} = 10 \text{ mole } H_2O$$

Unit conversion involving more than one conversion factor

Often, there is no direct conversion factor between the given unit and the desired unit. In this situation, convert the given unit to another unit that can, later on, be linked with the desired unit, as explained in the following examples.

Example 3.1.5.5

How many µg are in 10.0 mg?

Solution

Problems like this can be solved in two steps: i) by converting the given unit to the base unit and then, ii) converting the base unit to the desired unit:

$$10.0 \cancel{mg} \times \frac{10^{-3}\ g}{1\ \cancel{mg}} = 1.00 \times 10^{-2}\ g$$

$$1.00 \times 10^{-2} \cancel{g} \times \frac{1\ \mu g}{10^{-6}\ \cancel{g}} = 1.00 \times 10^{4}\ \mu g$$

Note that the first conversion factor converts mg to g, and then the second conversion factor converts g to the desired unit µg. The same calculation can be done using two conversion factors in a rwo:

$$10.0 \cancel{\text{mg}} \times \frac{10^{-3}\ \cancel{\text{g}}}{10\ \cancel{\text{mg}}} \times \frac{1\ \mu g}{10^{-6}\ \cancel{\text{g}}} = 1.00 \times 10^{4}\ \mu g$$

Example 3.1.5.6

What is 100. km/h speed in m/s, where h represents hours?

Solution

This problem asks to convert two units, i.e., km to m and h to s. First convert one unit and then follow on to convert the second unit as:

$$\frac{100.\ \cancel{\text{km}}}{\cancel{\text{h}}} \times \frac{10^{3}\ \text{m}}{1\ \cancel{\text{km}}} \times \frac{1\ \cancel{\text{h}}}{60\ \cancel{\text{min}}} \times \frac{1\ \cancel{\text{min}}}{60\ \text{s}} = 27.8\ \text{m/s}$$

Note that the first conversion factor converts km to m, and then two conversion factors are needed to convert h to s via min.

Example 3.1.5.7

A prescription says a dosage of 0.225 mg of Synthroid to be taken once a day. If tablets in stock contain 75 µg of Synthroid, how many tablets are needed per day?

Solution

The given mass is in mg, while the equality "1 tablet = 75 µg" takes mass in µg. First convert mg to base unit, i.e., g, then from g to needed unit, i.e., µg, and finally take appropriate conversion factor from the two given by the equality to convert µg to tablet, i.e., three conversion factors in a row:

$$0.225\ \cancel{\text{mg}} \times \frac{10^{-3}\ \cancel{\text{g}}}{1\ \cancel{\text{mg}}} \times \frac{1\ \cancel{\mu\text{g}}}{10^{-6}\ \cancel{\text{g}}} \times \frac{1\ \text{tablet}}{75\ \cancel{\mu\text{g}}} = 3.0\ \text{tablet}$$

Example 3.1.5.8

A healthy person has 16% body fat by mass. Calculate the mass of fat in kg of a person who weighs 180. lb?

Solution

Given: mass of a person = 180. lb, Desired: mass of body fat in kg.

16% body fat by mass means: 16 lb body fat = 100 lb body mass, and the equality kg and lb is: 1 kg = 2.20 lb. Take one conversion factor from each equality such that the units cancel out leaving the desired unit in the answer:

$$180.\ \cancel{\text{lb body mass}} \times \frac{16\ \cancel{\text{lb body fat}}}{100\ \cancel{\text{lb body mass}}} \times \frac{1\ \text{kg body fat}}{2.20\ \cancel{\text{lb body fat}}} = 16\ \text{kg body fat}$$

Note that there are three inexact numbers in the calculation, i.e., 180, 16, and 2.20, and the answer as two significant figures in agreement with the smallest significant figure among the inexact numbers.

This page titled 3.1.5: Unit conversions is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.1.6: Equations and graphs

Equalities where both sides have a single term, i.e., monomial, lead to the conversion factors. If one or both sides of the equality have more than one term, i.e., polynomial, it leads to a formula that does the same job, i.e., to convert units. Temperature conversion equations and ideal gas equations and the manipulation procedure of these equations are described below as examples of how to manipulate the questions needed for making calculations.

Temperature conversion equations

There are three temperature scales in common use: Celsius (^{o}C), Kelvin (K), and Fahrenheit (^{o}F), illustrated in Fig. 1.6.1.

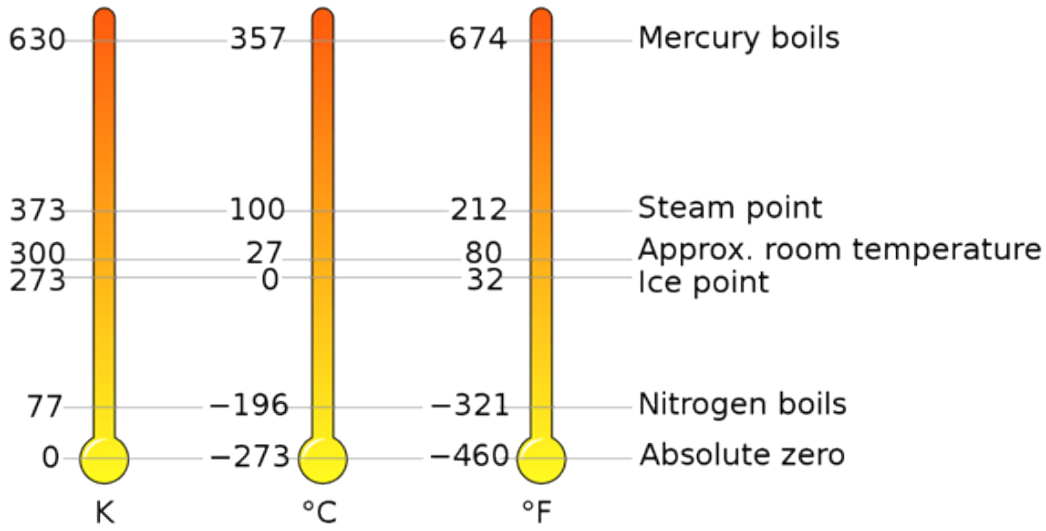


Figure 3.1.6.1: Relations among temperature scales in common use. Source: MikeRun / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Celsius (^{o}C)

The Celsiusscale has O ^{o}C at the freezing point of water and 100 ^{o}C at the boiling point of water. Celsius is the base unit of temperature in the metric system.

The Kelvin scale

Kelvin is the base unit of temperature in the SI. The freezing point of water is 273.15 K, and the boiling point of water is 373.15 K. For most practical purposes, the freezing point of water is reported as 273 K and the boiling point 373 K, i.e., accurate to three significant figures. Celsius scale units are the same size but shifted up by 273 compared to the Kelvin scale. So, the relationship between Kelvin and Celsius is:

$$T_K = T_C + 273,$$

where T_K is the temperature in Kelvin, and T_C is the temperature in degrees Celsius. This equation converts temperature in Kelvin to temperature in Celsius.

A 0 K, also called absolute zero, is the temperature of a matter at which no energy can be removed as heat from the matter. There is no negative temperature on the Kelvin scale.

Fahrenheit (^{o}F)

Fahrenheit is the base unit of the English system, with 32 ^{o}F at the freezing point of water and 212 ^{o}F at the boiling point of water. Fahrenheit is $\frac{5}{9}$ times shorter and shifted up by 32 than Celsius. So the relationship between the two is:

$$T_F = \frac{9}{5} \times T_C + 32,$$

where T_F is the temperature in Fahrenheit, and T_C is the temperature in degrees Celsius. This equation converts temperature in Celsius to temperature in Fahrenheit.

Manipulating temperature conversion equations

The equation for converting Celsius to Fahrenheit is:

$$T_F = \frac{9}{5} \times T_C + 32,$$

Addition or subtraction of the same number on the two sides of an equation does not change the equality. Subtracting 32 from both sides of the above equation leads to:

$$T_F - 32 = \frac{9}{5} \times T_C \cancel{+32} \cancel{-32},$$

$$T_F - 32 = \frac{9}{5} \times T_C.$$

Multiplication or division by the same number on both sides of an equation does not change equality. Remember that multiplication or division should apply to every term on either side of the equality. Enclose the side with more than one term in small brackets and then do the multiplication of division operation so that it applies to each term in the bracket. Multiplying both sides of the above equation with $\frac{5}{9}$ leads to:

$$\frac{5}{9} \times (T_F - 32) = \frac{\cancel{5}}{\cancel{9}} \times \frac{\cancel{9}}{\cancel{5}} \times T_C$$

$$\frac{5}{9} \times (T_F - 32) = T_C$$

Swapping the sides of an equation does not change equality. Swapping the sides in the above equation to bring T_C to the left:

$$T_C = \frac{5}{9} \times (T_F - 32)$$

This is the equation for Fahrenheit to Celsius conversion.

The procedure of rearranging an equation described above applies to all algebraic equations. For example, start with a relationship that converts Celsius to Kelvin:

$$T_K = T_C + 273,$$

subtract 273 from both sides:

$$T_K - 273 = T_C \cancel{+273} \cancel{-273},$$

$$T_K - 273 = T_C,$$

and finally swap the left and right side to bring T_C to the left:

$$T_C = T_K - 273.$$

This is the equation for Kelvin to Celsius conversion.

Ideal gas equation

The ideal gas equation relates more than two variables:

$$PV = nRT,$$

where P is pressure, V is volume, n is the amount of gas in moles, T is the temperature (in K), and R is the proportionality constant called an **ideal gas constant**. Dividing both sides of the equation with V leads to:

$$P \times \frac{V}{V} = \frac{nRT}{V},$$

$$P = \frac{nRT}{V}.$$

It allows calculating the pressure of a gas sample if the amount in moles, temperature in kelvin, and volume of the gas sample are known, along with the value of the constant R in the consistent units. Similarly, rearranging the equation leads to formulas for calculating, V, T or n of a gas sample:

$$V = \frac{nRT}{P}, \quad T = \frac{PV}{nR}, \quad \text{and} \quad n = \frac{PV}{RT}$$

Graphs

The graph is a visual presentation of a relationship between two variables. Fig. 1.6.2 shows a graph that presents a relationship between the volume and pressure of a given amount of gas at a constant temperature, known as Boyle's law.

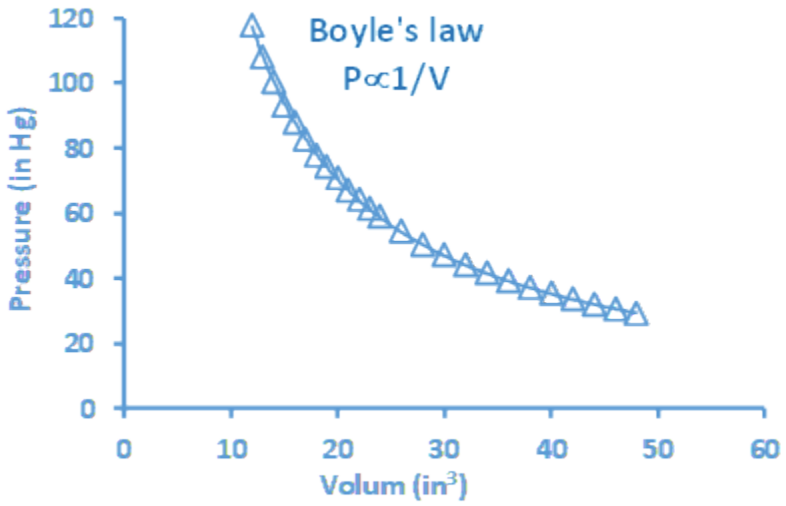


Figure 3.1.6.2: Boyle's law –the relationship between pressure and volume of a given amount of gas at a constant temperature.

The typical components of a graph are the following.

1. A title that tells what the graph is about, e.g., "Boyle's Law: " in the graph of Fig. 1.6.2.
2. Axes: the x-axis is a horizontal line, and the y-axis is a vertical line. Axes usually have an evenly distributed scale starting from zero. The x-axis represents the independent variable, and the y-axis represents the dependent variable. For example, volume is independent, and pressure is the dependent variable in Fig. 1.6.2.
3. Axes labels that tell the variable's name and the units of measurement. For example, volume (in^3) and pressure (in Hg), where in^3 and in Hg are the units of the variables.
4. Symbols representing experimental points. For example, Δ symbols in Fig 1.6.2, at the crossing of a vertical line starting from an experimental value independent variable on the x-axis and a horizontal line starting from the corresponding value of the dependent variable on the y-axis.

5. A curve connects the experimental points and shows the trend in the relationship. For example, the cure in Fig. 1.6.2 tells that the pressure decreases as the volume increases.

Interpretation of a graph

Interpretation is reading the trend or relationship between the variables plotted. For example, Fig. 1.6.2 shows that the pressure decreases as the volume of gas increases. The curve also allows reading the value of one variable from the value of the other. For example, if the volume is 30 in^3,the pressure would be ~50 in Hg. To read: draw a vertical line from the given value on the x-axis and a horizontal line from the point where the vertical line crosses the curve. Then, read the value where the horizontal line meets the y-axis. The process is reversed when the given value is of the dependent variable on the y-axis, and the desired is the corresponding value of the independent variable on the x-axis. For example, if the pressure is 40 in Hg, the volume is about 35 in^3. Fig. 1.6.3 is another example of a graph that shows a relationship between the molar mass and density of the gases at a constant temperature. The cure in this graph tells that the density of gases increases as the molar mass increases.

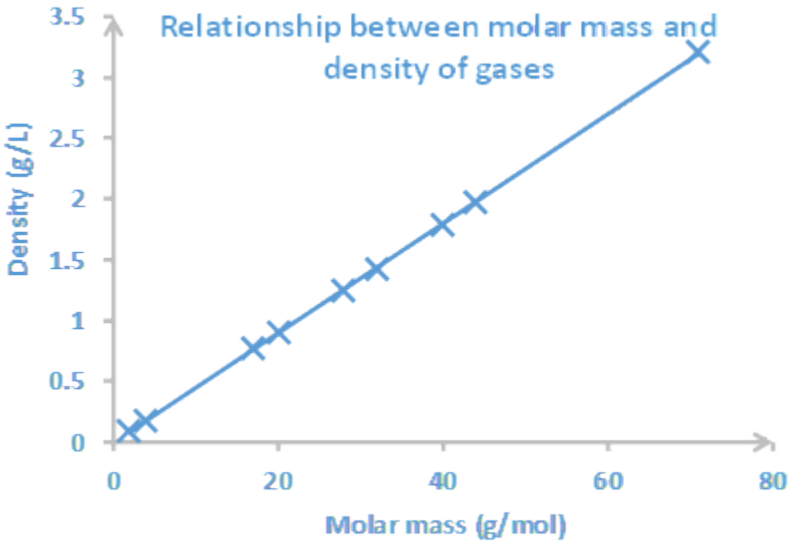


Figure 3.1.6.3: Curve showing gases having lower molar masses are less dense at the same temperature.

This page titled 3.1.6: Equations and graphs is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.1.7: Density and specific gravity measurements

Density

> Density is the mass-to-volume ratio of a substance.

Density is a physical characteristic of matter. Each substance has a characteristic density that can be used as one hint in identifying a substance.

Gases have very low density, usually expressed in g/L. For example, air density is around 1.224 g/L at sea level and 15 ^{o}C. The density of liquids and solids is usually expressed in g/mL. For example, the density of water at 4 ^{o}C is 1.00 g/mL.

Objects that are less dense than water float, and the denser objects than water sink in the water. For example, oil is less dense than water and floats on water. Metals are denser than water and sink in water. The density of some common substances is listed in Table 1

Table 1: Densities of some common substances

Substance	Density (g/mL)
hydrogen	0.000089
carbon dioxide	0.0019
ethyl alcohol	0.7893
water	1.00
magnesium	1.74
table salt	2.16
aluminum	2.70
iron	7.86
copper	8.92
silver	10.50
lead	11.34
mercury	13.59
gold	19.30

Density measurement

Density (d) is calculated from the mass (m) and volume (V) of a substance by the formula:

$$d = \frac{m}{V}$$

Mass is usually measured using an analytical balance. The volume of liquids can be measured using a graduated cylinder, pipet, or density bottle. The volume of regular solids can be calculated from the geometric parameters. For example, the volume of a rectangle is equal to length x width x height. The volume of a cube is equal to the edge length cubed.

The volume of an irregular shaped sold is usually measured because the substances that are denser than water sink and displace an equal amount of water. Fig. 1.7.1 illustrates the density measurement of an irregular-shaped solid object that sinks in water, as explained in the following example.

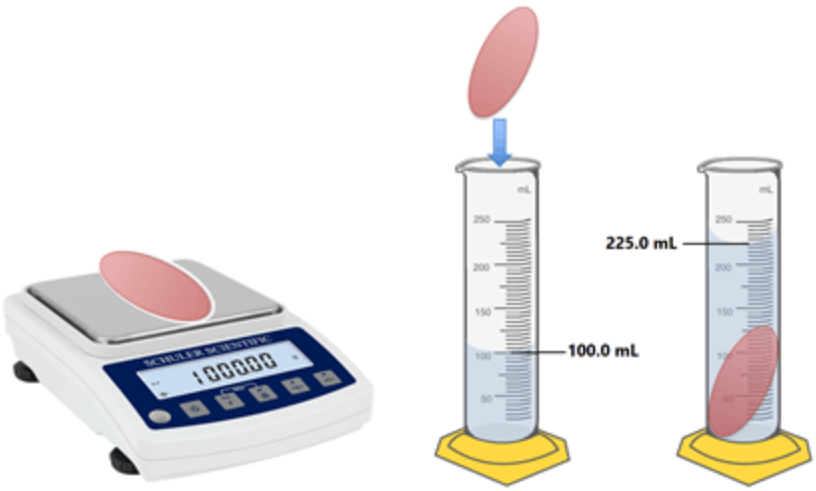


Figure 3.1.7.1: Measuring mass, volume, and density of an irregular solid object. Source: modified images from https://www.hiclipart.com

Example 3.1.7.1

Wwhat the density of the object in Fig. 1.7.1?

Soution

the mass (m) of the object on balance is 1000.00 g. The volume of the object is equal to the volume of the water displaced by the object, which is 225.0 mL – 100.0 mL = 125.0 mL.

$$d = \frac{m}{V} = \frac{1000.00\ \text{g}}{125.0\ \text{mL}} = 8.000 \frac{\text{g}}{\text{ml}}$$

Caution

Most of the time, Calculators give more significant numbers and sometimes less than needed; both need correction. In example 1.7.1, the calculator displays 8, i.e., one significant figure, but three zeros are added to make four significant figures.

Bone density and osteoporosis

Osteoporosis is a bone disease associated with decreased bone density, particularly in older adults. Bones always lose and gain calcium, magnesium, and phosphate. In childhood, the bones build faster than decay, but in old age, the process reverses, and the bones start to thin, loos strength, and become more prone to fracture, as illustrated in Fig. 1.7.2. Hormonal changes, diseases, and some medications contribute to bone thinning. Severe loss of bone density is called osteoporosis.

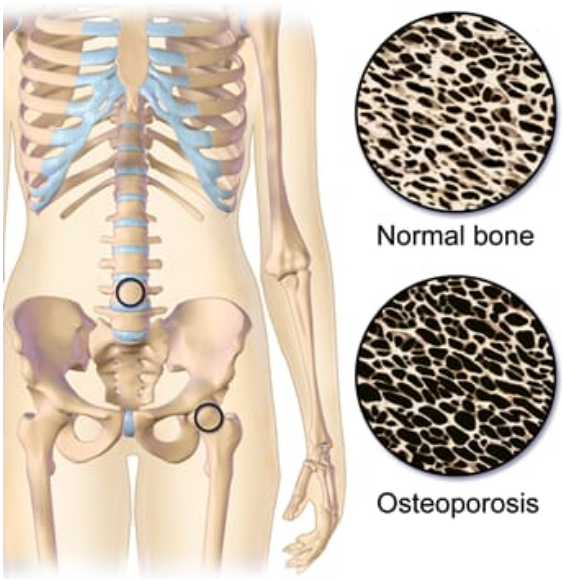


Figure 3.1.7.2: Illustration of Osteoporosis as a result of bone density decrease. Source: BruceBlaus / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Specific gravity

Specific gravity is the ratio of the object's density to the density of water, i.e.:

$$\text{Specific gravity} = \frac{\text{Density of an object}}{\text{Density of water}}$$

Specific gravity is the ratio of the object's density to the density of water, i.e.:

$$\text{Specific gravity} = \frac{\text{Density of an object}}{\text{Density of water}}$$

The units cancel out in the ratio. Therefore, the specific gravity is a unitless number. The density of water is 1.0 g/mL at room temperature, so the specific gravity is equal to the density of the object expressed without a unit.

When substances dissolve in water, the density of the solution is usually different from pure water. For example, the density of whole blood for humans is ~1.060 g/mL. The density of urine varies in the range of 1.0050 g/mL to 1.030 g/mL. Both the blood and urine have dissolved substances in water that increase the density from that of pure water. Both high and low density or specific gravity than the normal range of urine indicates medical problems. An increase in the specific gravity of urine indicates that it is due to an increase in the solutes caused by dehydration, diarrhea, or infection. Similarly, a decrease in solute concentration decreases the specific gravity of urine, which indicates medical problems like renal failure.

Specific gravity measurement

The specific gravity is usually measured using an instrument called a hydrometer. The hydrometer partially submerges in the liquid sample, and the reading on the scale at the air-water junction point is recorded, as illustrated in Fig. 1.7.3.

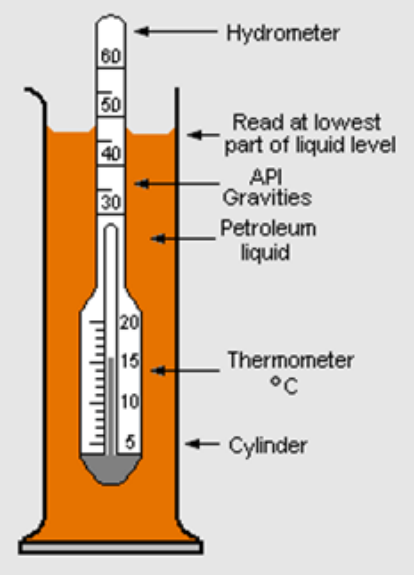


Figure 3.1.7.3: Schematic drawing of an API hydrometer combined with a thermometer. Source: Milton Beychok, CCO

This page titled 3.1.7: Density and specific gravity measurements is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.1.8: Heat and its measurements

Heat

Heat is energy in transfer to and from a thermodynamic system by a mechanism other than the work or the transfer of matter.

Energy exists in different forms, but energy transforms from one form to another through work or heat. In chemical reactions, a significant portion of energy transfer happens through heat. Green leaves in plants trap the energy from light and store it as chemical energy in the form of glucose synthesized from water and carbon dioxide, as illustrated in Fig. 1.8.1.

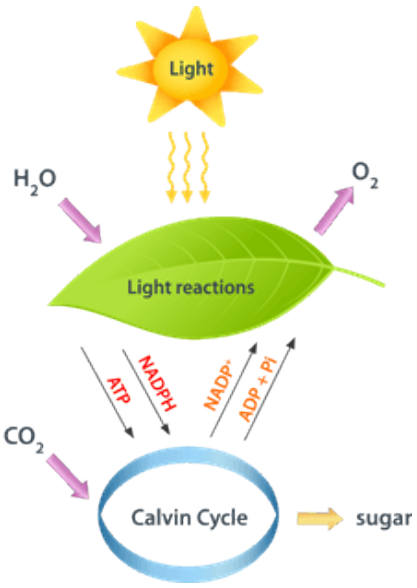


Figure 3.1.8.1: Photo-synthesis traps the light energy and stores it as chemical energy. Source: https://www.hiclipart.com/free-trans...gzitr/download

A reverse process happens when we eat food, i.e., the food substances are converted to carbon dioxide, water, and energy, as illustrated in Fig. 1.8.2. Some energy is used to maintain our body temperature at an average level. Another portion is used to drive chemical reactions that consume energy and synthesize substances we need.

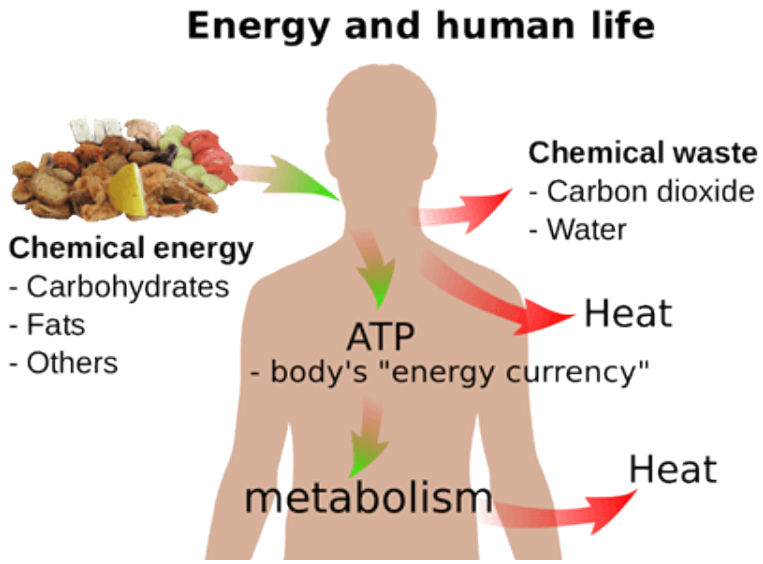


Figure 3.1.8.2: Basic overview of energy and human life.

Source: Mikael Häggström / Public domain

Body Temperature

The human body continually loses energy as heat to the environment. The heat released by exothermic reactions in the food digestion process supplies the heat needed to maintain the body temperature at an average level. Human body temperature varies over a small range, as illustrated in Fig. 1.8.3. Hypothermia is <35.0 °C (95.0 °F), normal body temperate is 36.5–37.5 °C (97.7–99.5 °F), fever and hyperthermia is >37.5 or 38.3 °C (99.5 or 100.9 °F), Hyperpyrexia is >40.0 or 41.0 °C (104.0 or 105.8 °F).

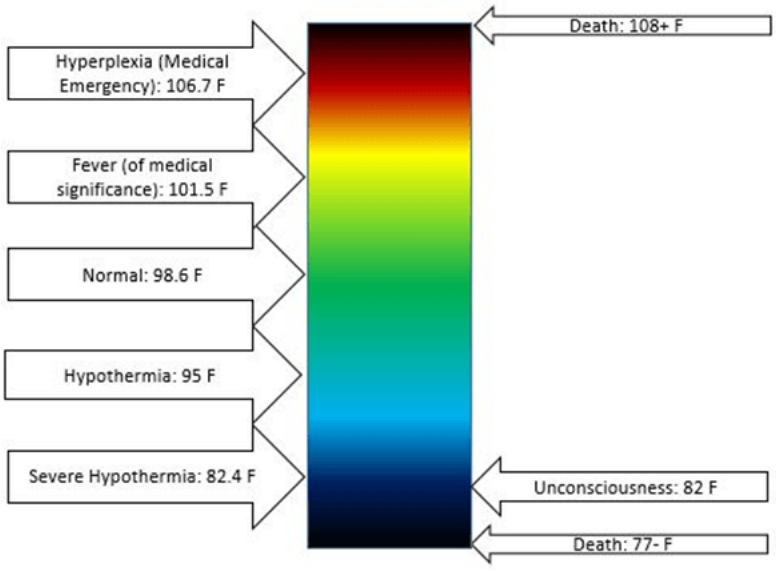


Figure 3.1.8.3: Variation in human body temperate. Source: Foxtrot620 / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Measurement of heat

Temperature is a manifestation of the thermal energy of an object, but the temperature and the energy are not the same. Temperature is a measure of hotness or coldness, and it is intensive, and energy is an extensive physical property of matter. For example, 1 g of water has some energy, but 2 g of water at the same temperature has twice the energy.

Specific heat

The calory (c or cal) is a non-SI unit of heat, work, and energy.

The calory (c or cal)

The energy needed to raise the temperature of 1 g of water by 1 ^{o}C is 1 calory

The heat needed to raise the temperature of 2 g of water by 1 ^{o}C is 2 calory, and to raise the temperature of 2 g of water by 2 ^{o}C is 4 calory. That is, heat energy (q) is directly proportional to both the mass (m) and change in temperature (ΔT) of an object:

$$q \propto m\Delta T$$

Introducing a constant of proportionality, i.e., specific heat (C_s) changes the proportionality to equality, that is known as the **heat equation**:

$$q = C_s m\Delta T$$

, where C_s is Specific heat.

Specific heat of a substance

The heat energy needed to raise the temperature of 1g of a substance by 1 ^{o}C (or 1K) is specific heat if the substance.

The mathematical form of the definition of specific heat is:

$$C_S = \frac{q}{m\Delta T}$$

The units of specific heat are $\frac{cal}{g.^{\circ}C}$, where cal is a non-SI unit of energy. The SI unit of energy is the joule (J). The following is an exact relationship between calory and joule:

$$1 \text{ cal} = 4.184 \text{ J (exact)}$$

So, the specific heat (C_s) in SI units is: $\frac{J}{g.^{o}C}$

The energy unit in nutrition and food is food calory or Calory (C) that is written with capital C, and it is equal to 1000 cal, i.e.,

$$\text{C} = 1000 \text{ cal} = 1 \text{ kcal} = 4184 \text{ J}$$

, where the capital letter C is food calory.

The specific heat is a characteristic physical property of a material. It varies because the energies of vibrational motions of bonds and other motions that contribute to the thermal energy content differ from material to material. Table 1 lists specific heats of some familiar materials.

Water has higher specific heat than other materials, i.e., it absorbs or releases more heat for the same change in temperature. The environmental temperature variation remains moderate in coastal areas because the temperature change of water is less for the same amount of energy absorbed or released. Note that the specific heat of liquid water is different from solid water (ice) and gaseous water (steam).

Table 1: Specific heat of some common substances

Substance	Specific heat ($\frac{J}{g.^{o}C}$)
Al	0.902
C (graphite)	0.720
Fe	0.451
Cu	0.385
Au	0.128
NH_3 (ammonia)	4.70
H_2O (liquid)	4.184
H_2O (ice)	2.06
H_2O(steam)	2.00
C_2H_5OH (l) (ethanol)	2.46
$(CH_2OH)_2$ (l) (ethylene glycol, antifreeze)	2.42
CCl_4 (carbon tetrachloride)	0.861
Wood	1.76
Concrete	0.88
Glass	0.84

Calculations using specific heat

The specific heat values are used in the heat equation ($q = C_s m \Delta T$) to calculate the temperature change for a given amount of heat or the heat energy needed for a given temperature change of a material. A coffee cup calorimeter illustrated in Fig. 1.8.4 is commonly used. It consists of a coffee cup with a lid. It contains water in which the hot or cold object is immersed, or a chemical reaction is performed for heat exchange with the water. A wire loop is used to stir the water. A thermometer records the change in the temperature of the water. The heat absorbed or released by the water is calculated from the data of mass, specific heat, and temperature change of the water.



Figure 3.1.8.4: Photo of coffee cup calorimeter set. Source: Community College Consortium for Bioscience Credentials / CC BY (https://creativecommons.org/licenses/by/3.0)

Example 3.1.8.1

Immersing a hot object in 50.0 grams of water in a coffee cup calorimeter increased the water temperature from 22.0 ^{o}C to 28.8 ^{o}C. How much heat is gained or lost by water?

Solution

Given: m = 50.0 g water, C_s = 4.184 (\frac{J}{g \cdot^{o} C}\), ΔT = T_f – T_i = 28.8 ^{o}C – 22.0 ^{o}C = 6.8 ^{o}C, Desired: heat gained or lost by water, i.e, q_{water} = ?

Calculations:

$$q_{water} = C_s m \Delta T = 4.184 \frac{J}{g^{\circ}C} \times 50.0\ g \times 6.8^{\circ}C = +1400\ J$$

The +ve sign tells water gained 1400 J heat.

Example 3.1.8.2

Ammonium nitrate (NH_4NO_3) is used as a cold pack in hospitals to decrease the temperature of a targeted skin area. When 3.21 g of NH_4NO_3 is dissolved in 50.0 g of water in a coffee cup calorimeter at 25.0 ^{o}C, the temperature decreases to 20.4 ^{o}C. What is the amount of heat absorbed or released by water?

Solution

Given: m of solution = 50.0g H_2O + 3.21 g NH_3NO_3 = 53.2 g, C_s = 4.184 (\frac{J}{g \cdot^{o} C}\), ΔT = T_f – T_i =20.4 ^{o}C - 25.0 ^{o}C = -4.6 ^{o}C, Desired: heat gained or lost by the solution, i.e, $q_{solution}$ = ?

Calculations:

$$q_{solution} = C_s m \Delta T = 4.184 \frac{J}{g^{\circ}C} \times 53.2 \text{ g} \times (-4.6^{\circ}C) = -1.0 \times 10^3 \text{ J}$$

The –ve sign tells water lost heat

Note

In the problems involving the heating of solution, the mass of solution includes the mass of water and the solute. It is assumed that the specific heat and density of the dilute solution are the same as that of pure water.

Example 3.1.8.3

Calcium chloride ($CaCl_2$) is used as a hot pack, i.e., to raise the temperature of a targeted portion of the skin. When $CaCl_2$ dissolves in water, heat is released by the process and absorbed by the surrounding water, increasing the temperature of the water.

When 5.00 g of $CaCl_2$ is dissolved in 50.0 g of water at 23°C in a coffee cup calorimeter, the temperature rises to 39.2^{o}C. What is the amount of heat absorbed or released by water?

Solution

Given: m of solution = 50.0g H_2O + 5.00 g $CaCl_2$ = 55.0 g, C_s = 4.184 (\frac{J}{g \cdot^{o} C}\), ΔT = T_f – T_i =39.2 ^{o}C - 23.0 ^{o}C = 16.2 ^{o}C, Desired: heat gained or lost by the solution, i.e, $q_{solution}$ = ?

Calculations:

$$q_{solution} = C_s m \Delta T = 4.184 \frac{J}{g.^{\circ}C} \times 55.0 \text{ g} \times (16.2^{\circ}C) = 3.7 \times 10^3 \text{ J}$$

The +ve sign tells water gained heat.

The energy released from foods

Foods and fossil fuels are carbon and hydrogen compounds and may also contain some other elements like oxygen, nitrogen, phosphorous, and sulfur. Fuels are used in combustion reactions to release heat for different purposes. The same combustion reactions happen in living systems in several steps to get energy from the foods. For example, the chemical equation for the combustion of glucose is:

$$C_6H_{12}O_6\,(s) + 6\,O_2\,(g) \longrightarrow 6\,CO_2\,(g) + 6\,H_2O\,(l) \quad \Delta H^{\circ} = -2803 \text{ kJ}$$

, where ΔH^{o} is the enthalpy of reaction, i.e., the heat evolved or absorbed at constant pressure conditions. This reaction releases 2803 kJ of energy as heat per mole of glucose. The –ve sign of ΔH^{o} tells that the reaction releases the energy, i.e., the reaction is exothermic. The same amount of energy is released when glucose is digested and converted to the same products in the living system under the same conditions.

Measurement of energy released from food

The heat released from the combustion of a substance is measured using a bomb calorimeter illustrated in Fig. 1.8.5. The

substance is placed in a particular cell surrounded by an excess of oxygen. An electric spark ignites the substance, and the heat released causes a temperature increase in the water and the other materials in the meter. The heat capacity (C_{cp}), i.e., the product of the meter's average specific heat and mass and the water components, is calibrated before the experiment. The heat released from the given amount of the substance is calculated from the heat capacity and the measured temperate increase.

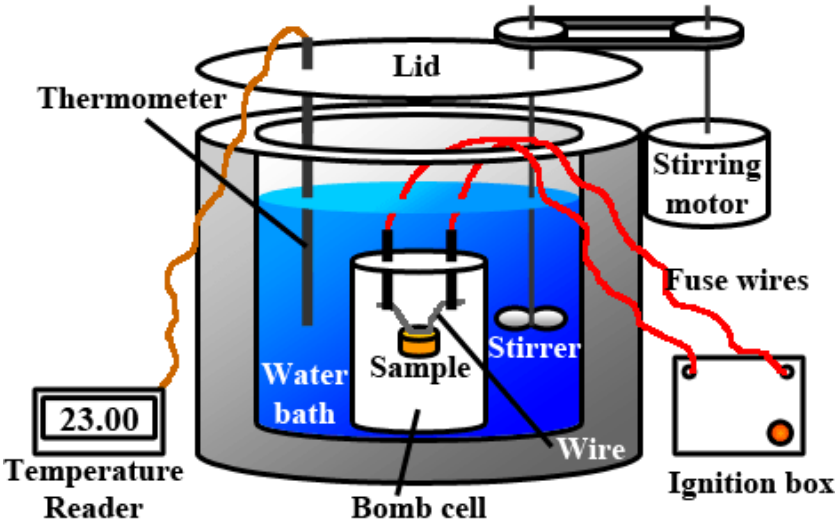


Figure 3.1.8.1: Bomb calorimeter. Source: Lisdavid89, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0>, via Wikimedia Commons

Fuel value

The heat of combustion per g of a food is called the fuel value of the food

Glucose is one of the **carbohydrates**. Carbohydrates are food components that are compounds of carbon, hydrogen, and oxygen in which the ratio of hydrogen-to-oxygen is 2-to-1. The average fuel value of carbohydrates is 17 kJ/g.

The second major class of food is **fat**. For example, the chemical equation of tristearin $C_{57}H_{110}O_6$, which is a fat, is:

$$2\,C_{57}H_{110}O_6\,(s) + 16\,O_2\,(g) \longrightarrow 114\,CO_2\,(g) + 110\,H_2O\,(l) \quad \Delta H^\circ = -71609\ \text{kJ}$$

The –ve sign with energy value tells the energy is released, or the reaction is exothermic. The average fuel value of fats is 38 kJ/g.

The third primary type of food is **protein**. Protein is mainly used as a building material in living systems, but it is also used as a source of energy. The average fuel value of protein is 17 kJ/g.

The average fuel values of the three major food components are listed in Table 2. These values are used to calculate the fule value of food servings, as explained in the following examples.

Table 2: Average fuel value of food components

Food	Fuel value kJ/g
Carbohydrates	17
Proteins	17
Fats	38

Example 3.1.8.1

Calculate the energy released from one cup of orange juice that contains 26 g of carbohydrates, no fat, and 2 g of protein.

Solution

Carbohydrate: $26\ \text{g} \times 17\ \text{kJ/g} = 442\ \text{kJ}$
Protein: $2\ \text{g} \times 17\ \text{kJ/g} = 34\ \text{kJ}$
Total energy released $= 476\ \text{kJ}$

The energy value in kJ can be converted to food calories (C) by using the conversion factor based on the following relations.

$$\text{Food calory C } = 1000\ \text{cal} = 1\ \text{kcal} = 4184\ \text{J}$$

This page titled 3.1.8: Heat and its measurements is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.1.9: Heat and changes in physical states of matter

Among the four physical states of matter, solid has the lowest thermal energy. Intermolecular forces in solids are strong and do not let the molecules slide past each other. The molecules and the bonds in them can still have vibrational motions that account for the thermal energy contents of the material.

The temperature reflects the thermal energy content of the material—the addition of heat increase the vibrational motions, and temperature increases. Ultimately, the solid changes to a liquid and the liquid changes to a gas phase as more heat is added, as illustrated in Figure 1.9.1.

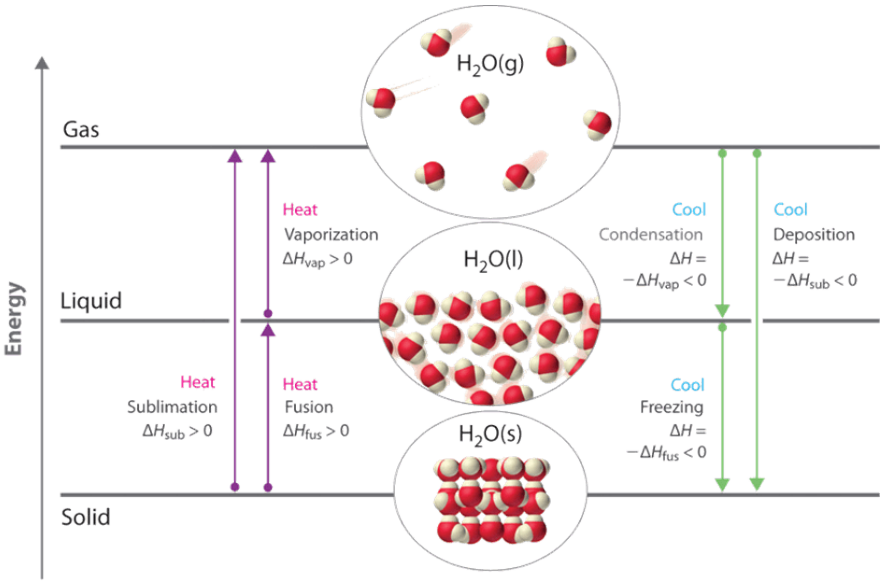


Figure 3.1.9.1: Illustration of the relationship between energy and phase changes of matter. Source: https://www.hiclipart.com/free-trans...poitg/download

Melting and freezing

When the temperature reaches the melting point of the solid upon heating, the temperature does not increase further, but the sold changes gradually to the liquid phase. The heat added at the melting point is used to change the particles from a well-arranged form in the solid to an irregular arrangement in the liquid phase. This process is called the melting of solid.

Heat of fusion (ΔH_{fus})

The energy needed to melt a unit amount of the substance is the heat of fusion (ΔH_{fus}).

The heat of fusion is usually expressed in the units of joules per gram ($\frac{J}{g}$) for the unit amount in grams or in joules per mole ($\frac{J}{mol}$) for the unit amount in moles.

If heat is removed from a substance at its melting point, the reverse of melting, i.e., freezing, happens, i.e., the liquid gradually changes from liquid to solid phase. The energy equal to the heat of fusion is released during the freezing process. Fig. 1.9.2 shows ice and water at 0 ^{o}C –an example of melting and freezing.



Figure 3.1.9.2: Ice and water at 0 °C –an example of melting and freezing. Source: Ulflund / CC0

Vaporization and condensation

After melting, the heat addition causes an increase in the temperature of the liquid until the boiling point is reached. Some of the molecules in a liquid have high enough kinetic energy to cross the liquid-gas boundary and become gas phase. This process is called vaporization.

Heat of vaporization (ΔHvap

The energy needed to evaporate a unit amount of a liquid is called the heat of vaporization (ΔH_{vap}).

The heat of vaporization is usually expressed in the units of joules per gram ($\frac{J}{g}$) for the unit amount in grams or joules per mole ($\frac{J}{mol}$) for the unit amount in moles.

The reverse of evaporation is called condensation, which releases heat equal to the heat of vaporization. Fig. 1.9.3 demonstrates the co-existence of liquid and gas-phase bromine at room temperate through the simultaneous evaporation and condensation processes.

When the temperature reaches the boiling point of the liquid, the temperature does not increase further, but the added heat is used to evaporate the liquid. Heating increases the temperature of the gas phase after all of the liquid has changed to the gas phase.



Figure 3.1.9.3: Bromine at room temperature –an example of evaporation and

condensation. Source: Alchemist-hp (pse-mendelejew.de) / CC BY-SA 3.0 DE (https://creativecommons.org/licenses...3.0/de/deed.en)

Sublimation and deposition

The solid can change directly to the gas phase without going through the liquid phase. This process is called sublimation.

> The energy required in sublimation (ΔH_{sub}) is the addition of the heat of fusion and the heat of vaporization, i.e.,:
>
> $$\Delta H_{\text{sub}} = \Delta H_{\text{fus}} + \Delta H_{\text{vap}}$$

The reverse of the sublimation is called deposition, i.e., the gas phase changes directly to the solid phase. Fig. 1.8.4 shows the sublimation of iodine crystals on a hot plate and deposition of iodine gas on an ice-cold watch glass.

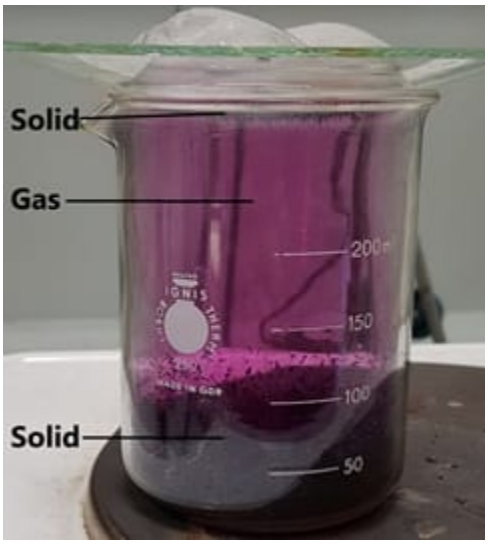


Figure 3.1.9.4: Iodine crystals on hot plate sublime by the heat and the gaseous iodine deposits on ice-cold watch galss. Source: Alvy16 / CC BY (https://creativecommons.org/licenses/by/4.0)

The sublimation is responsible for drying clothes below 0 ^{o}C conditions in cold areas. Sublimation is also used in freeze-drying vegetables and other foods. Bacteria can not grow on dried foods because they need some moisture to grow. Fig. 1.9.5 shows the terminologies related to the phase changes described in the previous paragraphs.

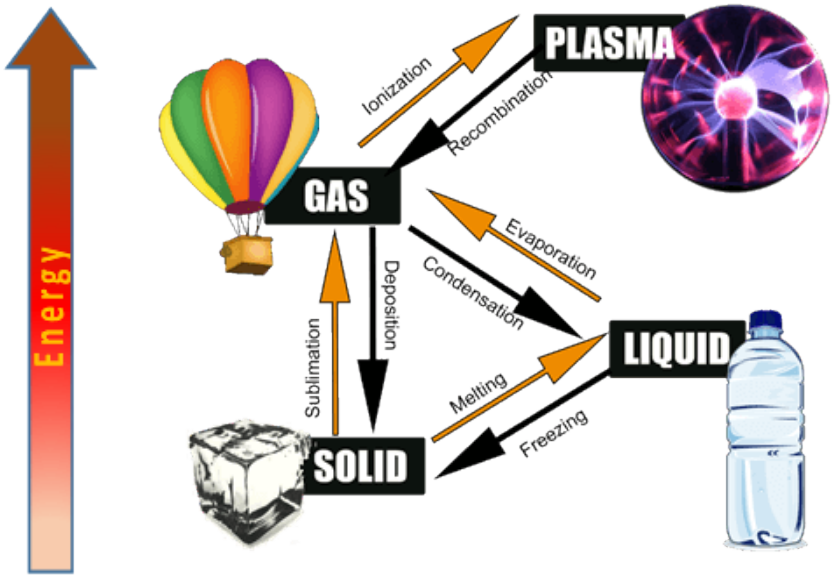


Figure 3.1.9.5: Nomenclature of the phase changes of matter. Source: https://www.hiclipart.com/free-trans...tfefi/download.

Heating curve

A graphical presentation of the relationship of heat added versus the temperature change and phase changes of a matter is called a **heating curve**.

Fig. 1.9.6 shows the heating curve of water. The curve shows the heating of ice initially, followed by co-existing of solid and liquid at the freeing point, then hating of liquid water, then co-existing of liquid and gas phases at the boiling point, and finally the heating of steam –the gas phase of water. The reverse of the heating curve is called the **cooling curve**.

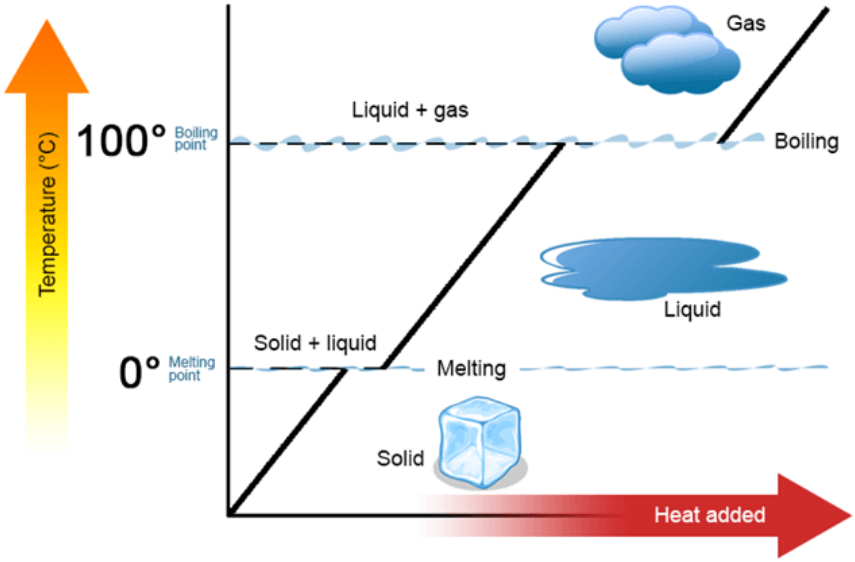


Figure 3.1.9.6: Heating curve of water. Source: Community College Consortium for Bioscience Credentials / CC BY (https://creativecommons.org/licenses/by/3.0)

Heat calculations on heating or cooling curves

The heat required or released can be calculated by using the specific heat of the substance's solid, liquid, and gas phases. The heat of fusion is needed at the freezing point, and the heat of evaporation is needed at the substance's boiling point. The heat calculation is explained in the following example.

Example 3.1.9.1

Calculate the energy required to heat 10.0 g of ice from -20.0 ^{o}C to steam (water vapor) at 110 ^{o}C?

Solution

1st step –heating the ice from -20.0 ^{o}C to the melting point of the ice, i.e., 0.00 ^{o}C:

m = 10.0 g, C_s of ice = 2.06 $\frac{J}{g\cdot^{\circ}C}$, ΔT = 0.00 ^{o}C – (-20.0 ^{o}C) = 20.0 ^{o}C

$$q_1 = C_s m \Delta T = 2.06 \frac{J}{g\cdot^{\circ}C} \times 10.0\ g \times 20.0^{\circ}C = 412\ J$$

2nd step – melting of ice, multiply the heat of fusion with the amout of substance:

m = 10.0 g, ΔH_{fus} = 334 $\frac{J}{g}$.

$$q_2 = \Delta H_{fus} \times m = 334 \frac{J}{g} \times 10.0\ g = 3340\ J$$

3rd step –hating of the water from 0.00 °C to the boiling point of water, i.e., 100 °C:

m = 10.0 g, C_s of liquid water = 4.184 $\frac{J}{g \cdot °C}$, ΔT = 100 °C – 0.00 °C = 100 °C

$$q_3 = C_s m \Delta T = 4.184 \frac{J}{g \cdot °C} \times 10.0\ g \times 100°C = 4180\ J$$

4th step – boiling of liquid water, multiply the heat of vaporization with the amount of the substance:

m = 10.0 g, ΔH_{vap} = 2260 $\frac{J}{g}$

$$q_4 = \Delta H_{vap} \times m = 2260 \frac{J}{g} \times 10.0\ g = 22600\ J$$

5th step –hating of the stem from 100 °C to 110 °C:

m = 10.0 g, C_s of steam = 2.00 $\frac{J}{g \cdot °C}$, ΔT = 110 °C – 100 °C = 10.0 °C

$$q_5 = C_s m \Delta T = 2.00 \frac{J}{g \cdot °C} \times 10.0\ g \times 10.0°C = 200\ J$$

Total heat needed = $q_1 + q_2 + q_3 + q_4 + q_5 = 412\ J + 3340\ J + 4180\ J + 22600\ J + 200\ J = 30700\ J$

$$30700\ J \times \frac{1\ kJ}{1000\ J} = 30.7\ kJ$$

Note

The most significant portion of the heat is consumed in boiling the water to steam, i.e., 22.6 kJ out of 30.7 kJ total. The same amount of heat is released when the steam condenses. That is why the steam burn is much more severe than the burn by hot water.

This page titled 3.1.9: Heat and changes in physical states of matter is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 3.2: ELEMENTS

3.2.1: Dalton's atomic theory
3.2.2: Subatomic particles and a modern view of an atom
3.2.3: Atoms of elements
3.2.4: The periodic table
3.2.5: Electrons in atoms
3.2.6: The periodic trends in properties of the elements

This page titled Chapter 3.2: Elements is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.2.1: Dalton’s atomic theory

Atom

Atom is the smallest particle of an element that retains the element's characteristics.

For example, gold is an element. Magnifying a section of the gold surface large enough would look like a packing of atoms, as illustrated in Fig. 2.1.1.

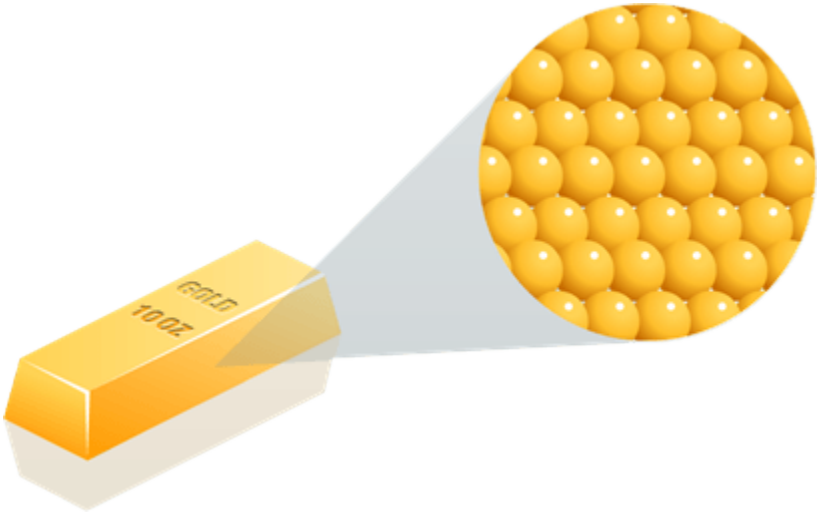


Figure 3.2.1.1: A piece of gold (right) is an element that is composed of particles called atoms, as shown in the model (left). Source: https://www.hiclipart.com/free-trans...ghixq/download.

Dalton's atomic theory was the first significant attempt to explain the basic knowledge of atoms gained over time. Postulates of Dalton's atomic theory are the following.

Postulates of Dalton's atomic theory

1. Elements are composed of tiny indivisible particles called atoms.
2. Atoms of an element are identical but different from atoms of any other element.
3. Atoms of different elements react with each other in a fixed whole numbers proportion to produce a compound.
4. Atoms in a compound can separate and recombine to give new substances. Still, the atoms are neither created nor destroyed in the reaction.

Dalton's atomic theory is the basis of the current atomic theory, though the atoms are no more considered ‘indivisible.’ According to the current knowledge, **subatomic particles** like electrons, protons, and neutrons compose the atoms. However, the subatomic particles do not represent the element.

This page titled 3.2.1: Dalton’s atomic theory is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.2.2: Subatomic particles and a modern view of an atom

Atoms are composed of fundamental subatomic particles: electrons, protons, and neutrons.

Electron was the first subatomic particle discovered. The discovery of electrons is related to the study of cathode rays and the basic knowledge of charges, i.e., there are two types of charges +ve and –ve; like charges repel each other; opposite charges attract each other as illustrated in Fig. 2.2.1; electric and magnetic field deflects the moving charges.

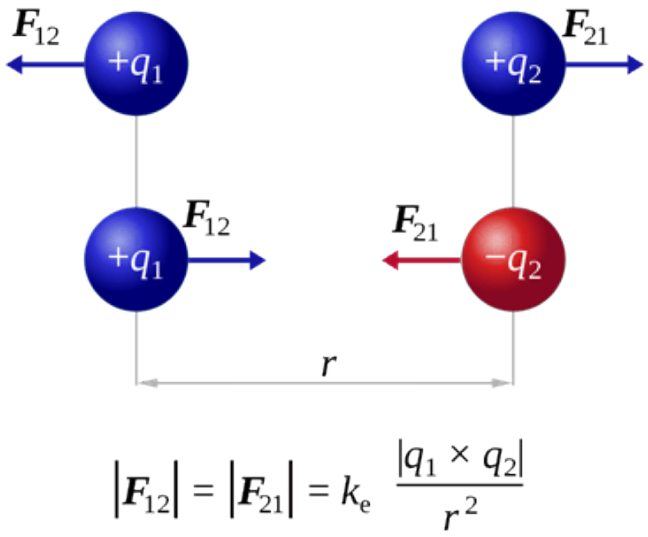
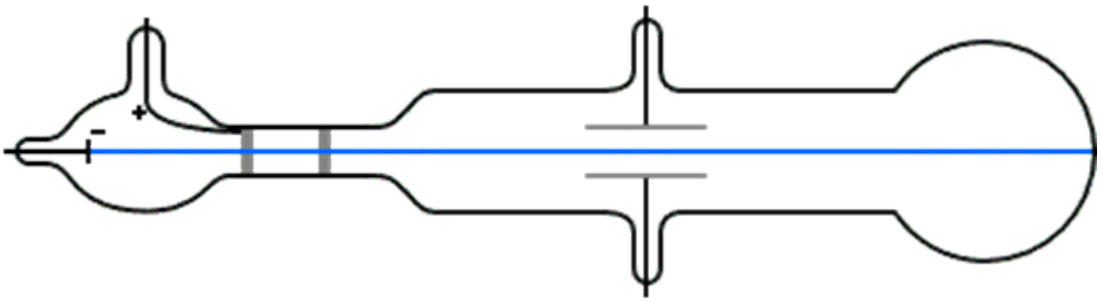


Figure 3.2.2.1: This diagram describes two equal (like) point charges repelling each other and two opposite charges attracting each other, with an electrostatic force F which is directly proportional to the product of the magnitudes of each charge and inversely proportional to the square of the distance between the charges. K_e is Coulomb's constant. Source: File:CoulombsLaw.svg: User:Dna-Dennis / *derivative work RJB1 / CC BY (https://creativecommons.org/licenses/by/3.0)

The discovery of the electron

Cathode rays

Cathode rays are a type of radiation emitted from a cathode (negatively charged electrode) when a high electric field is applied across a pair of electrodes under reduced pressure conditions, as illustrated in Fig. 2.2.2. The cathode rays travel in the inter-electrode space. If there is a hole in the anode, the cathode rays can pass through the hole and keep moving in a straight path. The cathode rays emit light when they strike a fluorescent screen.



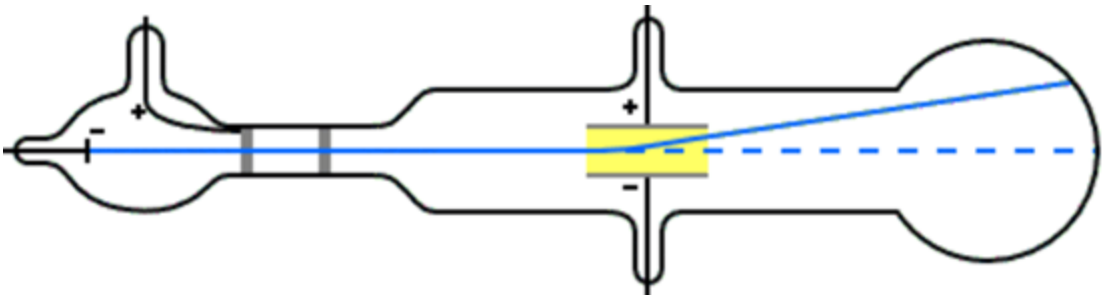


Figure 3.2.2.2: Cathode ray tube with cathode-ray shown deflected by an electric field. *Source: Kurzon / Public domain*

J.J Thomson's experiments -the discovery of electrons

J.J. Thomson studied the cathode rays and found that an electric field deflects the cathode rays towards the positive electrode. This observation indicates that the cathode rays were negative charges. Changing the cathode material did not change the properties of the cathode rays. J.J. Thomson concluded from these observations that the cathode rays were streams of particles, called **electrons**, that are present in the atoms of all elements.

Further discoveries revealed that the charge on the electrons is 1.602×10^{-19} C, and their mass is 9.10×10^{-28} g. The electrons are incredibly light, about two thousand times lighter than the lightest atom.

Plum Pudding Model of an atom

Based on the information from cathode ray experiments, J.J. Thomson concluded that there are –ve charge electrons and a +ve matter representing almost all the atom's mass. He proposed the plum-pudding model of atoms, i.e., the positive matter is like a diffused cloud or jelly that occupies the atom's space, and electrons are embedded in it like fruits in jelly in the case of the plum-pudding dessert dish as shown in Fig. 2.2.3.

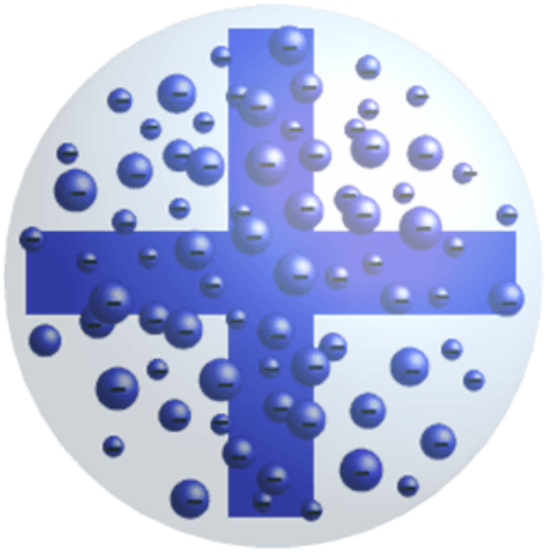


Figure 3.2.2.3: Plum pudding model of an atom. Atoms were initially thought to contain many hundreds or thousands of electrons as shown in this schematic representation of the plum pudding model. *Source: Tjlafave / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)*

The discovery of the nucleus of an atom

α-Rays

α-Rays, pronounced as alpha-rays, are high-energy radiations emitted from some radioactive sources. The α-rays are composed

of α-particles that are helium atoms without any electrons.

Rutherford's gold foil experiment -the discovery of the nucleus

Rutherford tested the plum pudding model of the atom by bombarding α-rays on a thin gold foil. He expected that α-particles to pass through the gold foil un-deflected like bullets fired through a Styrofoam sheet. He observed that although most of the α -particles passed through the gold-foil un-deflected, one in ~20,000 deflected at larger angles, as illustrated in Fig. 2.2.4.

Since the plum-pudding model of an atom could not explain the deflection of the α-particles, Rutherford concluded that there was a tiny but very dense region in the center of an atom, now called the nucleus, that deflected the α-particles. Rutherford's gold foil experiment led to the discovery of the nucleus.

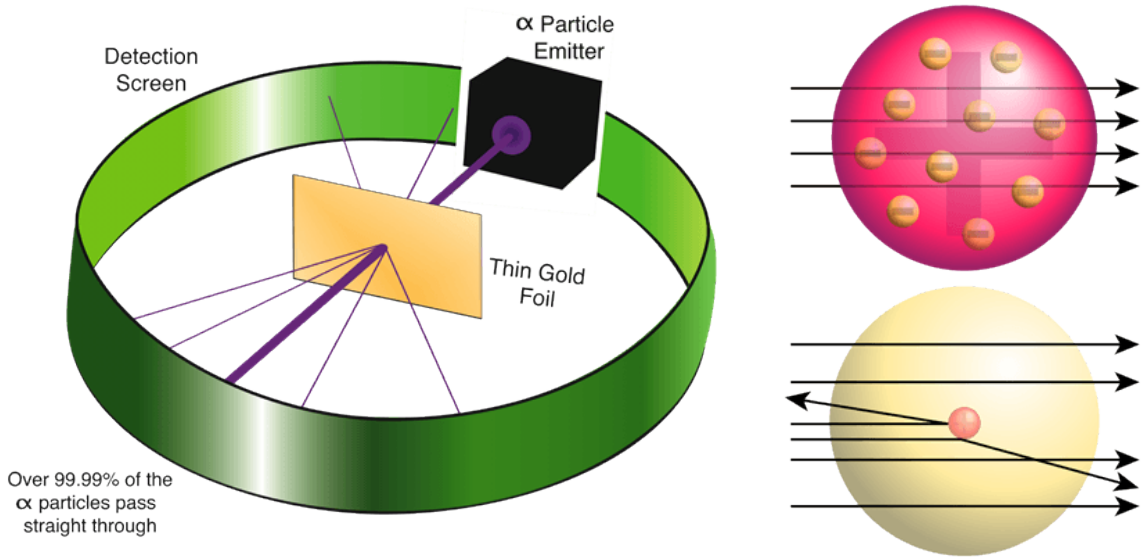


Figure 3.2.2.4: Illustration of Rutherford's gold foil experiment (left). Interpretation of the results – Almost all the atom's mass is in a tiny nucleus space. α-particles that come close to the atom's nucleus are deflected (right). Source: https://www.hiclipart.com/free-trans...xivja/download and https://www.hiclipart.com/free-trans...txlqr/download

The discovery of proton

Rutherford predicted that a positively charged fundamental particle of atoms should reside in the nucleus. Later on, Rutherford observed that shining α -rays on nitrogen gas produced positively charged particles called protons. The protons are about 20,000 times heavier than electrons but carry a +ve charge equal in magnitude to the –ve charge on an electron.

The discovery of the neutron

The mass of protons and electrons did not account for an atom's overall mass, which led to a search for another subatomic particle. James Chadwick discovered that bombarding α -rays on a beryllium target produced a highly penetrating radiation consisting of a beam of neutral particles, now called neutrons. The presence of neutrons in the nucleus accounts for the missing mass of atoms.

Other sub-atomic particles have been discovered, e.g., quarks that constitute protons and neutrons, but their knowledge is not critical for understanding basic chemistry.

Modern view of an atom

Protons and neutrons reside in the nucleus with a diameter of 10^{-15} m. Electrons occupy the region outside the nucleus with a diameter of 10^{-10} m, as illustrated in Fig. 2.2.5. If the nucleus is about the size of a marble, the atom would be about the size of a

soccer field.

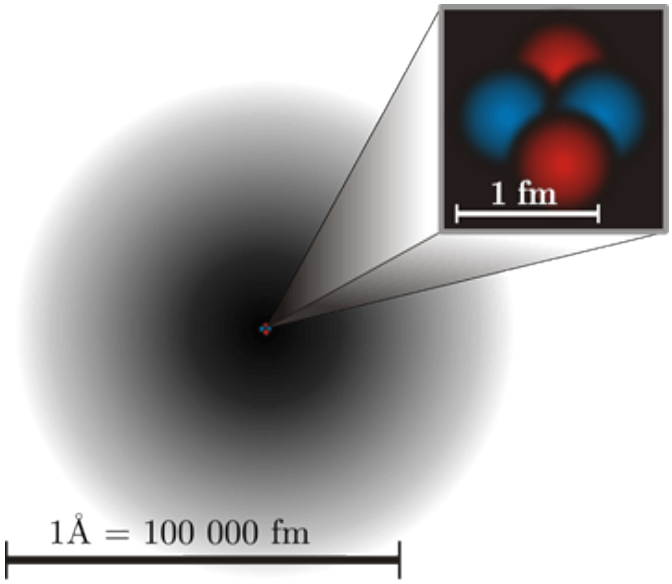


Figure 3.2.2.5: depiction of the atomic structure of the helium atom. The darkness of the electron cloud corresponds to the line-of-sight integral over the probability function of the 1s atomic orbital of the electron. The magnified nucleus is schematic, showing protons in pink and neutrons in purple. Source: User:Yzmo / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/)

Properties of the subatomic particles

The basic SI units of mass, electric charge, and distance are too big for atomic-scale measurements. New units are defined for this purpose, i.e.:

Atomic mass unit (amu)

Atomic mass unit (amu) which is 1/12 of the mass of a single carbon atom that has 6 protons and 6 neutrons in it.

$$1 \text{ amu} = 1.660539606660(50) \times 10^{-27} \text{ kg}$$

Electron charge (e)

Charge on one electrons is electron charge (e).

$$1\ e = 1.602176634 \times 10^{-19}\ \text{C}$$

Angstrom(Å)

$$1 \text{ Å} = 10^{-10} \text{ m}$$

These units are usually used for masses, charges, and diameters of atoms. Table 1 lists the basic properties of the subatomic particles.

Table 1: Basic properties of subatomic particles

Particle	Charge (e)	Mass (amu)
Proton	+1	1.0073

Particle	Charge (e)	Mass (amu)
Neutron	0	1.0078
Electron	-1	5.486×10^{-4}

How the subatomic particles are held in the atom

The gravitational force is negligible in the case of the behavior of subatomic particles in the atoms. Although electrons repel other electrons, they stay in a tiny space due to attraction towards the nucleus. Similarly, protons repel each other, but the electrical force is small compared to the strong nuclear force that holds protons and neutrons together in the nucleus.

This page titled 3.2.2: Subatomic particles and a modern view of an atom is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.2.3: Atoms of elements

Symbol of an element

Elements are represented by symbols, the first alphabet of their English or non-English name, written in capital letters. For example, C for carbon, N for nitrogen, and I for iodine. Usually, another alphabet is also chosen from the element's name and written as a small letter, e.g., Al for aluminum, Zn for zinc, and Ca for calcium. Some element symbols are derived from non-English names, e.g., Na for sodium from the Latin name *natium*, Cu for copper from Latin *cupurum*, and Ag for silver from Latin *argentum*.

Atoms of the same element have the same number of protons, and atoms of different elements have different numbers of protons. In other words, the number of protons in an atom defines the element. There are 118 elements known at this time; the number of protons in atoms varies from 1 for hydrogen to 118 for Oganesson (Og).

Atomic number

The number of protons in an atom is called the atomic number.

The atomic number defines the element. A subscript to the left of the symbol of an element represents the atomic number. For example, $_{1}\mathrm{H}$ shows one proton in a hydrogen atom, and $_{6}\mathrm{C}$ shows 6 protons in a carbon atom.

Mass number

The number of protons plus the number of neutrons in an atom is the mass number.

A superscript to the left of the symbol of an element represents the mass number. For example, $^{1}_{1}\mathrm{H}$ is a hydrogen atom with atomic number 1, mass number 1, and no neutrons, while $^{19}_{9}\mathrm{F}$ is a fluorine atom with 9 protons and 10 neutrons.

Number of electrons

The number of electrons in an atom equals the number of protons minus the charge on the atom.

The number of electrons is equal to the number of protons in the case of a neutral atom, as there is no charge on a neutral atom.

Cation

A neutral atom can lose some electrons and become a positively charged particle, called a **cation**.

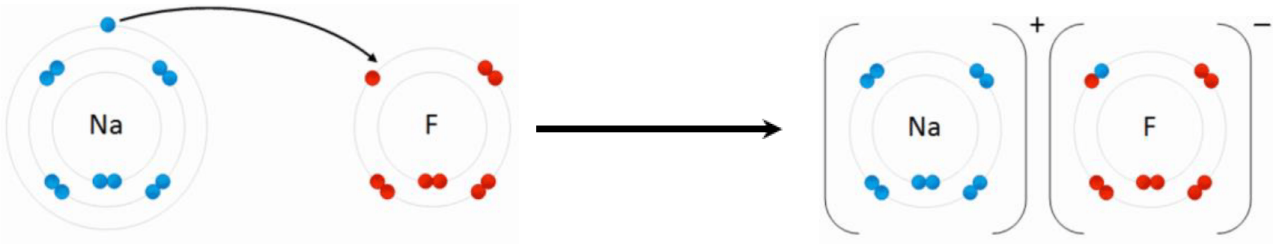
The charge is represented as a superscript on the right side of the element symbol, e.g. $^{1}_{1}\mathrm{H}^{+}$ is hydrogen without any electron, i.e., 1 proton, 0 neutrons, and 0 electrons. $^{20}_{40}\mathrm{Ca}^{2+}$ is calcium with two fewer electrons than protons, i.e., 20 protons, 20 neutrons, and 18 electrons.

Anion

An atom can gain electrons and become a negatively charged particle, called an **anion**.

For example, $^{16}_{8}\mathrm{O}^{2-}$ is oxygen with two more electrons than protons on it, i.e., 8 protons, 8 neutrons, and 10 electrons. A $^{19}_{9}\mathrm{F}^{-}$ is fluorine with one more electron than protons on it, i.e., 9 protons, 10 neutrons, and 10 electrons. Fig. 2.3.1 illustrates the gain or

loss of electrons from neutral atoms.



Symbol:	$^{23}_{11}Na$	$^{19}_{9}F$	$^{23}_{11}Na^{1+}$	$^{19}_{9}F^{1-}$
Protons	11	9	11	9
Neutrons	12	10	12	10
Electrons	11	9	10	10

Figure 3.2.3.1: Comparison of subatomic particles in neutral atoms and ions formed from them after the gain or loss of an electron. Source: modified from: Wdcf / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Calculating the number of protons, neutrons, and electrons

In general, an atom of a hypothetical element X is represented as $^{A}_{Z}X^{I\ charge}$ where *Z* is the atomic number, *A* is the mass number, and *I* is an integer equal to charge number and *charge* the sign of the charge number: + or -.The number of protons, neutrons, and electrons is calculated by using the following formulas:

Number of protons = *A*,

Number of neutrons = *A* – *Z*, and

Number of electrons = *Z* – (*charge I*),

where a *charge* is a + or - sign of the charge number.

Example 3.2.3.1

Calculate the number of protons, neutrons, and electrons in $^{16}_{8}\mathrm{O}$, $^{16}_{8}\mathrm{O}^{2-}$, and $^{16}_{8}\mathrm{O}^{+}$?

Solution

$^{16}_{8}\mathrm{O}$: number of protons = Z = 8, number of neutrons = *A* – *Z* = 16-8 = 8,

and number of electrons = *Z* – (*charge I*) = 8 – 0 = 8.

$^{16}_{8}\mathrm{O}^{2-}$: number of protons = Z = 8, number of neutrons = *A* – *Z* = 16-8 = 8,

and number of electrons = *Z* – (*charge I*) = 8 – (-2) = 8 + 2 = 10.

$^{16}_{8}\mathrm{O}^{1+}$: number of protons = Z = 8, number of neutrons = *A* – *Z* = 16-8 = 8,

and number of electrons = *Z* – (*charge I*) = 8 – (+1) = 8 – 1 = 9.

Note

If charge number *I* is 1 in $^{A}_{Z}X^{I\ charge}$, it is usually not written, but a number more than one is written. For example, $^{16}_{8}\mathrm{O}^{-}$ has change = -1, $^{16}_{8}\mathrm{O}^{2-}$ has charge charge -2, and $^{16}_{8}\mathrm{O}^{+}$ has charge = +1.

Isotopes

All atoms of the same element have the same number of protons but can have a different number of neutrons. For example, ${}^{1}_{1}\mathrm{H}$, ${}^{2}_{1}\mathrm{H}$, and ${}^{3}_{1}\mathrm{H}$ have neutrons equal to 0, 1, and 2, respectively.

> Atoms of the same element that have a different number of neutrons are called **isotopes**.

For examples, ${}^{1}_{1}\mathrm{H}$, ${}^{2}_{1}\mathrm{H}$, and ${}^{3}_{1}\mathrm{H}$ are isotopes of hydrogen illustrated in Fig. 2.3.2. Another example is ${}^{6}_{3}\mathrm{Li}$, and ${}^{7}_{3}\mathrm{Li}$ are two isotopes of lithium. Natural samples of elements usually have almost constant ratios of isotopes. Table 1 lists some Isotopes of elements and their percent abundance in typical natural samples.

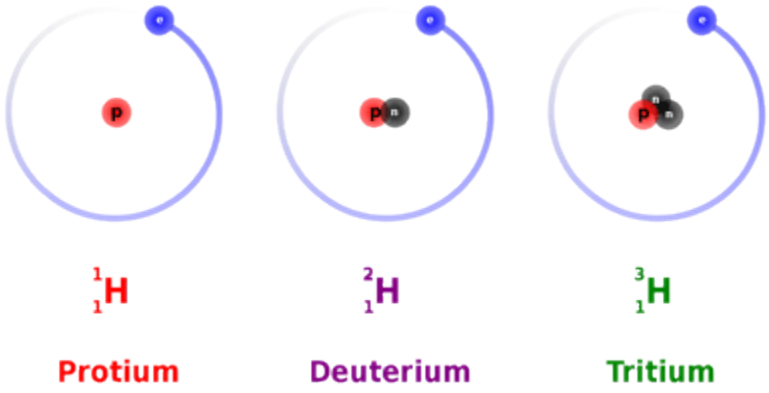


Figure 3.2.3.2: Isotopes of hydrogen. Source: Dirk Hünniger; Derivative work in english - Balajijagadesh / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Table 1: Some of the important isotopes of elements with abundance in typical natural samples

Element	Isotopes	Abundance (%)
Hydrogen	${}^{1}_{1}\mathrm{H}$	99.99
	${}^{2}_{1}\mathrm{H}$	0.01
	${}^{3}_{1}\mathrm{H}$	Negligible
Lithium	${}^{6}_{3}\mathrm{Li}$	7.6
	${}^{7}_{3}\mathrm{Li}$	92.4
Carbon	${}^{12}_{6}\mathrm{C}$	98.93
	${}^{13}_{6}\mathrm{C}$	1.07
	${}^{14}_{6}\mathrm{C}$	Negligible
Chlorine	${}^{35}_{17}\mathrm{Cl}$	75.78
	${}^{37}_{17}\mathrm{Cl}$	24.22

Element	Isotopes	Abundance (%)
Bromine	$^{79}_{35}\mathrm{Br}$	50.69
	$^{81}_{35}\mathrm{Br}$	49.31
Uranium	$^{235}_{92}\mathrm{U}$	0.72
	$^{238}_{92}\mathrm{U}$	99.28

Atomic mass

The **atomic mass** listed in the periodic table is the weighted average of the masses of the isotopes present in a natural sample of the element.

The following formula calculates the atomic mass:

$$\text{Atomic mass } = \sum[(\text{ mass of isotope }) \times (\text{ fractional abuncance of the isotope })$$

, where $\sum$ means summation over all isotopes of the element, the fractional abundance of the isotope is the % abundance divided by 100. Fig. 2.3.3. illustrates how the atomic mass is listed in a periodic table.

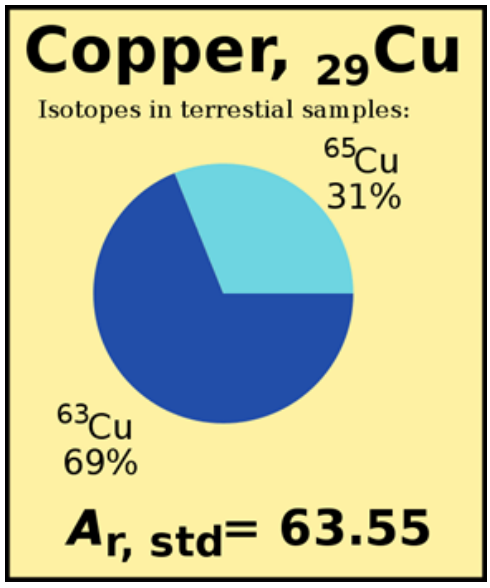


Figure 3.2.3.3: Two isotopes are present: copper-63 (62.9 amu) and copper-65 (64.9 amu), in abundances 69% + 31%. The standard atomic weight for copper is the average, weighted by their natural abundance. Source: DePiep / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Example 3.2.3.2

Calculate the atomic mass of chlorine with two isotopes in nature samples, i.e., $^{35}_{17}\mathrm{Cl}$ of mass 34.969 amu and % abundance 75.78% and $^{37}_{17}\mathrm{Cl}$ of mass 36.996 amu and % abundance 24.22%.

Solution

Formula: $\text{Atomic mass } = \sum[(\text{ mass of isotope }) \times (\text{ fractional abuncance of the isotope })$

Plug in the given values in the formula and calculate:

$$\left(34.969\ \text{amu} \times \frac{75.78}{100}\right) + \left(36.996\ \text{amu} \times \frac{24.22}{100}\right) = 35.45\ \text{amu}$$

Note

1. Atomic masses of isotopes are close to but not the same as their mass numbers. For example, $^{35}_{17}\text{Cl}$ has mass number mass = 35, but the atomic mass of this isotope is 34.969 amu as shown in the above example.
2. The weighted average atomic mass is usually closer to the mass number of the most abundant isotope, e.g., 35.45 amu in the above example is close to the mass number 35 of $^{35}_{17}\text{Cl}$ isotope, which is the most abundant isotope.
3. The periodic table reports the atomic mass as calculated in the above example, i.e., the weighted average of the masses of the isotopes present in the natural sample of the element.

This page titled 3.2.3: Atoms of elements is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.2.4: The periodic table

Early developments

The discoveries of elements happened over a long time. As the list of known elements grew, scientists tried to arrange them based on their properties. **Mendeleev** arranged elements in a table based on atomic masses. It resulted in elements having similar properties placed next to each other in most cases. There were gaps intentionally left in the table for the elements that were predicted based on the knowledge from the periodic table but not yet discovered.

Few exceptions were there observed where the properties of the elements did not agree with the group in which they were placed based on their atomic masses. **Mosely** developed a method to measure atomic numbers based on X-ray spectroscopy. The arrangement of the elements based on the atomic number instead of atomic masses removed the discrepancies in Mendeleev's periodic table.

The modern periodic table of elements

- The modern periodic table of elements arranges the elements according to the increasing order of the atomic number starting from atomic number 1 for H and ending with atomic number 118 for Og, as shown in Fig. 2.4.1.
- The elements are arranged in horizontal rows called **periods** and vertical columns called **groups**.

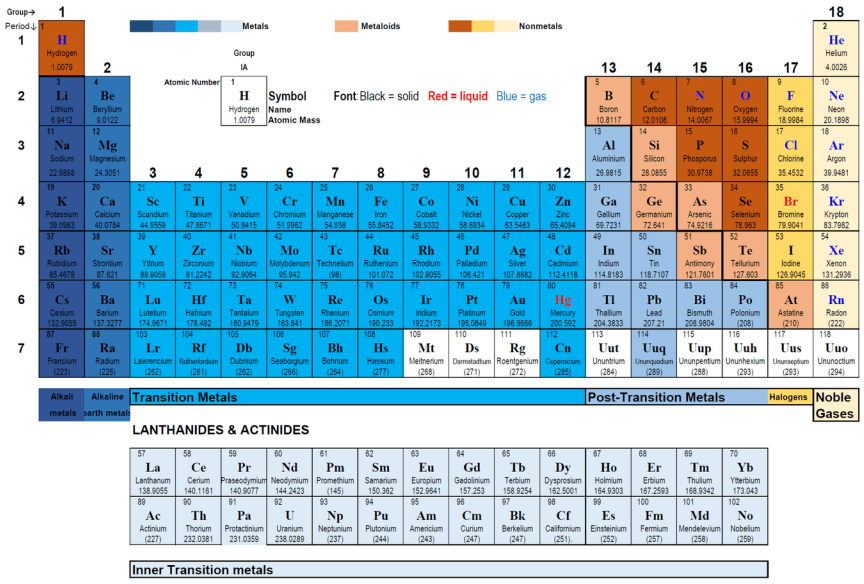


Figure 3.2.4.1: Periodic table of elements. Fort color of element symbols: blue are gases, red are liquids and black are solids at room temperature. Source: Modified from https://www.exceltemplates.org/category/academic)

Periods

The periodic table has seven horizontal rows called periods**.** The periods are numbered: 1 at the top to 7 at the bottom.

1. The 1st period has only two elements: hydrogen in group 1 and helium in group 18, with a gap from group 2 to group 17.
2. The 2nd and the 3rd periods have eight elements each, filling groups 1 and 2 followed by groups 13 to 18, leaving a gap from group 3 to group 12.
3. The periods 4th and 5th periods have eighteen elements that are filled successively from group 1 to group 18.

4. The periods 6th and 7th have 32 elements, each: the first two in groups 1 and 2, the next fourteen elements in separate rows below the table. These two rows of 14 elements each are called **Actinides** and **Lanthanoids**, respectively. Then the next sixteen elements fill the groups 3 to 18.

Groups

The periodic table has 18 vertical columns called groups or families. The groups are numbed starting from 1 on the leftmost and going through to 18 at the rightmost.

Alkali metals

The 1st group is called alkali metals. The alkali metals include lithium (Li), sodium (Na), potassium (K), rubidium (Rb), cesium (Cs), and francium (Fr), shown in Fig. 2.4.2. The alkali metals are the most reactive among the metals in the periodic table. They react vigorously with water, as shown in Fig. 2.4.3.

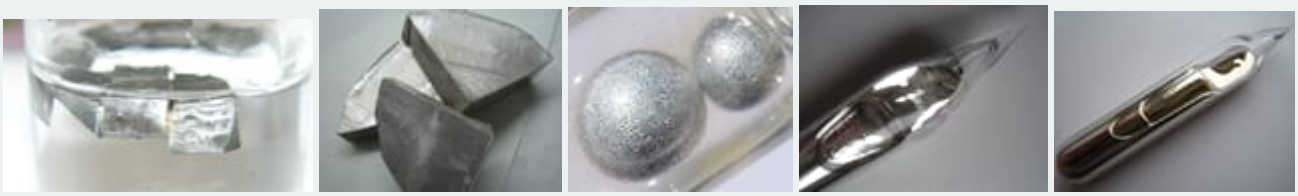


Figure 3.2.4.2: Alkali metals, from left to right: lithium, sodium, potassium, rubidium, and cesium. Source: Tomihahndorf at German Wikipedia, Dnn87 Contact email: Dnn87@yahoo.dk, and http://images-of-elements.com/potassium.php.

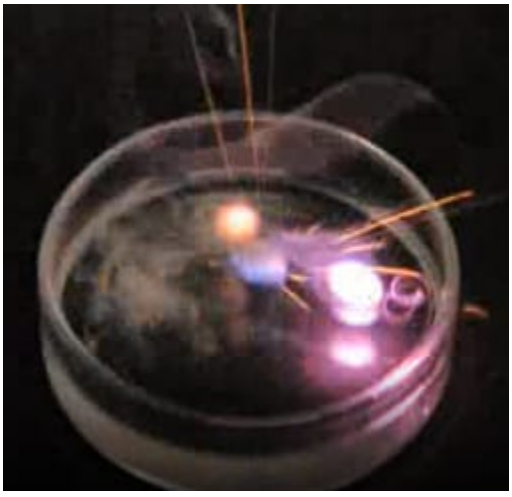


Figure 3.2.4.3: Potassium reacting with water. Source: Ozone aurora / Philip Evans / CC BY-SA (https://creativecommons.org /licenses/by-sa/3.0)

Hydrogen

Hydrogen is in group 1 but is not included in alkaline earth metals. Hydrogen is a nonmetal and has properties quite different from alkali metals or any other group of elements.

Alkaline earth metals

The 2nd group is called alkaline earth metals. It includes beryllium (Be), magnesium (Mg), calcium (Ca), strontium (Sr), barium (Ba), and radium (Ra). They are reactive metals but less reactive than alkali metals. Alkaline earth metals impart characteristic color to a flame. Salts of alkali metals are used in firework formulation to give distinctive colors to the firework, as shown in Fig. 2.4.4.



Figure 3.2.4.4: New years eve fire works at Al Majaz waterfront, Sharjah UAE. Metal salts including alkali metals are used to give distinctive colors in fireworks. Source: Fariz Safarulla / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Transition metals

Groups 3 to 12 are called transition metals. They include precious metals like gold, silver, platinum, and construction metals like iron. Some make catalysts and are found in enzymes and other bio-molecules, like hemoglobin and chlorophyll.

Group 13 to 16

Group 13 to group 16 does not have a unique name. They comprise nonmetals at the top and metals at the bottom of each group called **post-transition metals**. Important nonmetals include carbon, nitrogen, oxygen, phosphorous, and sulfur.

Halogens

Group 17 elements are called halogens. The halogens include fluorine (F), chlorine (Cl), bromine, iodine (I), and astatine (At). The halogens are highly reactive nonmetals. Chlorine is gas, bromine is liquid, and iodine is solid at room temperature, as shown in Fig. 2.4.5.



Figure 3.2.4.5: From left to right: chlorine, bromine, and iodine. Source: W. Oelen / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Noble gases

Group 18 is called noble gases. They include helium (He), neon (Ne), argon (Ar), krypton (Kr), xenon (Xe), and radon (Rn). They are the least reactive of all the elements. Noble gases are used to create an inert atmosphere for chemical reactions. Noble gases are also used in the lighting system because of their chemically inert nature , as illustrated in Fig. 2.4.6.

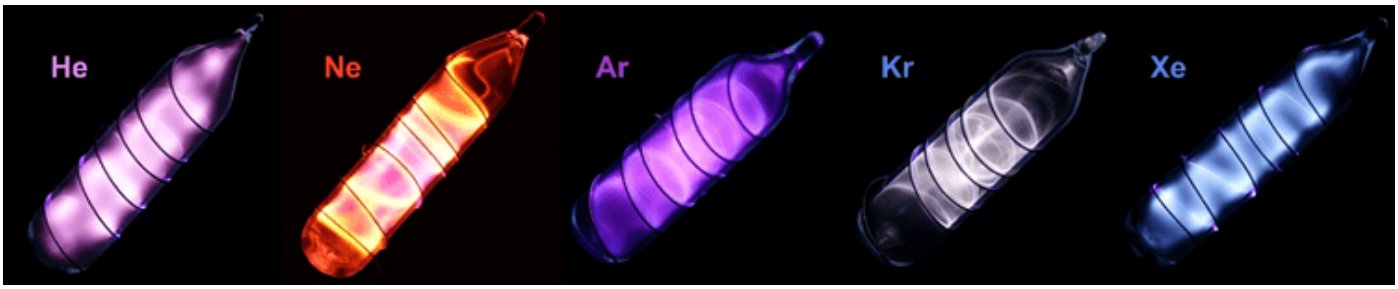


Figure 3.2.4.6: Glowing noble gases in glass vials with low pressure inside. Power supply: 5 kV, 20 mA, 25 kHz. Source: New work Alchemist-hp (talk) www.pse-mendelejew.de); original single images: Jurii, http://images-of-elements.com. / CC BY (https://creativecommons.org/licenses/by/3.0)

Metals, metalloids, and nonmetals

Metals

The elements towards the right-bottom corner of the periodic table are metals except for hydrogen, which is a nonmetal. Metals have common characteristics, including:

1. they are shiny,
2. solid at room temperature (except mercury which is liquid),
3. malleable (can be hammered into sheets) and ductile (can be drawn into wires),
4. good conductors of heat and electricity, and
5. tend to lose electrons and form ionic compounds when they react with nonmetals.

The metals of group 1 are called alkali metals; group 2 are called alkaline earth metals, group 3 to group 13 are called transition metals, the two-row below transition metals in the periodic table are called **lanthanoids**, and **actinides** or **inner transition metals**, and metals in the group 13 to group 16 are called **post-transition metals**. Elements other than transition or inner-transition metals, i.e., the elements of groups 1 and 2 and groups 13 to 18, are collectively called **main group elements or representative elements**.

Metalloids

The dividing line between metals and nonmetals is a staircase line starting from $_{5}\mathrm{B}$ and ending at $_{85}\mathrm{At}$. The elements on the staircase line are metalloids except for aluminum and polonium, which are considered metals. Metalloids have properties in-between metals and nonmetals; e.g., they have moderate heat and electrical conductivity.

Nonmetals

Elements towards the top-right corner of the periodic table and hydrogen are called nonmetals.

1. The nonmetals usually have properties opposite to metals, e.g., they are not typically shiny, brittle if solid, and poor conductors of heat and electricity.
2. Nonmetals tend to make ionic compounds by accepting electrons from metals and making molecular compounds by reacting with each other.

Two groups in nonmetals also have unique names, i.e., group 17 is called halogens**,** and group 18 is called Noble gases.

The most reactive and nonreactive elements

Generally, alkali metals are the most reactive, followed by alkaline earth metals, and halogens are the most reactive nonmetals. Noble gases are the least reactive nonmetals, also called inert gases.

Fig. 2.4.7 shows examples of a metal, a metalloid, and a nonmetal.



Figure $3.2.4.7$: Gold nugget –a metal, pure silicon –a metalloid, and sulfur –a nonmetal (from left to right). Source: Gold: Rob Lavinsky, iRocks.com – CC-BY-SA-3.0 / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0), Silicon: Enricoros at English Wikipedia / Public domain, Sulfur: Benjah-bmm27 / Public domain

Elements essential for life

The elements that are the main constituents of humans and other living organisms are oxygen, carbon, hydrogen, nitrogen, and sulfur. Phosphorous is present in bone, teeth, and DNA. Calcium and magnesium are the main constituents of bones and teeth and perform some other body functions. Sodium and potassium cations are the main electrolytes in body fluids, and chloride anion balances the charge. Iron is present in hemoglobin that carries oxygen to the cells. These elements are essential to life, and they are macronutrients. Besides these, several other elements are needed in a small amount. They are essential for life and called micronutrients. Fig. 2.4.8 shows the macronutrients in pink and micronutrients in blue color in a periodic table.

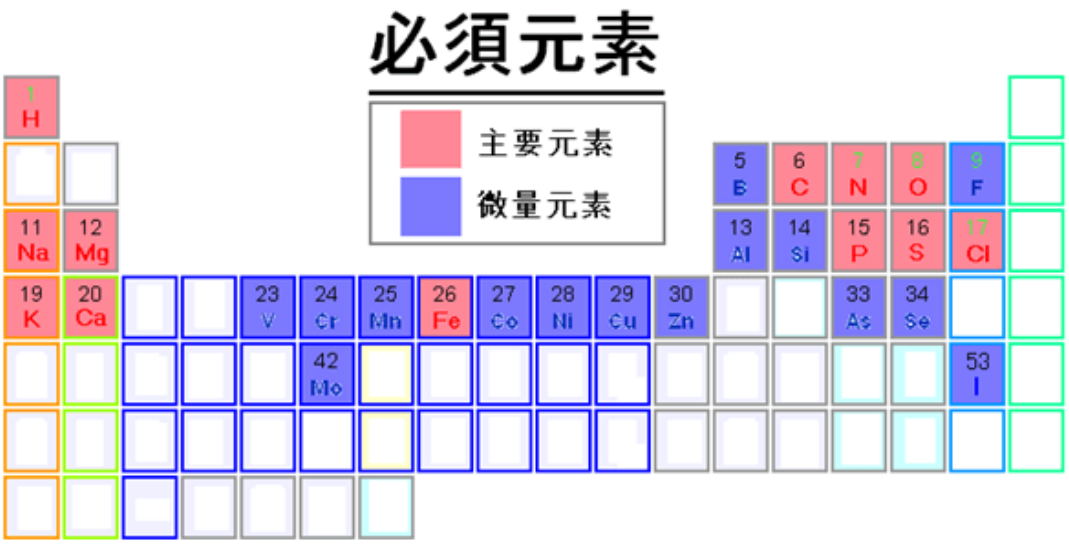


Figure 3.2.4.8: A table of elements important to human life. Pink elements are essential that are macroneutrients; blue elements are important that are microneutrients. Source: Tosaka / Public domain

Physical and chemical states of elements

Elements in gases state at ambient conditions

The term molecule is generally used for an electrically neutral group of two or more atoms held together by chemical bonds. In the kinetic molecular theory of gases, a **molecule** is the smallest particle of an element or compound with a stable and independent existence. Atoms of noble gases exist as independent species in the gas phase at room temperature, i.e., as **monoatomic molecules** like He, Ne, Ar, Kr, Xe, and Rn. Other elements that are gases at room temperature or **diatomic molecules**, i.e., H_2, N_2, O_2, F_2, and Cl_2, are called dihydrogen, dinitrogen, dioxygen, and dichlorine, respectively. Note that usually prefix di is not used for the name of the element, i.e., these are usually called hydrogen, nitrogen, oxygen, fluorine, and chlorine, respectively. Monoatomic species of these elements, i.e., H, N, O, F, and Cl, exist, but they are very reactive species called free radicals, and they do not survive for a long time.

Elements in the liquid state at ambient conditions

Two elements exist as liquids at room temperature, i.e., mercury (Hg), a metal, and bromine (Br_2), a no metal that exists as diatomic molecules. Four elements, francium, cesium, gallium, and rubidium, are solid metals at 25 ^{o}C, but become liquid when the temperature is slightly warmer.

Elements in the solid-state at ambient conditions

All elements not mentioned in the previous two sections are solid at room temperature. They may be diatomic or polyatomic molecules, e.g., I_2, O_3, P_4, S_8, diiodine, ozone, tetraphosphorous, and octasulfur, with two or three, four, and eight atoms, respectively, in a molecule.

Allotropes

> Different forms of the same element in the same physical state are called allotropes.

The allotropes are different structural modifications of the element. For example, carbon exists in several allotropic forms; two of them are shown in Fig. 2.4.9. Another example is O_2, and O_3 are gaseous forms of oxygen.

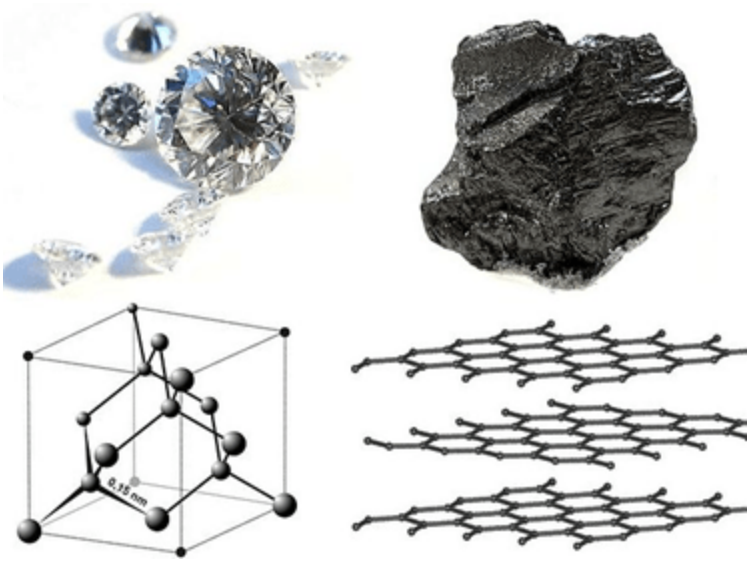


Figure 3.2.4.9: Diamond and graphite are two of the allotropes of carbon: pure forms of the same element that differ in structure. Source: Diamond_and_graphite.jpg: User:Itubderivative work: Materialscientist / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/)

Elements that exist as giant molecules

Some elements exist as **giant molecules**, i.e., a collection of many atoms bonded with each other through a 3D network of bonds. For example, carbon is a giant molecule in several allotropic forms, including diamond, graphite, carbon nanotubes, and fullerenes. Fig. 2.4.9 shows the bonding in diamond and graphite allotropes. The whole piece of diamond is one molecule with carbon atoms interconnected in a 3D network of bonds.

Metals also exist as a collection of many atoms bonded together by metallic bonds. The metal atoms exist as +ions arranged in a well-defined 3D arrangement called crystal lattice with some of the outermost electrons roaming around in the whole piece of the metal, forming a sea of electrons around the metal atoms, as illustrated in Fig. 2.4.10.

The elements in giant molecules and metals are represented by element symbols without any subscript, e.g., C is carbon, Fe is iron, Au is gold, etc.

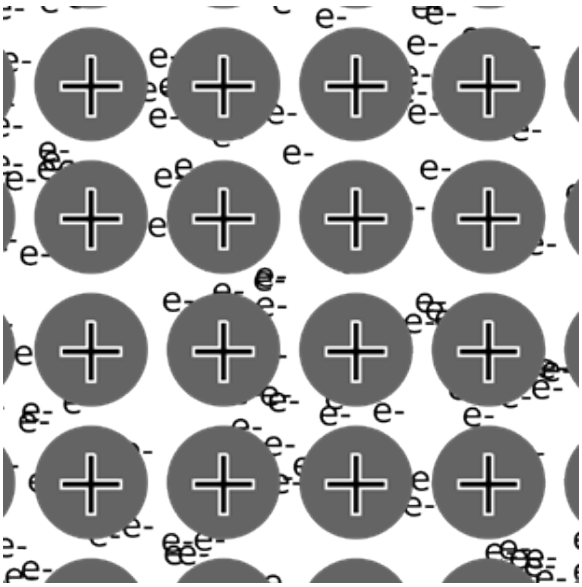


Figure 3.2.4.10: Metal ions in a sea of electrons. Source: Steven Legg (Sjlegg) / *Public domain*

This page titled 3.2.4: The periodic table is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.2.5: Electrons in atoms

The electronic structure of an atom determines the properties of the element. Knowledge of electromagnetic radiation is described first as it plays an essential role in understanding the electronic structure of atoms.

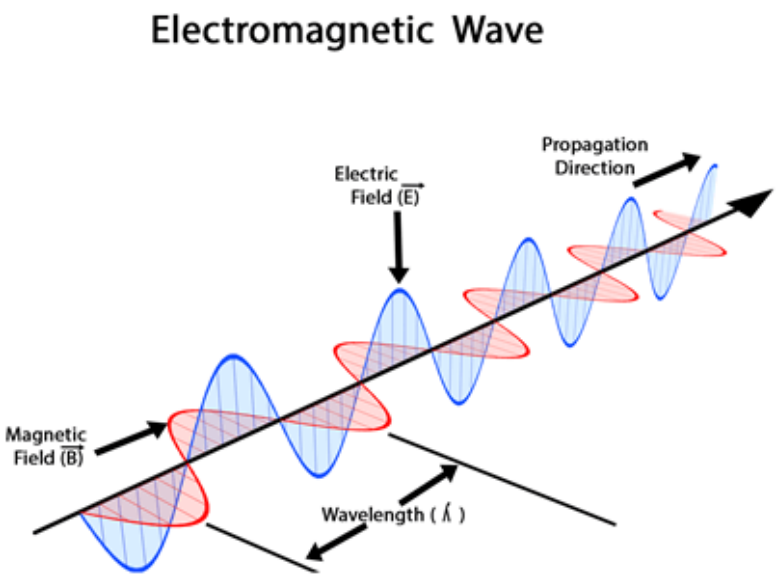


Figure 3.2.5.1: Illustration of electric field, magnetic field, propagation direction and wavelength of an electromagnetic wave. Source: DECHAMMAKL / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Electromagnetic radiations

Electromagnetic radiations are waves that are oscillating electric and magnetic fields. The wave propagates in one direction, e.g., along the x-axis. The electric field oscillates perpendicular to it, e.g., along the y-axis. The magnetic field oscillates perpendicular to both, e.g., along the z-axis, as illustrated in Fig. 2.5.1. The distance between two consecutive maxima or between any two consecutive same phase points along the wave is called **wavelength** (λ,pronounced 'lambda'). The number of waves that pass a reference point in one second is called **frequency** (ν, pronounced 'nu'). The speed of electromagnetic radiation is called the speed of light (c). The speed of light is the product wavelength and frequency, i.e.,

$$c = \lambda\nu$$

The speed of light (c) is 3.00 x 10^8 m/s in a vacuum. The **energy** (*E*) of electromagnetic radiation is directly proportional to frequency, i.e.,

$$E = h\nu$$

, where h is a constant, called plank's constant. Replacing ν with $\frac{c}{\lambda}$ shows that the energy is inversely proportional to the wavelength.

$$E = \frac{hc}{\lambda}$$

Fig. 2.5.2 illustrates the range of electromagnetic radiations that differ from each other concerning wavelength, frequency, or energy. The separation of the radiations based on their wavelength gives a **spectrum**. Visible light is a small portion of the spectrum of electromagnetic radiations, as illustrated in Fig. 2.5.2.

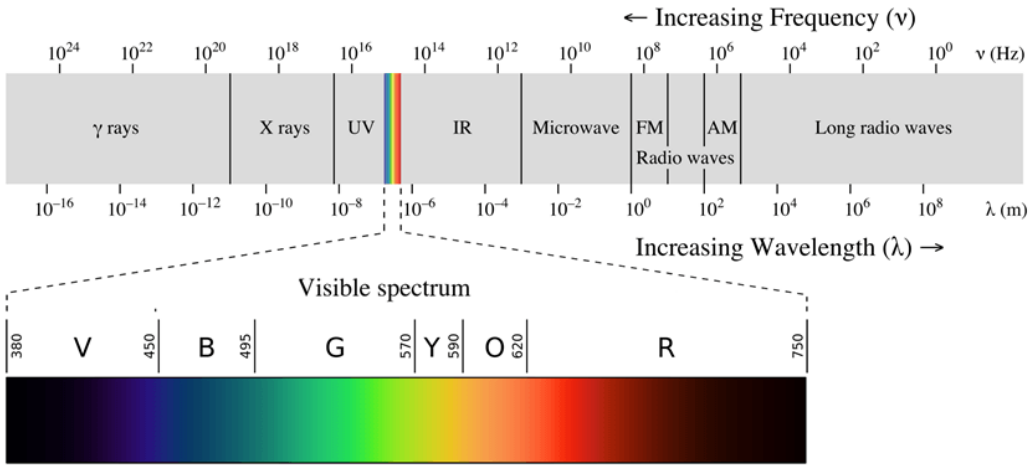


Figure 3.2.5.2: A spectrum of electromagnetic radiations. The visible range of the spectrum is expanded with the wavelength shown in nm units. Source: Philip Ronan, Gringer / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Continuous and line spectrum

The spectrum of sunlight contains radiations of all wavelengths within the visible range. The spectrum that contains all wavelengths in the range is called a continuous spectrum. Spectrum from sodium lamps or emissions from other atoms contains only some discrete wavelengths. The spectrum that contains discrete wavelengths is called the line spectrum. For example, as illustrated in Fig. 2.5.3, the emission from the hydrogen atom is a line spectrum.



Figure 3.2.5.3: Continues spectrum of visible light (400-700 nm range, top) and line spectrum from hydrogen (bottom). Source: it:Utente:Sassospicco / Public domain and Merikanto, Adrignola / CC0

Energy levels of electrons in an atom

The question is why the atomic emission has a discrete wavelength or discrete energy. The answer to this question came from the discoveries that concluded that the electron in an atom does not have continuous energy values; they have discrete energy values called shells and subshells.

The shell

Quantum numbers determine the allowed energy values of an electron in an atom.

- Principal quantum number (*n*) can have any integer value starting from 1, i.e., 1, 2, 3, 4, and so on.
- The smaller the *n*, the lower is the energy state, and the closer the electron is to the nucleus, the more tightly held the electron is by the nucleus.
- The value of *n* defines the shell, i.e., 1st shell has *n* = 1, 2nd shell has *n* = 2, 3rd shell has *n* =3, and so on.

Bohr introduced this concept of quantization of electronic energy levels. Fig. 2.5.4 illustrates Bohr's atomic model.

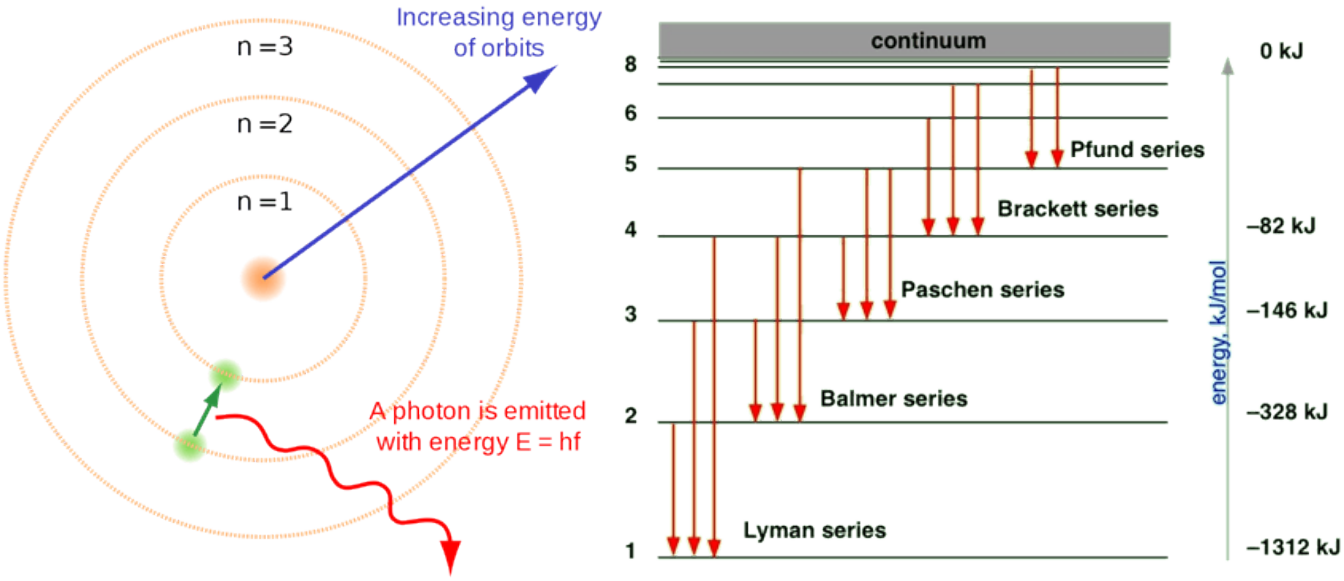
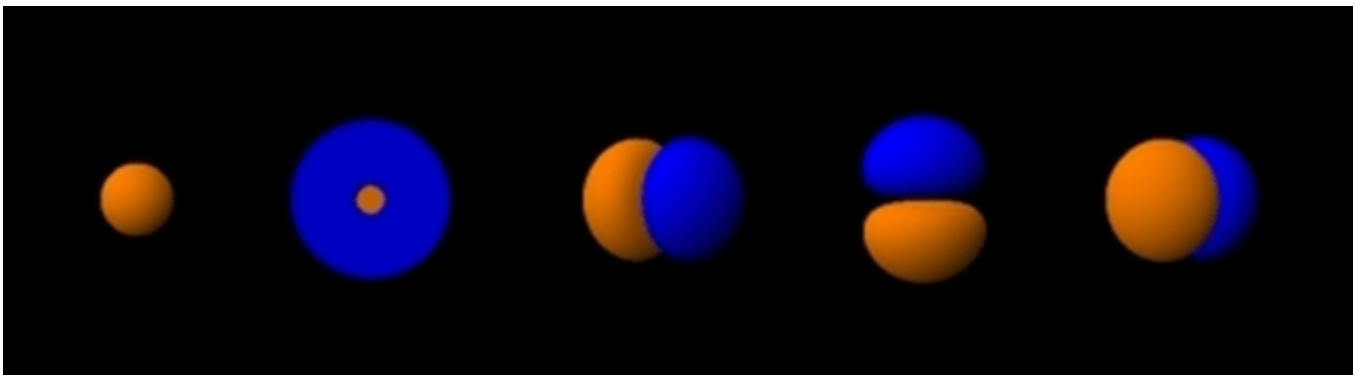


Figure 3.2.5.4: Bohr model of the atom generalized (left), and electron energy levels of a hydrogen atom with energy values and spectral emission series. Source: Brighterorange / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/), and https://www.hiclipart.com/free-trans...cnxhb/download

- When an electron jumps from a lower shell to a higher shell, it absorbs electromagnetic radiation of energy equal to the energy gap between the initial and the final shell.
- When an electron jumps from a higher shell to a lower shell, it emits radiation equal to the energy gap between the initial and the final shell.

Fig. 2.5.3 illustrates the emission of radiation from atoms –it is a line spectrum because only discrete energy levels, called shells, are allowed to electrons in an atom.



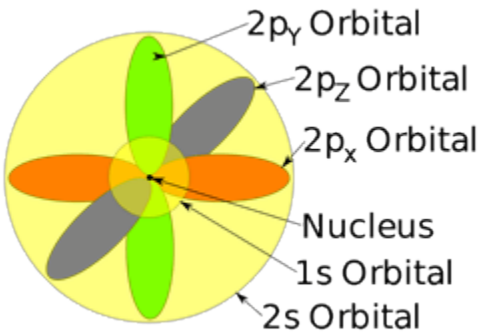


Figure 3.2.5.5: The shapes of the first five atomic orbitals are: (from left to right) 1s, 2s, $2p_x$, $2p_y$, $2p_z$, and of the five togather with the nucleus at the center. Source: Neon_orbitals.JPG, public domain, and John Trombley / CC BY (https://creativecommons.org/licenses/by/4.0)

The subshell

A second quantum number, called Azimuthal quantum number (l) defines subshells within a shell.

The subshells are usually designated as s, p, d, f, ...

Each shell has subshells equal to the shell number. For example, 1st shall have only one subshell, i.e., s. It is designated 1s, where the number is the principal quantum number, and the letter is the subshell. The 2nd shell has two subshells 2s and 2p; the 3rd shell has three subshells 3s, 3p, and 3d; and the 4th shell has four subshells 4s, 4p, 4d, and 4f.

The energy order of subshells is 1s<2s<2p<3s<3p<4s and so on.

Fig. 2.5.6 helps in remembering the order. This figure is drawn by placing the orbitals in columns and shell numbers in rows in increasing order of n from top to bottom, starting from 1s orbitals in the first column and first row, p orbitals in the second, d in the third, and f in the fourth. The filling of electrons follows arrows going from corner to corner, starting from the top left corner of the topmost cell, as described in more detail in a later section.

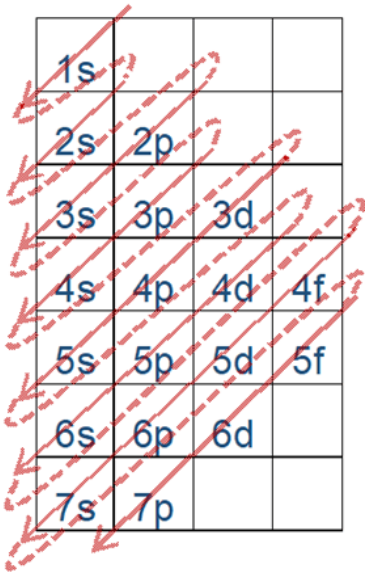


Figure 3.2.5.6: Order of filling of electrons in subshells. Start from the top and follow the arrows.

Orbital

The orbital is the region in space around the nucleus of an atom where electrons are most likely found.

Each subshell has a certain number of orbitals in them. The orbitals have a specific shape and orientation. The s subshell has only one orbital spherically symmetrical, like a ball with a nucleus at the center. The 1s orbital is smaller than 2s, and 2s is smaller than 3s, but they all have a spherically symmetrical shape, as illustrated in Fig. 2.5.5. The p subshell has three orbitals. Each p orbital is a dumbbell shape with two lobes, i.e., p_x oriented along the x-axis, p_y along the y-axis, and p_z along the z-axis, as illustrated in Fig. 2.5.5.

Degenerate orbitals

A set of orbitals having the same energy is called degenerate orbitals.

All the three p orbitals in the same shell have the same energy, i,e, $2p_x$, $2p_y$, and $2p_z$ is a set of degenerate orbitals. The d subshell has five degenerate orbitals, and f subshell has seven degenerate orbitals, as shown in Fig. 2.5.7. Their shapes and orientations are more complex and not shown here.

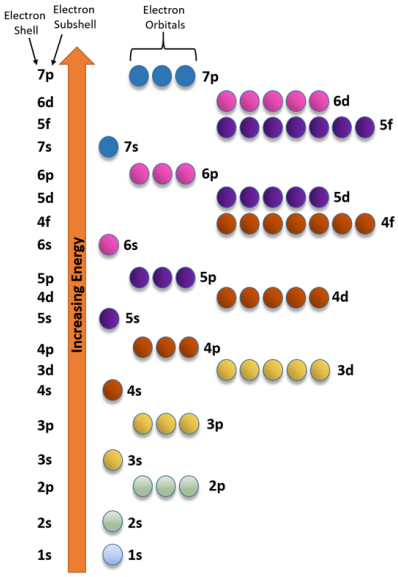


Figure 3.2.5.7: Relative energies of shells, subshells, and orbitals in multi-electron atoms. Source: https://www.hiclipart.com/free-trans...yfbwx/download

The electron configuration of atoms

The electron configuration is the distribution of electrons in the atom's orbitals.

Rules for distributing electrons in atom's orbitals

1. Each orbital can take a maximum of two electrons.
2. The orbitals are filled in the order of increasing energy: the lowest energy orbital is filled with electrons before the next higher energy orbital starts filling up.

The energy order of orbitals is 1s<2s<2p<3s<3p<4s and so on, as shown in Fig. 2.5.6 and Fig. 2.5.7.

The electron configuration of 1st-row elements

The hydrogen atom has one electron, and it occupies 1s orbital, i.e., $1s^1$ represents the electron configuration of the hydrogen atom. The superscript in $1s^1$shows the number of electrons in the sub-shell. A helium atom has two electrons, and both occupy 1s orbital, giving helium the electron configuration of $1s^2$. Hydrogen and helium are in the 1st-row of the periodic table and have the 1st-shell containing electrons.

The electron configuration of 2nd-row elements

The 2nd row starts from lithium with 3 electrons: the first two occupy 1s, and the third occupy 2s giving lithium the electron configuration $1s^2$ $2s^1$. The next element is beryllium with 4 electrons with the configuration $1s^2$ $2s^2$. The next element is a boron

with 5 electrons and the electron configuration $1s^2\ 2s^2\ 2p^1$. Carbon has 6 electrons with the electron configuration $1s^2\ 2s^2\ 2p^2$. Remember that the s subshell has one orbital and can take a maximum of two electrons, but the p subshell has three orbitals and can take a maximum of six electrons, i.e., two per orbital. Nitrogen, oxygen, fluorine and neon have 7, 8, 9, and 10 electrons and have the electron configuration $1s^2 2s^2 2p^3$, $1s^2 2s^2 2p^4$, $1s^2 2s^2 2p^5$, and $1s^2 2s^2 2p^6$, respectively. Atomic number 3 to 10, i.e., lithium to neon, completes the 2nd-row.

Valence electrons and core electrons

The outermost shell is called the valence shell, and electrons in the valence shell are called valence electrons. 1st shell is the valence shell for 1st-row elements hydrogen and helium. 2nd-row elements have an inner shell with configuration $1s^2$, and a valence shell containing 2s and 2p being filled. The inner shell is also called the core-shell, and the electrons in the core-shell are called the core electrons.

The electron configuration of 3rd-row and 4th-row elements

The 3rd-row starts with sodium atomic number 11 and ends with argon atomic number 18. The electron configuration of the 3rd-row elements repeats the pattern of the 2nd-row, i.e., a set of full core-shells $1s^2 2s^2 2p^6$, and a valence shell having 3s or 3p being filled. The first two elements of the 4th-row are potassium and calcium with atomic numbers 19 and 20 having configurations $1s^2 2s^2 2p^6 3s^2 3p^6 4s^1$ and $1s^2 2s^2 2p^6 3s^2 3p^6 4s^2$, respectively. The electron configuration of elements with atomic numbers beyond 20 is more complicated, involving d and f subshell, and they are out of the scope of this book. Fig. 2.5.8 shows the electron configuration of the first twenty elements described above.

Group→ Row↓	1	2	13	14	15	16	17	18
1	1 H $1s^1$							2 He $1s^2$
2	3 Li $1s^2 2s^1$	4 Be $1s^2 2s^2$	5 B $1s^2 2s^2 2P^1$	6 C $1s^2 2s^2 2P^2$	7 N $1s^2 2s^2 2P^3$	8 O $1s^2 2s^2 2P^4$	9 F $1s^2 2s^2 2P^5$	10 Ne $1s^2 2s^2 2P^6$
3	11 Na $1s^2 2s^2 2P^6$ $3s^1$	12 Mg $1s^2 2s^2 2P^6$ $3s^2$	13 Al $1s^2 2s^2 2P^6$ $3s^2 3p^1$	14 Si $1s^2 2s^2 2P^6$ $3s^2 3p^2$	15 O $1s^2 2s^2 2P^6$ $3s^2 3p^3$	16 S $1s^2 2s^2 2P^6$ $3s^2 3p^4$	17 Cl $1s^2 2s^2 2P^6$ $3s^2 3p^5$	18 Ar $1s^2 2s^2 2P^6$ $3s^2 3p^6$
4	19 K $1s^2 2s^2 2P^6$ $3s^2 3p^6 4s^1$	20 Ca $1s^2 2s^2 2P^6$ $3s^2 3p^6 4s^2$						

Figure 3.2.5.8: The electron configuration of the first twenty elements in the periodic table of elements. Valence electrons are shown in red color fonts.

This page titled 3.2.5: Electrons in atoms is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.2.6: The periodic trends in properties of the elements

Properties of elements generally show a periodic trend that correlates with their position in the periodic table. The properties and their periodic trends are described below.

Valence electrons

Electrons in the outermost shell are the valence electrons. Fig. 2.5.8 shows the valence electrons of the first twenty elements in red fonts.

Periodic trend of valence electrons

All elements in a group have the same number of valence electrons equal to the first digit of their group number. For example, 1st group of hydrogen and alkali metals has one valence electron, 2nd group of alkali metals has two valence electrons, halogens in 17th group have seven valence electrons, and noble gases in 18th group have eight valence electrons.

Caution

The transition metals in groups 3 to 12, and inner-transition metals, i.e., lanthanoids and actinoids -the two rows of elements placed below the periodic table, are the exception to the general trend of valence electrons described above. Valence electron configurations of transition metals and inner-transition metals are not described here; it is beyond the scope of this book.

The valence electrons mainly determine the chemical properties of the elements. The elements in the same group have similar chemical properties because they have the same valence shell electron configuration. The elements in a row show a gradual change in chemical properties because their valence shell electron configuration changes gradually along the row.

Lewis symbols

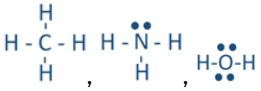
1. Lewis symbols show the valence electrons as dots around the symbol of an element. One dot represents one valence electron, e.g., $\mathrm{H}\bullet$.
2. The dots are shown on any of the four sides of the symbol.
3. A single dot on the top, bottom, left, or right is shown four valence electrons. Then start pairing the dots beyond four valence electrons, as shown in Fig. 2.6.1 for the first twenty elements.
4. Helium is an exception that has only two valence electrons, but they are shown paired.

Group→ Row↓	1	2	13	14	15	16	17	18
1	H•							He:
2	Li•	•Be•	•B•	•C•	•N•	•O•	:F•	:Ne:
3	Na•	•Mg•	•Al•	•Si•	•P•	•S•	:Cl•	:Ar:
4	K•	•Ca•						

Figure 3.2.6.1: Lewis symbols or electron-dot symbols of the first twenty elements in the periodic table.

The electron dots in the Lewis structure are a convenient way to determine how many bonds an atom of an element can make.

> Generally, each unpaired dot can make one bond.

For example, a hydrogen atom with one unaired dot can make one bond as in H-H. A bond is represented by a line between the bonded atoms. A bond is formed by sharing unpaired valence electrons. It is called a covalent bond. Carbon, nitrogen, oxygen, and fluorine with 4, 3, 2, and 1 unpaired dot can make 4, 3, 2, and 1 bond, e.g., in the following molecules:  , , , and  .

> - Each line in these molecules represents a **bonding electron pair**, and
> - the pair of dots represent valence electrons that are not involved in bonding, called **lone pair of electrons**.

Atomic size

Electrons exist around the nucleus in a cloud-like appearance with no clearly defined boundaries. Therefore, the atomic size generally refers to the **covalent radius** of an atom that is one-half of the distance between the nuclei covalently bonded in a homonuclear molecule, like Cl_2, I_2, H_2, as illustrated in Fig. 2.6.2.

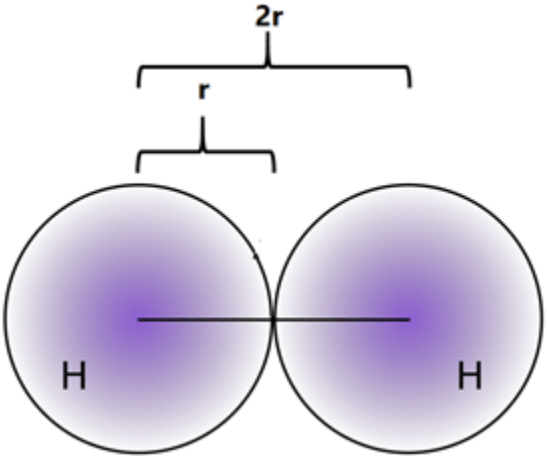


Figure 3.2.6.2: Diagram showing the atomic radius of H_2. Source: modified from CK-12 Foundation / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Periodic trend of atomic sizes

1. The atomic size generally increases from top to bottom in a group because the valence electrons add in the higher shell in each consecutive member of a group down the column.
2. The atomic size generally decreases from left to right in a row because the valence electrons are in the same shell while more protons add to the nucleus, increasing the pull on the valence electrons, as illustrated in Fig. 2.6.3.

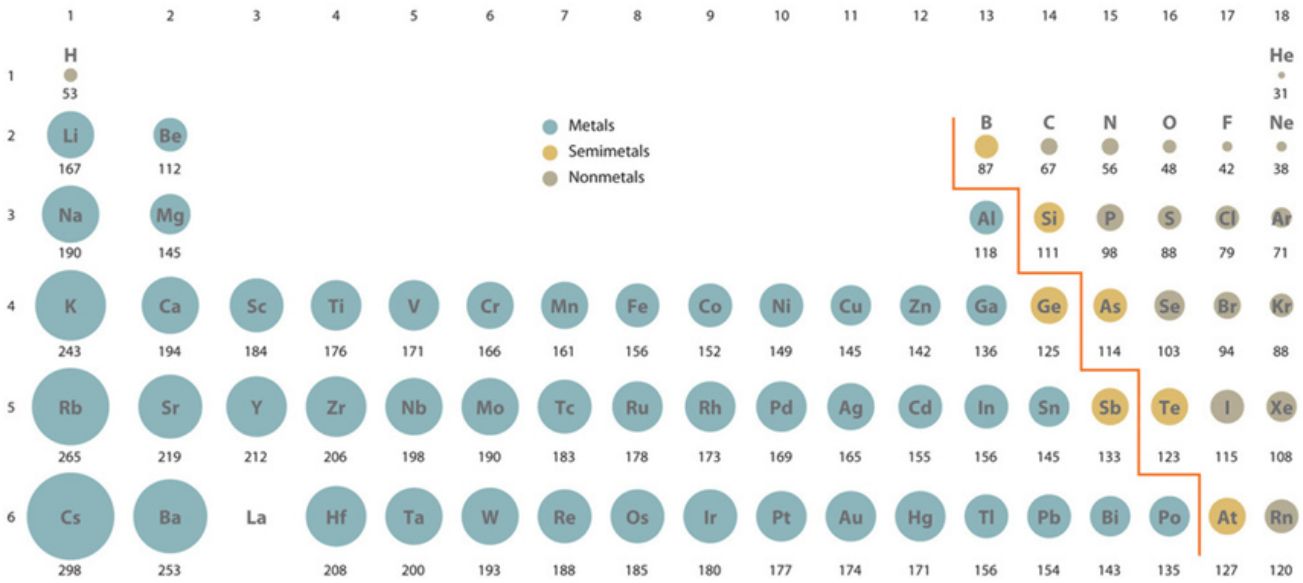


Figure 3.2.6.3: Illustration of the periodic trend of atomic sizes, Calculated Atomic Radii (in Picometers). Source: https://chem.libretexts.org/@api/dek...jpg?revision=1

Ionization energy

The positively charged nucleus attracts the negatively charged electrons. Therefore removal of an electron from the atom requires energy. The ionization produces a cation with fewer electrons than the parent neutral atom, i.e., cations.

Ionization energy

Ionization energy is the energy needed to remove an electron from a neutral atom, as in the following reaction.

$$\mathrm{Na} + \text{ ionization energy } \rightarrow \mathrm{Na}^{1+} + \mathrm{e}^{-}$$

Periodic trend in ionization energy

1. The ionization energy generally decreases from top to bottom in a column because the valence electrons are further away and experience less pull to the nucleus down the column.
2. The ionization energy generally increases from left to right in a row because the valence electrons are in the same shell while more protons add to the nucleus, which increases the pull on the valence electrons.

Fig. 2.6.4 illustrates the periodic trend in the ionization energy.

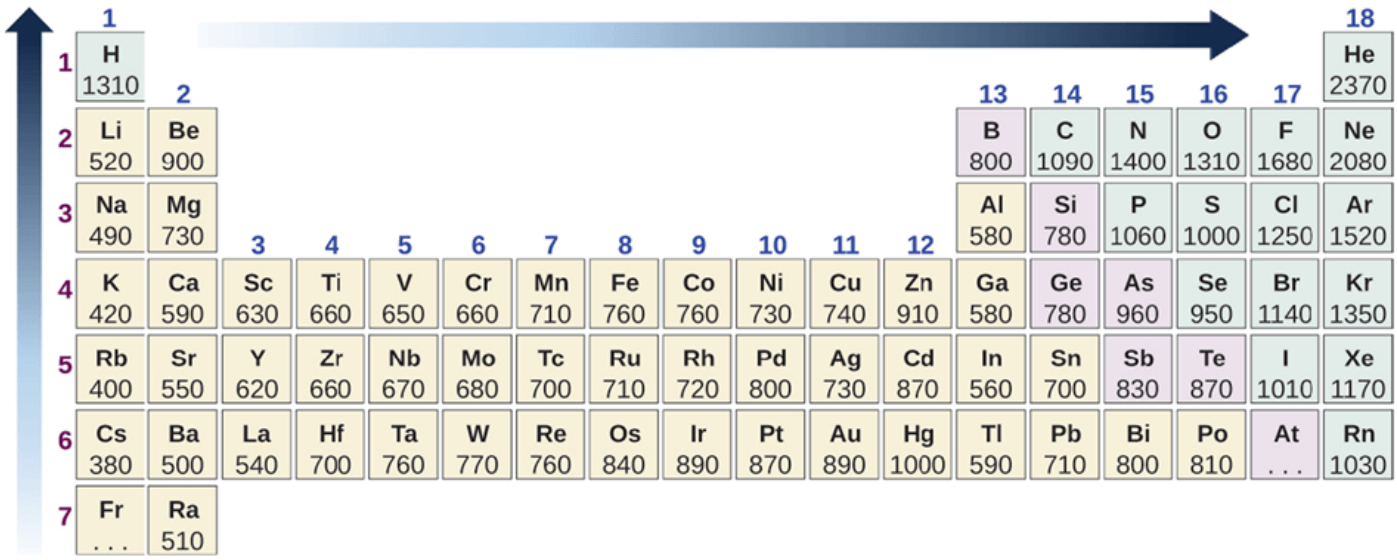


Figure 3.2.6.4: The 1st ionization energy in kJ/mol. Source: Download for free at https://openstax.org/details/books/chemistry

Electronegativity

Definition: Electronegativity

Electronegativity is the ability of an atom in a compound to attract the bonding electron pair to itself. Electropositivity is the opposite of electronegativity. Electronegativity is a property of an atom in a compound, i.e., a bonded atom, not a feature of an individual atom.

There are several electronegativity scales. The most commonly used is the Pauling electronegativity scale. Fig. 2.6.5 shows the electronegativities values on Pauling's electronegativity scale.



Figure 3.2.6.5: Periodic table of elements by Pauling scale. Source: https://en.Wikipedia.org/wiki/Electronegativity

Periodic trend in electronegativity

1. The electronegativity generally decreases from top to bottom in a column because the atomic size increases down the column making the nucleus less effective in pulling the bonding electrons.
2. The electronegative generally increases from left to right in a row because the atomic size decrease from left to right in a row, making the nucleus more effective in pulling the bonding electron pair to itself.

Metallic character

The metallic character relates to the ease of losing an electron in a chemical reaction. The metallic character trend is opposite to the trend of ionization energy.

Periodic trend in metallic character

1. The metallic character generally increases from top to bottom in a column because the atomic size increases down the column, making the valence electrons less tightly held and easier to remove.
2. The metallic character generally decreases from left to right in a row because the atomic size decreases from left to right, making the valence electrons more tightly held and difficult to remove.

Summary of the periodic trends

The ionization energy and the electronegativity generally increase from left to right in a row and from bottom to top in a column. The atomic size and the metallic character are opposite, i.e., they increase from right to left in a row and from top to bottom in a column. Fig. 2.6.6 summarizes the periodic trend in the properties of the elements.

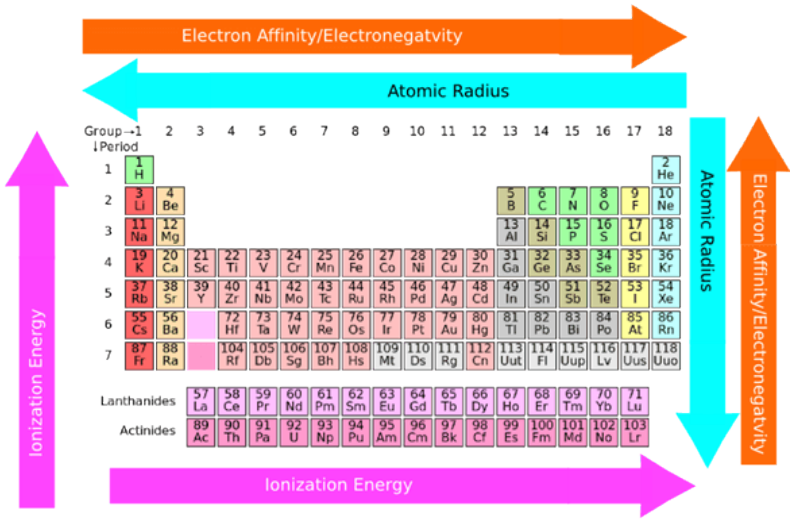


Figure 3.2.6.6: Summary of the periodic trend in the properties of elements. Source: https://www.hiclipart.com/free-trans...ntzga/download

This page titled 3.2.6: The periodic trends in properties of the elements is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 3.3: COMPOUNDS

3.3.1: Bonding in compounds

3.3.2: Naming binary ionic compounds

3.3.3: Polyatomic ions and their compounds

3.3.4: Naming acids

3.3.5: Naming binary covalent compounds

3.3.6: Lewis structures of molecules

3.3.7: Molecular shapes –Valence shell electron pair repulsion (VSEPR) theory

3.3.8: Polarity of molecules

3.3.9: Intramolecular forces and intermolecular forces

This page titled Chapter 3.3: Compounds is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.3.1: Bonding in compounds

Chemical bonds

What is a compound?

Compounds are a pure form of matter formed by atoms of more than one element combined in a constant whole number ratio.

The bonds connect the atoms in the compounds. Sharing or transferring some valance electrons from one atom to the other makes the bonds. Noble gases have a full valence shell of eight valence electrons, except helium which has a full valence shell of two valence electrons. The noble gases are the least reactive, i.e., the most inert group of elements.

Octet rule

The octet rule states that atoms of all elements other than noble gases tend to share, lose, or gain valence electrons to acquire the electron configuration of the nearest noble gas having eight valence electrons.

Covalent bonds

A bond formed by sharing electrons is a covalent bond.

When a nonmetal atom combines with another nonmetal atom, they usually make a covalent bond. A covalent bond is a pair of shared electrons, called a **bonding pair** of electrons, where each bonded atom contributes one electron. For example, chlorine has seven valence electrons and needs one more to complete its octet. Hydrogen has one valence electron and requires one more to acquire the electron configuration of helium, i.e., **duet** instead of the octet. Hydrogen and chlorine combine by sharing one electron to make the compound HCl. Similarly, oxygen, nitrogen, and carbon make 2, 3, and 4 covalent bonds with hydrogen to complete their octet and make compounds H_2O, NH_3, and CH_4, respectively, as illustrated in Fig. 3.1.1.

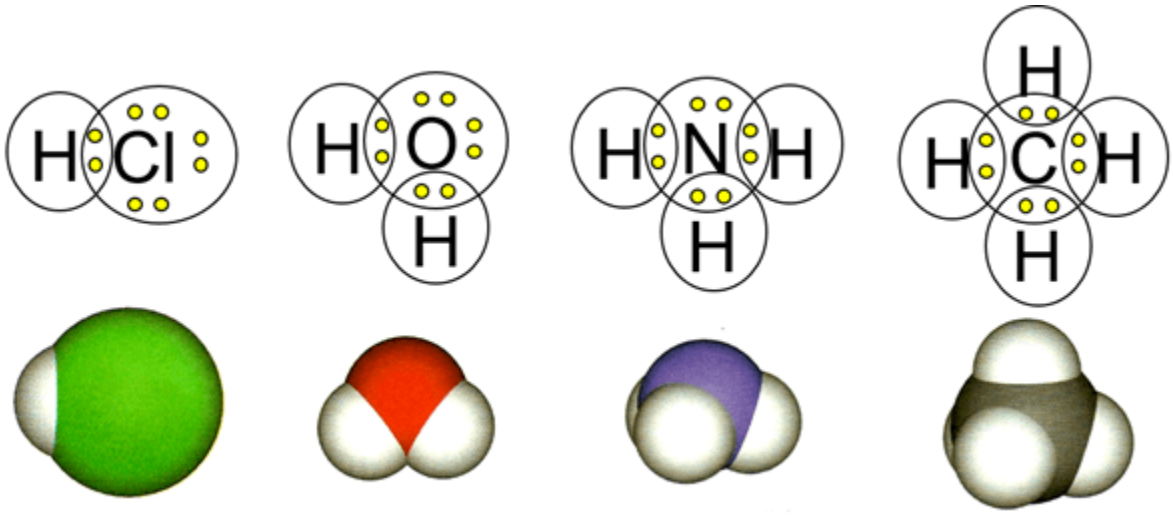


Figure 3.3.1.1: Examples of covalent bonds where Cl, O, N, and C acquire octet by sharing 1, 2, 3, and 4 electrons, respectively. Hydrogen shares one electron and acquires a duet, i.e., the electron configuration of helium. 2nd row is the space-filling model of the molecule where the white sphere is hydrogen, and the colored is the other atom. Source: "File:Kovalentne veze.png" by Drago Karlo CC BY-SA 4.0

Two atoms can share one, two, or three electrons to make a single, a double, or a triple covalent bond. For example, H_2 has a single bond (H-H), O_2 has a double bond (O=O), and N_2 has a triple bond (N≡N), where each line between the atoms represent one covalent bond. Fig. 3.1.2 illustrates the formation of three covalent bonds in N_2. A valence electron pair that is not involved in bonding is called a **nonbonding pair**. One nonbonding pair and three bonding pairs complete the octet of each nitrogen atom in the N_2 molecule, as shown in Fig. 3.1.2.

Note

The bonding pair of electrons counts towards the total valence electrons of each bonded atom., i.e., in H-H each hydrogen atom has two valence electrons, and in **:**N≡N**:** each nitrogen has eight valence electrons; two in the nonbonding pair and six in three bonding pairs.

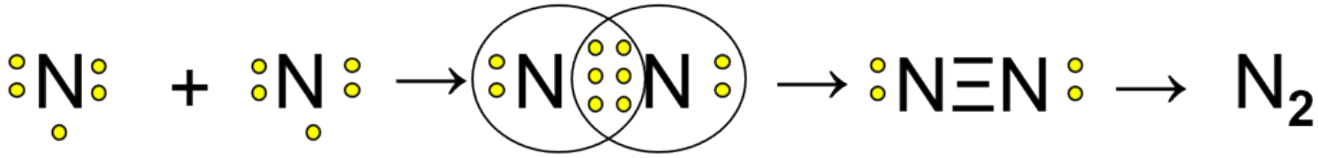


Figure 3.3.1.2: Illustration of three covalent bond formations between two nitrogen atoms. Source: File:Kovakentna veza dušika.png" by Drago Karlo is licensed under CC BY-SA 4.0

The formula of a covalent compound

A compound is represented by a chemical formula that combines the symbols of its constituent elements. More electropositive elements are usually written first, e.g., HF, NO, CO. Some exceptions to this rule, e.g., NH_3 and CH_4, have more electronegative elements written first. The majority of the covalent compounds exist as discrete molecules. A subscript to the right of the element's symbol represents the number of atoms of the component of the molecule. For example, H_2O has two hydrogen atoms and one oxygen atom in a water molecule. Note that subscript 1 is not written, i.e., the symbol of an element alone represents one atom.

A few covalently bonded compounds are giant molecules where the atoms are held together by a 3D network of bonds. The formula of these compounds shows the simplest whole-number ratio of elements in the compound. For example, Fig. 3.1.3 shows SiO_2 present in high-purity sand and quartz, a giant molecule.



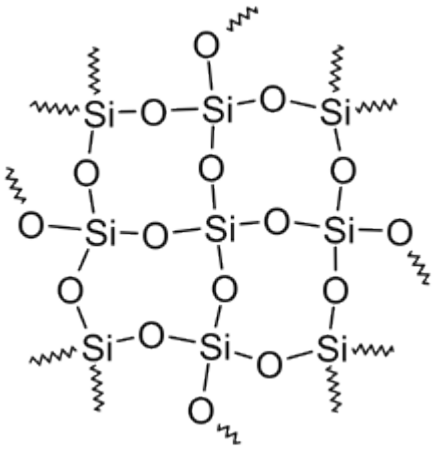


Figure 3.3.1.3: Quartz crystal composed of SiO_2, and an illustration of covalent bonds in a giant molecule of SiO_2. Source: Parent Géry, Public domain, via Wikimedia Commons, and Roland Mattern / Public domain

Ionic bond

A bond formed by the transfer of electrons from one atom to the other atoms is an ionic bond.

A compound that has ionic bonds is an ionic compound. Usually, metal atoms lose electrons and become cations, and nonmetal atoms gain electrons to become anions. The electrostatic attraction between the opposite charges holds the ions together in the ionic compound. For example, sodium (Na) loses one electron, and fluorine (F) gains one electron to make a compound sodium fluoride (NaF), as illustrated in Fig. 3.1.4.

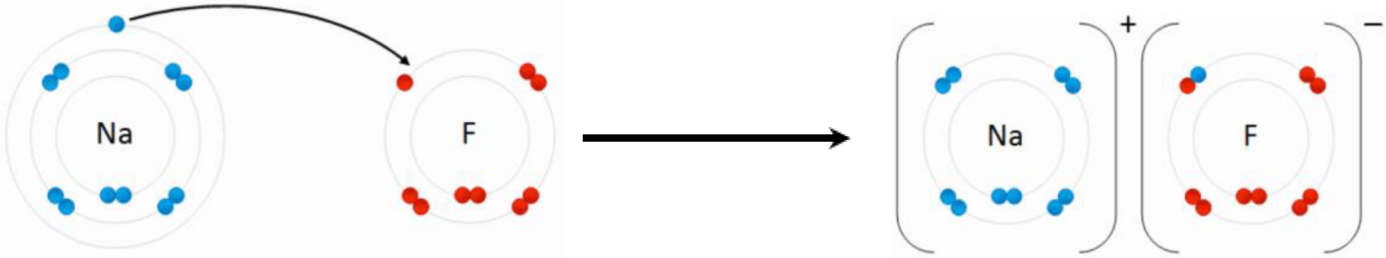


Figure 3.3.1.4: An example of ionic bond formation between Na and F. Source: modified from: Wdcf / CC BY-SA.

Ionic compounds

Table salt, i.e., NaCl is an example of an ionic compound. The Na completes its octet by losing one electron and becoming Na^+ cation. Losing electrons reduces the electron-electron repulsion, but the electron-nucleus attraction remains the same. Consequently, the electron cloud around the nucleus shrinks. Similarly, the chlorine atom has seven valence electrons. After gaining one electron, it becomes Cl^- anion with its octet complete. Gaining electrons increases the electron-electron repulsion, but the electron-nucleus attraction remains the same. Consequently, the electron cloud around the nucleus expands. Fig. 3.1.5 illustrates the formation of an ionic bond and the accompanying changes in the total number of electrons and sizes relative to the parent neutral atoms in the case of NaCl formation.

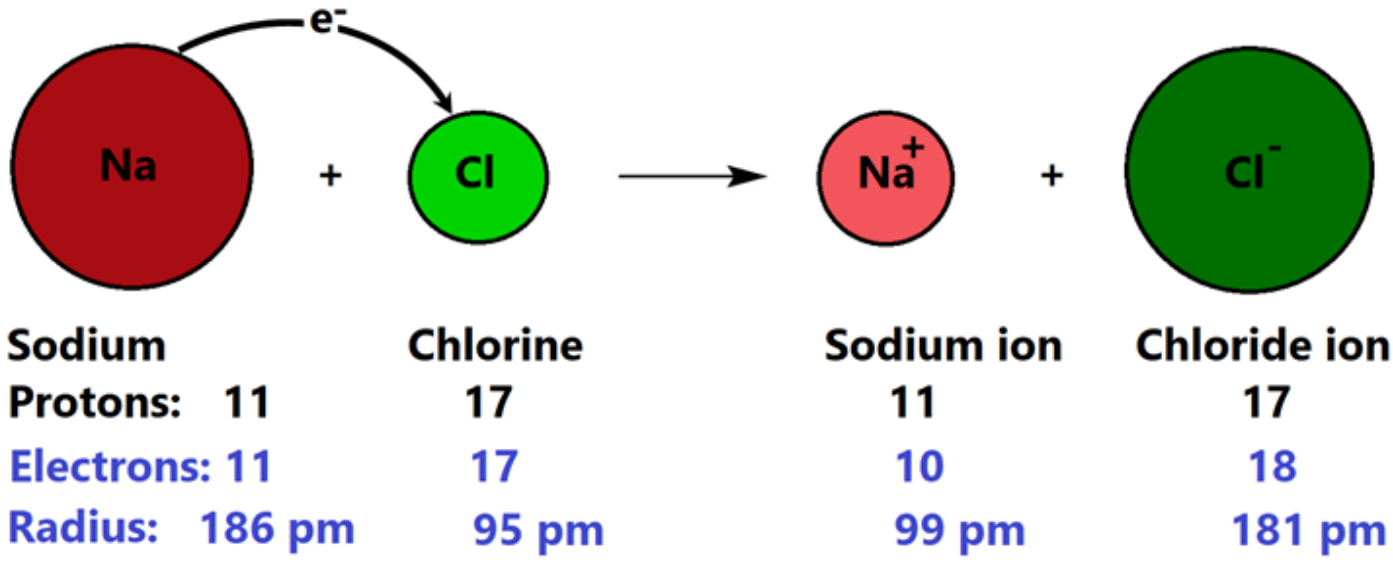


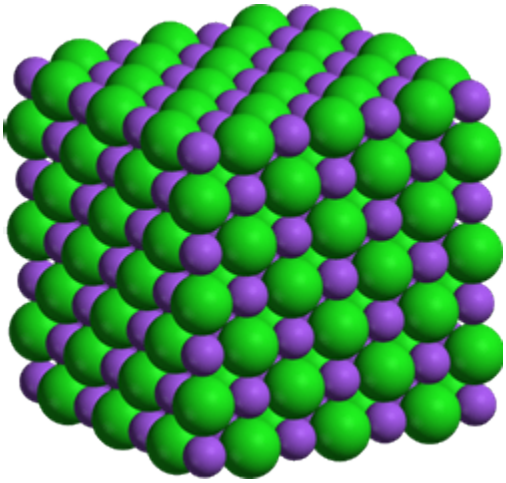
Figure 3.3.1.5: Illustration of an ionic compound formation by electron transfer from sodium (Na) to chlorine (Cl) making Na^+ cation with one less electron and Cl^- anion with one more electron than their parent neutral atoms. The Na^+ cation is smaller than Na, and the Cl^- anion is large than Cl atom.

The ionic bond is not localized or unidirectional. The electrostatic force is all around the ions. Therefore, the cations surround the anions, and the anion surrounds the cations in a regular array in a 3D crystal lattice. Fig 3.1.6 illustrates the structure of NaCl.

Note

1. The formula of the ionic compounds represents the simplest whole-number ratio of the atoms of the constituent elements.
2. A cation is always smaller in size than its parent neutral atom.
3. An anion is always larger than its parent neutral atom.





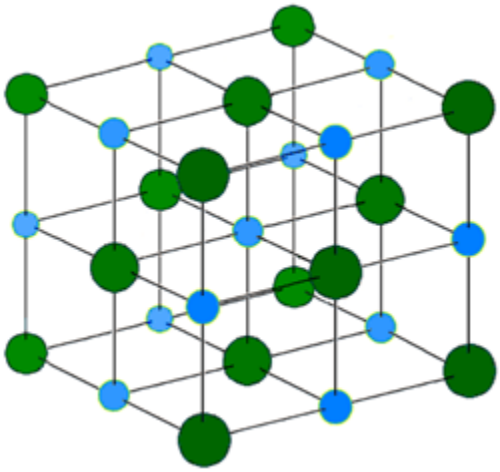


Figure 3.3.1.6: Sodium Chloride from left: actual crystals in a water bubble within a 50 mm metal loop, model, and crystal lattice. Na^+ is purple or blue, and Cl^- is green in the models. Source: Photograph by the NASA Expedition 6 crew, Public domain, via Wikimedia Commons, Benjah-bmm27 (talk · contribs) / Public domain, and File: H Padleckas / Public domain

Criteria to distinguish between ionic and covalent bond

Usually, a metal and a nonmetal bond is ionic, and the bond between two nonmetals is covalent. Better criteria are based on the difference in electronegativities of the bonded atoms. Electronegativity is the ability of an atom to attract a pair of bonded electrons to itself. If the electronegativity difference is significant, the bonding electrons completely transfer to the more electronegative atom, and the bond is ionic. There is no single value of electronegativity difference to separate ionic and covalent bonds, but usually, the electronegativity difference of more than 1.8 results in an ionic bond. Otherwise, a covalent bond, but the bonding electrons are more towards the more electronegative atom, making it a polar covalent bond. An electronegativity difference less than 0.5 is considered a noncovalent bond, but a true noncovalent bond forms when the bonded atoms are the same element.

Properties of compounds

The properties of the compounds are usually altogether different from the properties of their constituent elements. For example, hydrogen (H_2) is a gas that burns in oxygen, oxygen (O_2) is a gas that assists combustion, but water (H_2O) is a liquid that extinguishes fire. Similarly, sodium (Na) is a soft metal that melts at 97.79 ^oC, chlorine (Cl_2) is yellowish color gas, but sodium chloride (NaCl) is a transparent crystal that melts at 801 ^oC.

The intermolecular interactions in covalent molecules are weak to moderate relative to the strength of covalent bonds or ionic bonds. Therefore, the covalent molecules are usually gases like O_2, NH_3, CH_4, liquids like H_2O, or soft and low melting solids like waxes, glucose ($C_6H_{12}O_6$, mp 146^oC).

Ions in the ionic compounds are held together by strong ionic bonds in a 3D array of crystal lattices. Therefore, ionic compounds are usually hard solids with high melting points. For example, NaCl melts at 801 ^oC. Covalent compounds that exist as a 3D network of covalent bonds, i.e., as giant molecules, are usually hard materials having a higher melting point than ionic compounds. For example, SiO_2 present in sand and quartz is a hard solid that melts at 1,710 ^oC. Diamond –the hardest known substance, is a giant molecule of carbon atoms held together in a 3D network of covalent bonds that melts around 4,027 ^oC.

This page titled 3.3.1: Bonding in compounds is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.3.2: Naming binary ionic compounds

Binary ionic compounds are compounds composed of monoatomic cations and monoatomic anions. For example, NaCl is a binary ionic compound composed of monoatomic cations Na^+ and monoatomic anions Cl^-. Another example is $CaCl_2$ composed of monoatomic cations Ca^{2+} and monoatomic anions Cl^-.

Charge on monoatomic ions

Nonmetals and metalloids of group 14 to group 17 usually form monoatomic anions. The charge on the anions is equal to the group number minus eighteen. For example, halogens in group 17 have charge: 17-18 = -1, oxygen in group 16 has charge: 16-18 = -2, and nitrogen in group 15 has charge: 15-18 = -3. Metals usually form cations: metals of group 1 from +1, metals of group 2 form +2, and aluminum of group 13 form +3 charge on cations, as shown in Fig. 3.2.1. Other metals have variable charges in compounds. The charge of the metals having variable charge can be calculated from the compound's chemical formula because the total –ve charge should be equal to the total +ve charge to make the compound neutral.

Example 3.3.2.1

calculate the charge of iron ion in $FeCl_2$?

Solution

Three are two chloride anions, each with a -1 charge, making a total of -2. So the charge on cation has to be +2 to balance the negative charge. Answer: Fe^{2+}.

Example 3.3.2.2

Calculate the charge on an iron ion in Fe_2O_3?

Solution

There are three oxygen anions, each with a -2 charge, making -6. So the total charge on two iron atoms should be +6, i.e., the charge on iron atoms is +3. Answer: Fe^{3+}.

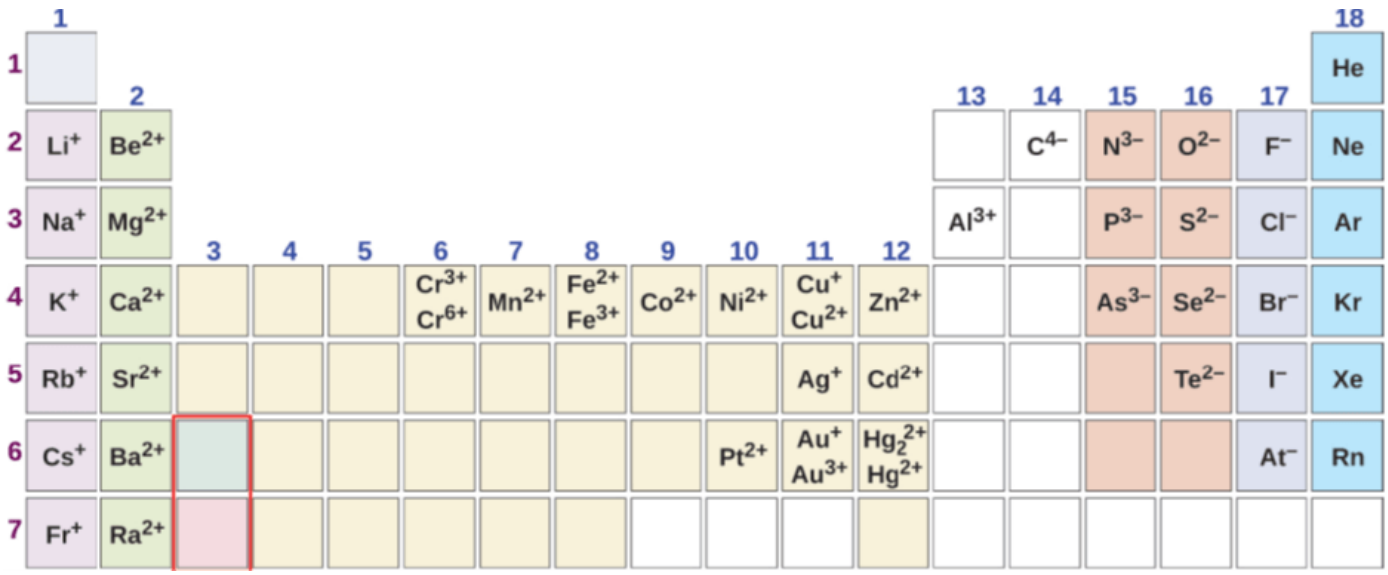


Figure 3.3.2.1: The charges on common ions. Download for free at https://openstax.org/details/books/chemistry.

Names of monoatomic ions

Name of a monoatomic anion

The name of a monoatomic anion is the element's name with the last syllable replaced with –ide ion. For example, Cl^- is a chloride ion, O^{2-} is an oxide ion, N^{3-} is a nitride ion, S^{2-} is sulfide ion, and C^{4-} is a carbide ion, derived from the element names chlorine, oxygen, nitrogen, sulfur, and carbon, respectively.

Name of a monoatomic cation having a fixed charge

Alkali metals have +1, alkaline earth metals have +2, and aluminum has +3 charge. Their name is the name of the element ending with ion. For example, Na^+ is a sodium ion, Ca^{2+} is a calcium ion, and Al^{3+} is an aluminum ion.

Name of a monoatomic cation having a variable charge

The names of the cations that have a variable charge is the name of the element followed by charge in roman numeral enclosed in small bracts and ending with ion. For example, Fe^{2+} is an iron(II) ion, and Fe^{3+} is an iron(III) ion, Cu^+ is a copper(I) ion, and Cu^{2+} is a copper(II) ion.

Ions in body fluids

The ions that are important in body fluids include sodium ion (Na^+), potassium ion (K^+), calcium ion (Ca^{2+}), magnesium ion (Mg^{2+}), and chloride ion (Cl^-), as shown in Fig. 3.2.2. Na^+ is present in fluids inside the cells. It regulates and controls body fluids. K^+ is present in fluids outside the cells and regulates body fluids and cell functions. Ca^{2+} and Ma^{2+} are present in the body fluids outside the cells, where Ca^{2+} is needed for muscle contraction, Mg^{2+} is needed for muscle contraction, nerve control, and enzymes. Cl^- is primarily present to balance the charge of the cations in the body fluids.

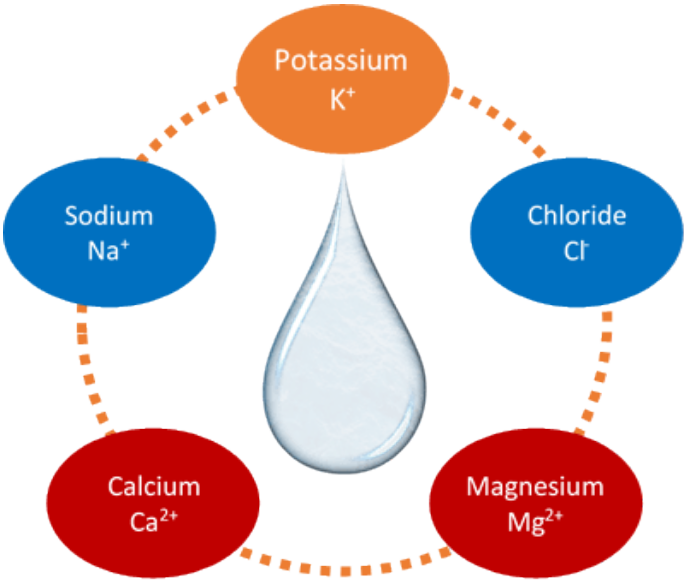


Figure 3.3.2.2: Electrolytes usually found in body fluids.

Writing the formula of a binary ionic compounds

The formula of an ionic compound is the symbol of the cation element with a subscript number followed by the symbol of the anion element with a subscript number. The formula shows the simplest whole-number ratio of the constituent elements in the subscripts, such that the total positive charge is equal to the total negative charge.script on the right of the symbols tell the number if they're more than one atom of the element.

Rules of writing the formula of a binary ionic compound

The rules are illustrated in Fig. 3.2.3:

1. Write cation followed by anion with charges,
2. swap the charge as a subscript of the opposite ion,
3. simplify the subscript to the simplest whole-number ratio,
4. use the simplified subscript in the final formula, and
5. do not write the subscript if it is one.

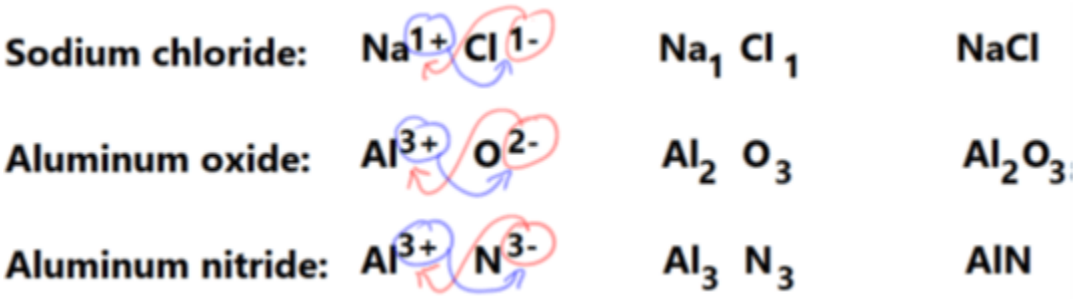


Figure 3.3.2.3: Rules to write the formula of the binary ionic compounds.

Writing the names of an ionic compound from the formula

Writing the name of ionic compounds of cations with fixed charge

If the cation has a fixed charge in compounds, the name starts with the name of the element of the cation, followed by the name of the anion without the world ion at the end. For example, KI is potassium iodide, and $CaCl_2$ is calcium chloride.

Writing the name of ionic compounds of cations with variable charge

Write the name of the cation, including the charge in roman numerals enclosed in small brackets but without the word ion at the end, followed by the name of the anion, without ion at the end. For example, $FeCl_2$ is iron(II) chloride, and Fe_2O_3 is iron(III) oxide. Additional examples are given in Table 1.

Transition metals with fixed charge

Silver, zinc, and cadmium cations have fixed changes: Ag^+, Zn^{2+}, and Cd^{2+}. Names of these cations are the names of the element with or without charge shown in roman numerals, both ways it is correct.

Table 1: Examples of writing names from the formulae of binary ionic compounds

Example#	Formula	Name of the cation	Name of the anion	Name of the compound
1	NaCl	Sodium ion	Chloride ion	Sodium chloride
2	Al_2O_3	Aluminum ion	Oxide ion	Aluminum oxide

Example#	Formula	Name of the cation	Name of the anion	Name of the compound
3	$FeCl_3$	Iron(III) ion	Chloride ion	Iron(III) chloride
4	CuO	Copper(II) ion	Oxide ion	Copper(II) oxide
5	AgCl	Silver(I) ion, or Silver ion	Chloride ion	Silver(I) chloride, or Silver chloride

This page titled 3.3.2: Naming binary ionic compounds is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.3.3: Polyatomic ions and their compounds

Polyatomic ions

Polyatomic ions are molecular ions composed of two or more atoms bonded by covalent bonds and acting as a single unit, but unlike molecules, they have a net charge on them.

The examples include cations like ammonium ion (NH_4^+), and hydronium ion (H_3O^+); and anions like hydroxide ion (OH^-), and cyanide ion (CN^-). Usually, the name of polyatomic cations ends with –ium, and the name of polyatomic anions end with –ide, except for oxyanions that have separate rules for their nomenclature.

Oxyanion

The oxyanions are oxides of nonmetals that are molecular ions. Examples include carbonate (CO_3^{2-}) , nitrate (NO_3^{2-}), phosphate (PO_4^{3-}), and sulfate (SO_4^{2-}). The following guidelines will help remember the names and charges of oxyanions in most cases.

Oxyanions having elements from groups 14, 15, and 16

1. the 2nd-row elements C, and N have three oxygen, and 3-row elements P, and S have four oxygen, i.e., one oxygen more than the row number,
2. the name of the polyatomic anion is the name of the atom other than oxygen with the last syllable replaced with –ate, and
3. the charge on these oxyanions equals the number of valence electrons on the non-oxygen atom minus twice the number of oxygen atoms. For example: in (PO_4^{3-}), phosphorous has 5 valence electrons and there are 4 oxygen atoms, so the charge = valence elections in the central atom -2 x number of oxygen atoms = 5-(2x4) = -3; in (SO_4^{2-}), sulfur has 6 valence electrons and there are 4 oxygen atoms, so the charge = 6-(2x4) = -2; in (CO_3^{2-}), carbon has 4 valence electrons and there are 3 oxygen atoms, so the charge = 4-(2x3) = -2; and in (NO_3^-), nitrogen 6 valence electrons and there are 3 oxygen atoms, so the charge = 6-(2x3) = -1.
4. There is a 2nd set of oxyanions of the elements mentioned above with one less oxygen but the same charge and the last syllable of the name changed from -ate to -ite. For example: (NO_3^-) is nitrate and (NO_2^-) is nitrite; (PO_4^{3-}) is phosphate and (PO_3^{3-}) is phosphite; and (SO_4^{2-}) sulfate and (SO_3^{2-}) is sulfite.

Oxyanions of halogens

Oxyanins of chlorine, bromine, and iodine are also common oxyanions with the following in common.

1. They have -1 charge,
2. a halogen with four oxygen is named by adding prefix "per-" to the name of the halogen with last syllable replaced with -ate, for example; (ClO_4^-) is perchlorate, (BrO_4^-) is perbromate, and (IO_4^-) is periodate.
3. a halogen with three oxygen is named as name of the halogen with last syllable replaced with -ate, for example; (ClO_3^-) is chlorate, (BrO_3^-) is bromate, and (IO_3^-) is iodate,
4. a halogen with two oxygen is named as name of the halogen with last syllable replaced with -ite; for example, (ClO_2^-) is chlorite,(BrO_2^-) is bromite, and (IO_2^-) is iodite,
5. a halogen with one oxygen is named by adding prefix "hypo-" to the name of the halogen with last syllable replaced with -ite; for example, (ClO^-) is hypochlorite, (BrO^-) is hypobromite, and (IO_3^-) is hypoiodaite.

Acids of oxyanions

Oxyanions are acids when their charge in neutralized with protons (H^+). Names of the acids are the names of oxyanions with -ate replaced with -ic acid and -ite replaced with -ous acid. For example: (HNO_3) is nitric acid and (HNO_2) is nitrous acid; (H_3PO_4) is phosphoric acid and (H_3PO_3) is phosphorous acid; and (H_2SO_4) sulfuric acid and (H_2SO_3) is sulfurus acid.

The prefixes "per-" and "hypo-" in the cases of oxyanions of halogens remain in the acid name. For example: ($HClO_4$) is perchloric acid; ($HClO_3$) is chloric acid; ($HClO_2$) is chlorous acid; and ($HClO$) is hypochlorous acid.

Oxyanions with one proton attached but charge one them not fully neutralized, i.e., they are still polyatomic anion are named beginning with hydrogen and ending with the name of the oxyanion. For example: (HSO_4^-) is hydrogen sulfate; (HSO_3^-) is hydrogen sulfite; and (HPO_4^{2-}) is hydrogen phosphate.

Oxyanions with two protons attached but charge one them not fully neutralized, i.e., they are still polyatomic anion are named beginning with dihydrogen and ending with the name of the oxyanion. For example, ($H_2PO_4^-$) is dihydrogen phosphate

fTwo oxyanions containing a transition metal as the central atom in common use as reagents are in chemistry are chromate (CrO_4^{2-}) and permanganate (MnO_4^-). Table 3.3.1 lists the formulas and names of some of the common polyatomic ions.

Table 1: Names of some of the common polyatomic ions

Formula	Name	Formula	Name
(NH_4^+)	Ammonium	(MnO_4^-)	Permanganate
(H_3O^+)	Hydronium	(BrO_4^-)	Perbromate
(HO^-)	Hydorxide	(IO_4^-)	Periodate
(CN^-)	Cynide	(CrO_4^{2-})	Chromate
(CO_3^{2-})	Carbonate	(CO_2^{2-})	Carbonite
(NO_3^-)	Nitrate	(NO_2^-)	Nitrite
(PO_4^{3-})	Phosphate	(PO_3^{3-})	Phosphite
(SO_4^{2-})	Sulfate	(SO_3^{2-})	Sulfite
(HCO_3^-)	Hydrogen carbonate	(ClO_4^-)	Perchlorate
(HSO_4^-)	Hydrogne sulfate	(ClO_3^-)	Chlorate
(HPO_4^{2-})	Hydrogenphosphate	(ClO_2^-)	Chlorite
($H_2PO_4^-$)	Dihydrogenphosphate	(ClO^-)	Hypochlorite

Names of compounds containing polyatomic ions

Rules for naming ionic compounds containing polyatomic ions are the same as binary ionic compounds. That is, write the name of the cation followed by the name of the anion. For cations with variable charge, keep the roman numeral in the compound's name. For example, $NaNO_3$ is sodium nitrate, $CaCO_3$ is calcium carbonate, $FeCO_3$ is iron(II) carbonate, NH4Cl is ammonium chloride.

Writing formulae of compounds containing polyatomic ions

The polyatomic ion acts as a single unit, i.e., they are molecular ions. The writing formula of compounds containing polyatomic ions is the same as writing the formula of a binary ionic compound, except that the polyatomic ions must remain intact as a unit. If a subscript is needed for the anion, place the polyatomic ion within small brackets and write the subscript outside the bracket. Just like the subscript to the right of the monoatomic anion tells how many atoms of the anions are there, the subscript to the right of the small bracket around a polyatomic anion tells how many polyatomic anions are there in the compound. For example, iron(III) nitrate is $Fe(NO_3)_3$; sodium carbonate is Na_2CO_3; ammonium phosphate is $(NH_4)_3PO_4$; potassium permanganate is

$KMnO_4$; and calcium phosphate is $Ca_3(PO_4)_2$. Note that in $Fe(NO_3)_3$, there are three nitrate ions, i.e., one iron atom, three nitrogen atoms, and nine oxygen atoms, in the formula unit of the compound.

This page titled 3.3.3: Polyatomic ions and their compounds is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.3.4: Naming acids

Acids donate produce protons and anions when dissolved in water. The naming is different for acids that have oxyanions and the others.

Naming acid that does not contain oxyanion

There are two nomenclatures, first as ionic compounds and second as acids.

Naming acids as ionic compounds

Wright the name of the cation element, i.e., hydrogen, followed by the name of anion without ion at the end. For example, HCl is hydrogen chloride, and HCN is hydrogen cyanide. The ionic name is preferred when the compound is not acting as an acid, e.g., pure HCl in the gas phase.

Naming as acids

Add prefix hydro- to the name of anion and replace the last syllable from –ide to –ic acid. For example, HCl is hydrochloric acid, HCN is hydrocyanic acid, HI is hydroiodic acid. The acid name is preferred when the compound acts as an acid, particularly when it is in solution form in water.

Naming acid that contains oxyanion

Begin the name with the name of oxyanion and change the last syllable from -ate to -ic acid or from -ite to -ous acid. If there is prefix per- or hypo to the name of the oxyanion, it stays in the acid name. For example: ($\mathrm{NO_3^-}$) is nitrate and ($\mathrm{HNO_3}$) is nitric acid ; ($\mathrm{NO_2^-}$) is nitrite and ($\mathrm{HNO_2}$) is nitrous acid; ($\mathrm{SO_4^{2-}}$) is sulfate and {H2SO4}\)) sulfuric acid; ($\mathrm{ClO_4^-}$) is perchlorate and ($\mathrm{HClO_4}$) is perchloric acid; and ($\mathrm{ClO^-}$) is hypochlorite and (HClO) is hypochlorous acid.

One of the commonly encountered oxyacid is acetic acid. Acetic acid is an organic acid with the formula CH_3COOH, where the last hydrogen attached with oxygen is the acidic proton, and the other three oxygen attached with carbon are not acidic. The anion from acetic acid is called acetate ion that has a formula CH_3COO^- which may also be written as $C_2H_3O_2^-$.

This page titled 3.3.4: Naming acids is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.3.5: Naming binary covalent compounds

What is a binary covalent compound?

Binary covalent compounds are atoms of two different elements held together by covalent bonds. Usually, they are composed of nonmetals elements, e.g., laughing gas NO, acid rain causing gas SO_2, etc.

Writing formulae of binary covalent compounds.

The molecular formula of a binary covalent compound shows the symbols of constituent elements, followed by a subscript showing how many atoms of the element are in the molecule. Usually, the symbol of an element closer to metals in a period or a group is written first, followed by the symbol of the other element. For example, CO_2, NO, P_2O_5, where carbon, nitrogen, and phosphorous are nearer to the metals in the periodic table than the other element. An exception occurs when the compound contains oxygen combined with chlorine, bromine, or iodine, where oxygen is closer to metals but is second in the formula, e.g., ClO_2.

Writing the names of binary covalent compounds

The name of binary covalent compounds contains prefixes, listed in Table 1, to indicate the number of atoms followed by the name of the elements according to the following rules:

1. name of the first element in the formula with a prefix showing the number of atoms, followed by,
2. the name of the second element with a prefix showing the number of atoms and its last syllable replaced with –ide.
3. Do not write mono- if it applies to the first element in the formula, but write mono- if it applies to the second element. Write all other prefixes.
4. If the prefix ends with a vowel and the element name begins with a vowel, drop the ending vowel of the prefix.

Table 3.5.1 lists the prefixes used to represent the number of atoms from 1 to 10. Examples of the names are: NO is nitrogen monoxide, CO_2 is carbon dioxide, PCl_3 is phosphorous trichloride, P_2O_5 is diphosphorus pentoxide, SiO_2 is silicon dioxide. Trivial names are well known for some molecular compounds and they are often used in the place of systematic names, e.g., water for H_2O and ammonia for NH_3.

Table 1: Prefixes used in naming binary covalent compounds

Prefix	Means
Mono-	1
di-	2
Tri-	3
tetra-	4
penta-	5
hexa-	6
hepta-	7
octa-	8
nona-	9
deca-	10

This page titled 3.3.5: Naming binary covalent compounds is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

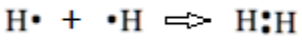
3.3.6: Lewis structures of molecules

Lewis symbols of elements

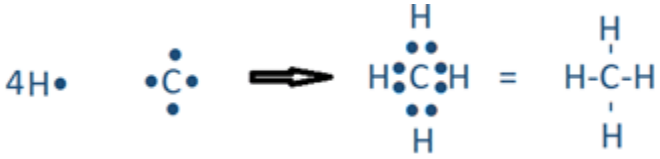
Lewis symbol of element shows the symbol of element with valence electrons shown as dots placed on top, bottom, left, and right sides of the symbol. Valence electrons up to four are shown as a single dot on either side of the symbol. The 5th, 6th, 7th, and 8th valence electron dots are paired with any of the first four dots. For example,  represent hydrogen, beryllium, boron, carbon, nitrogen, fluorine, and neon with 1, 2, 3, 4, 5, 6, 7, and 10 valence electrons, respectively. Lewis symbols of first twenty elements are shown in section 2.6, Table 1.

Lewis structures of simple molecules

An unpaired dot in the Lewis symbol of an element can make one bond by sharing it with an unpaired dot of another atom. The shared pair of dots represents a pair of bonding electrons, a covalent bond. For example, as shown below, a Hydrogen atom has one unpaired valence electron and makes a covalent bond with another hydrogen atom.



The covalent bond is usually represented by a single line between the bonded atoms, e.g., the H_2 molecule shown in the above equation is generally shown as H-H. An example is a reaction between hydrogen having one valence electron and carbon having four valence electrons react to form CH_4 molecule.



Similarly, hydrogen reacts with nitrogen, oxygen, and fluorine to form the following molecules:  ,  , and  . Each line in these molecules represents a **bonding electron pair**, and the pair of dots represent valence electrons that are not involved in bonding, called **lone pair of electrons**. The lone pair is usually omitted from the Lewis structure unless it is needed to emphasize their presence for some reason.

Procedure for writing Lewis structures of molecules

A systematic approach for writing the Lewis structure of molecules is explained with the help of the following example.

Example 3.3.6.1

Draw the Lewis structures of CH_4, PCl_3, CO_2, and HCN

Solution

Step 1: Add the valence electrons of all the molecules' atoms:

- CH_4 has 4 valence electrons in C, and 1 in each of the four H: = 4 + 1x4 = 8 valence electrons
- PCl_3 has 5 valence electros in P and 7 in each of the three Cl: = 5 + 7x3 = 26 valence electrons

- CO_2 has 4 valence electrons in C and 6 in each of the two O: = 4 + 6x2 = 16 valence electrons
- HCN has 1 valence electron in H, 4 in C, and 5 in N: = 1 + 4 + 5 = 10 valence electrons

Step 2: Place the element symbol with more valances, i.e., having more unpaired dots in its Lewis structure, in the center and the rest of the atoms on four sides:



Step 3: Draw a line between the outer atom and the central atom to represent a single covalent bond:



Step 4: Every single bond consumes two valence electrons. Subtract the total number of valence electrons consumed in all the bonds from the total valence electrons initially present in step 1:

- CH_4: 4 bond = 8 electrons consumed. 8 – 8 = 0 electrons left
- PCl_3: 3 bond = 6 electrons consumed. 26 – 6 = 20 electrons left
- CO_2: 2 bond = 4 electrons consumed. 16 – 4 = 12 electrons left
- HCN: 2 bond = 4 electrons consumed. 10 – 4 = 6 electrons left

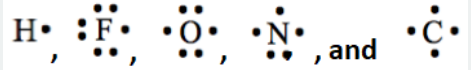
Step 5: Distribute the remaining electrons as lone pairs, first to the outer atoms to complete their octet (duet in the case of hydrogen) and then to the central atom to complete its octet

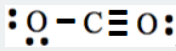


Step 6: Check that the octet of each atom is complete (duet for hydrogen). If yes, the Lewis structure is complete, e.g., as in the cases of CH_4 and PCl_3 in the present examples. If not, move one of the lone pair of electrons from a neighboring atom to make a double bond, as shown by the red color arrows in the figure in the previous step. If the octet is still not complete, move one more lone pair of electrons from a neighboring atom: from the same atom to make a triple bond, as in the case of HCN above, or from another neighboring atom to make two double bonds, as in the case of CO_2 above. The result is Lewis structures shown below.



Note

To draw the Lewis structure of the most stable form, try to keep covalent bonds with an atom equal to the number of unpaired dots in the Lewis symbol of the atoms. For example,  have 1, 1, 2, 3, and 4 unpaired electrons. So, hydrogen makes one, fluorine one, oxygen two, nitrogen 3, and carbon four covalent bonds in the stable molecules. When pulling a lone pair from neighboring atoms is needed to result in a double or triple bond, it is preferable

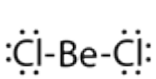
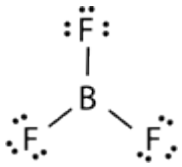
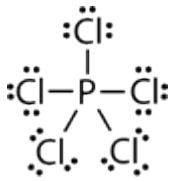
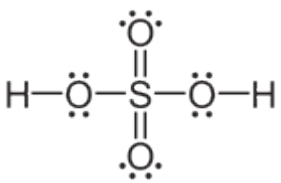
to keep the resulting number of covalent bonds equal to the number of unpaired electrons in the Lewis symbol of the atom. For example, in the case of CO_2 molecule, if both lone pairs were pulled from the same oxygen in step 6 above, the resulting Lewis structure would have been  , which is technically correct Lewis structure but not the most stable form of this molecule. Note that the latter structure has one covalent bond in one oxygen and three covalent bonds in the other instead of two covalent bonds needed for oxygen.

Exceptions to the octet rule

The octet rule generally applies in most cases, but there are exceptions in a few cases:

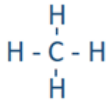
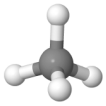
1. Atoms of hydrogen, lithium, and beryllium tend to share, lose, or gain valence electrons to acquire the electron configuration of the nearest noble gas helium having two valence electrons. This is called the **duet rule**.
2. Sometimes the atoms may end up in compounds with less than eight valence electrons. This often happens in the case of boron compounds and aluminum compounds of group 13. For example, boron has three valence electrons resulting in compounds like BF_3 with three covalent bonds and six instead of eight valence electrons around boron.
3. Atoms of elements in period three and beyond sometimes end up in compounds with more than eight valence electrons. For example, PCl_3 has octet complete, but PCl_5 has 10 valence electrons. Similarly, sulfur has its octet complete in H_2S but has 12 valence electrons in H_2SO_4. This happens because atoms in period thee and beyond have larger sizes and they have valence electrons in d or f orbitals in addition to the valence s and p orbitals. Table 1 lists some examples of exceptions to the octet rule.

Table 1: Some examples of exceptions to the octet rule.

Structure	H:H				
Name	Hydrogen	Beryllium dichloride	Boron trifluoride	Phosphorous pentafluoride	Sulfuric acid
Valence electrons on the central atom	2	4	6	10	12

This page titled 3.3.6: Lewis structures of molecules is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

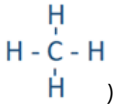
3.3.7: Molecular shapes –Valence shell electron pair repulsion (VSEPR) theory

The Lewis structure tells the connection between atoms and any lone pair present, but it does not tell the exact angles of bonds around the central atom or the actual shape of the molecule. The conventional way of presenting a Lewis structure of a molecule shows it as a planer, e.g., CH_4 as  which implies the molecule to be planar and H-C-H bond angles to be either 90° or 180°. The actual CH_4 molecule is nonplanar as  with all H-C-H bond angles 109.5^o. the following theory helps explain the actual shapes of molecules.

Valence shell electron pair repulsion (VSERP) theory

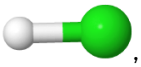
The valence shell electron pair repulsion theory predicts the shape of the molecule and bond angles based on the fact that valence shell electrons around the central atom in a molecule are grouped. The electron groups repel each other and go as far apart from each other as possible.

Electron groups

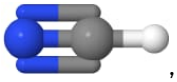
A lone pair, a single bond, a double bond, and a triple bond, each of these is one electron group. This is because two elections of single bond, four electrons of a double bond, and six electrons of a triple bond are located in the region along the axis of the bond, i.e., they are grouped together. Similarly, a lone pair is located in a defined space around the atom. For example, carbon in methane () has four electron groups that are the four single bonds (C-H bonds) around the carbon. Carbon in carbon dioxide (O=C=O) has two electron groups that are the two double bonds around the carbon, and in H-C≡N has two electron groups that are a single bond (C-H bond) and a triple bond (C≡N bond).

Molecular shapes and bond angles based on VSEPR theory

One electron group

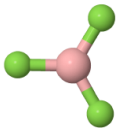
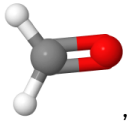
One electron group between two atoms is always a linear molecule. For example, H-H ,O=O, N≡N, and H-Cl  , are linear molecules, where hydrogen is white and chlorine is green in the H-Cl model.

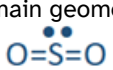
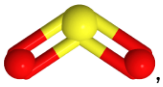
Two-electron groups

Two-electron groups are farthest apart in a **linear geometry** with the central atom in the middle of the line and the bond angles of 180° around the central atom. The examples include CO_2  and HCN  , where the central carbon atom is gray, hydrogen is white, nitrogen is blue, and oxygen is red.

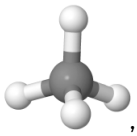
Three-electron groups

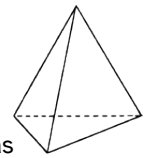
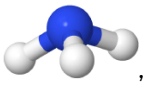
Three-electron groups are farthest apart when they are at the corners of a triangle in a **planar trigonal geometry** with the central

atom in the middle of the triangle and the bond angles of 120° around the central atom. Examples include BF_3  , and H_2CO  , where boron is pink, F is green, carbon is gray, oxygen is red, and hydrogen is white.

If one of the electron domains is a lone pair, the electron domain geometry remains the same, but the geometry of the atoms in the molecule, i.e., **molecule geometry,** is **bent**. For example,  has three electron domains and trigonal planar electron domain geometry, but there is one lone pair. So, the molecule geometry is **bent** as  , where sulfur is yellow, and oxygen is red (lone pair in not shown).

Four-electron groups

Four-electron groups are farthest apart when they are at the corners of a tetrahedron in a **tetrahedral geometry** with the central atom at the center of the tetrahedron and the bond angles of 109.5^{o} around the central atom as: . An example is methane CH_4  , where carbon is gray, and hydrogens are white.

If one of the electron domains is a lone pair, the electron domain geometry is still tetrahedral, but the molecule geometry is **trigonal pyramidal** as  with three pereferal atoms at the corners of the triangel and the central atom raised to the top of the pyramid. An example is ammonia ($:NH_3$)  , where nitrogen is blue, and hydrogens are white.

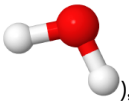
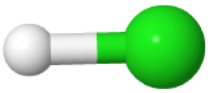
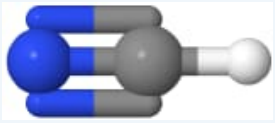
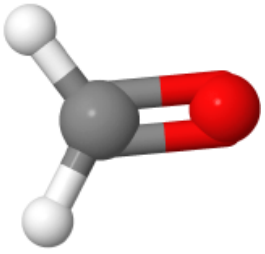
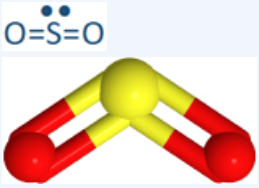
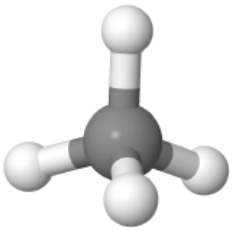
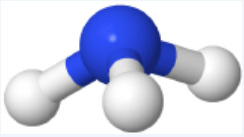
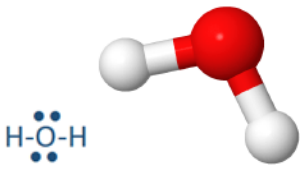
If two electron domains are lone pairs, the electron domains geometry is still tetrahedral, but the molecule geometry is bent. An example is water ( ), where oxygen is red, and hydrogens are white.

Table 1 is the summary of the electron domain geometries and the corresponding molecular geometries.

Table 1: Common molecular geometries around the central atom

Electron domain	Lone pair	Electron domain geometry	Molecule geometry	Bond angles	Examples
1	0	Linear	Linear	-	HCl 

Electron domain	Lone pair	Electron domain geometry	Molecule geometry	Bond angles	Examples
2	0	Linear	Linear	180°	CO_2  HCN 
3	0	Trigonal pyramidal	Trigonal pyramidal	120°	 H_2CO
3	2	Trigonal pyramidal	Bent	120°	
4	0	Tetrahedral	Tetrahedral	109.5°	 CH_4
4	1	Tetrahedral	Trigonal pyramidal	109.5°	:NH_3 
4	2	Tetrahedral	Bent	109.5°	

This page titled 3.3.7: Molecular shapes –Valence shell electron pair repulsion (VSEPR) theory is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.3.8: Polarity of molecules

The polarity of an atom

The negative charge of electrons balances the positive charge of protons in an atom. The electrons symmetrically distributed around the nucleus leave no negative or positive end. The atoms are nonpolar. Fig. 3.8.1 illustrates the polarity of a hydrogen atom with color codes.

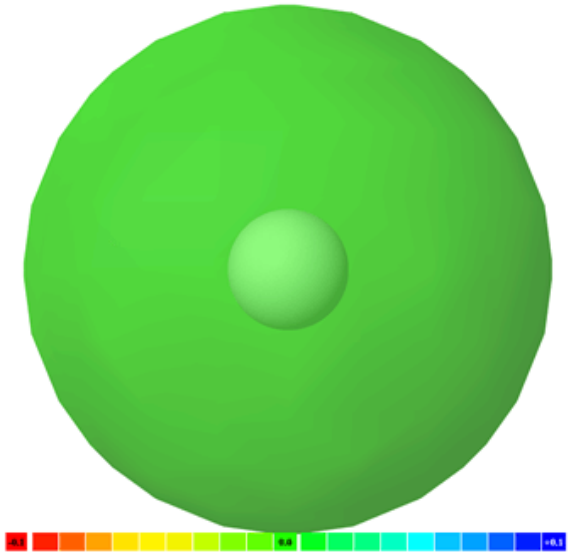


Figure 3.3.8.1: Electrostatic potential map of hydrogen atom, shown nonpolar by green color code. Source: drawn using free software https://chemagic.org/molecules/amini.html

The polarity of a covalent bond

When atoms of one element combine to make a covalent bond, e.g., H-H and F-F, the positive and negative charges are still symmetrical, and the bond is **nonpolar**, i.e., no negative end separated from a positive end. However, when atoms of different elements combine to make a covalent bond, the more electronegative atom attracts the bonding electron pair towards itself stronger than the other atom. The separation of positive and negative charges happens as the electrons shift more towards the electronegative atom. The bond becomes **polar** with a partially positive (δ+) end on the electropositive atom and a partially negative (δ-) end on the electronegative atom. For example, fluorine is more electronegative than hydrogen. Consequently, fluorine pulls the bonding electron pair towards itself in the H-F molecule, creating a partial negative charge (δ-) on fluorine and a partial positive charge (δ+) on hydrogen. The H-F bond is polar. Fig. 3.8.2 illustrates the polarity in H-F with color codes.

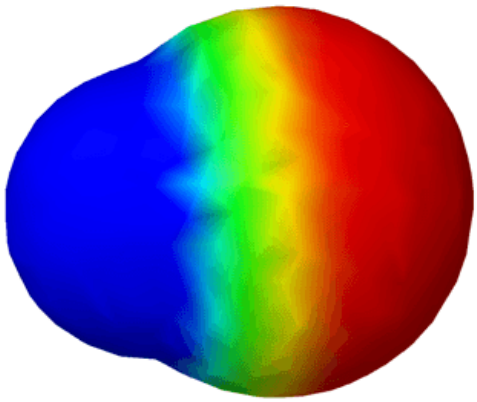


Figure 3.3.8.2: H-F molecule showing δ+ end in blue, δ- end in red and neutral part in green. Source: drawn using free software http://molview.org/?cid=14917

Bond polarity is a vector that has a magnitude and direction and can be represented by an arrow, like other vectors, as shown in Fig. 3.8.3 for the case of a water molecule.

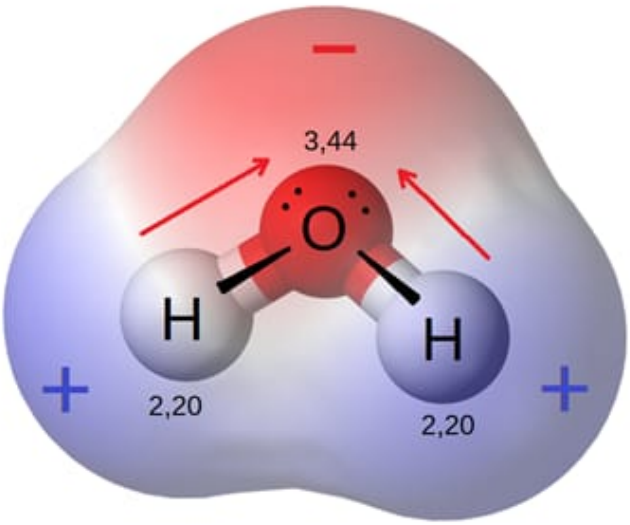


Figure 3.3.8.3: The water molecule is made up of oxygen and hydrogen, with respective electronegativities of 3.44 and 2.20. The electronegativity difference polarizes each H–O bond, shifting its electrons towards the oxygen (illustrated by red arrows). These effects add as vectors to make the overall molecule polar. Source: Riccardo Rovinetti / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

A bond is categorized as a nonpolar covalent, polar covalent, or an ionic bond based on the following convention: nonpolar covalent if the electronegative difference of the bonded atom is less than 0.5, a polar covalent if the electronegativity difference is between 0.5 to 1.9, and an ionic if the electronegativity difference is more than 1.9.

The polarity of a Molecule

The molecules fall into the following categories concerning molecular polarity.

The molecule is nonpolar if there is no polar bond in it, e.g., H-H, F-F, and CH_4 are nonpolar molecules. Fig. 3.8.4 illustrates CH_4 molecules with green color electron clouds that represent a nonpolar molecule.

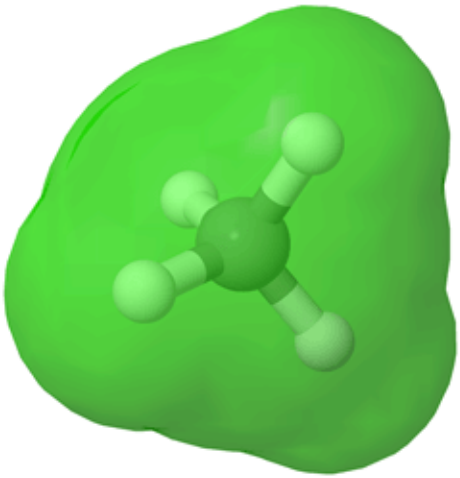


Figure 3.3.8.4: Methane (CH_4) with no polar bond is nonpolar. Source: drawn using free software https://chemagic.org/molecules/amini.html

If there is only one polar bond in a molecule, then the molecule is polar, e.g., the H-F molecule shown in Fig. 3.8.2.

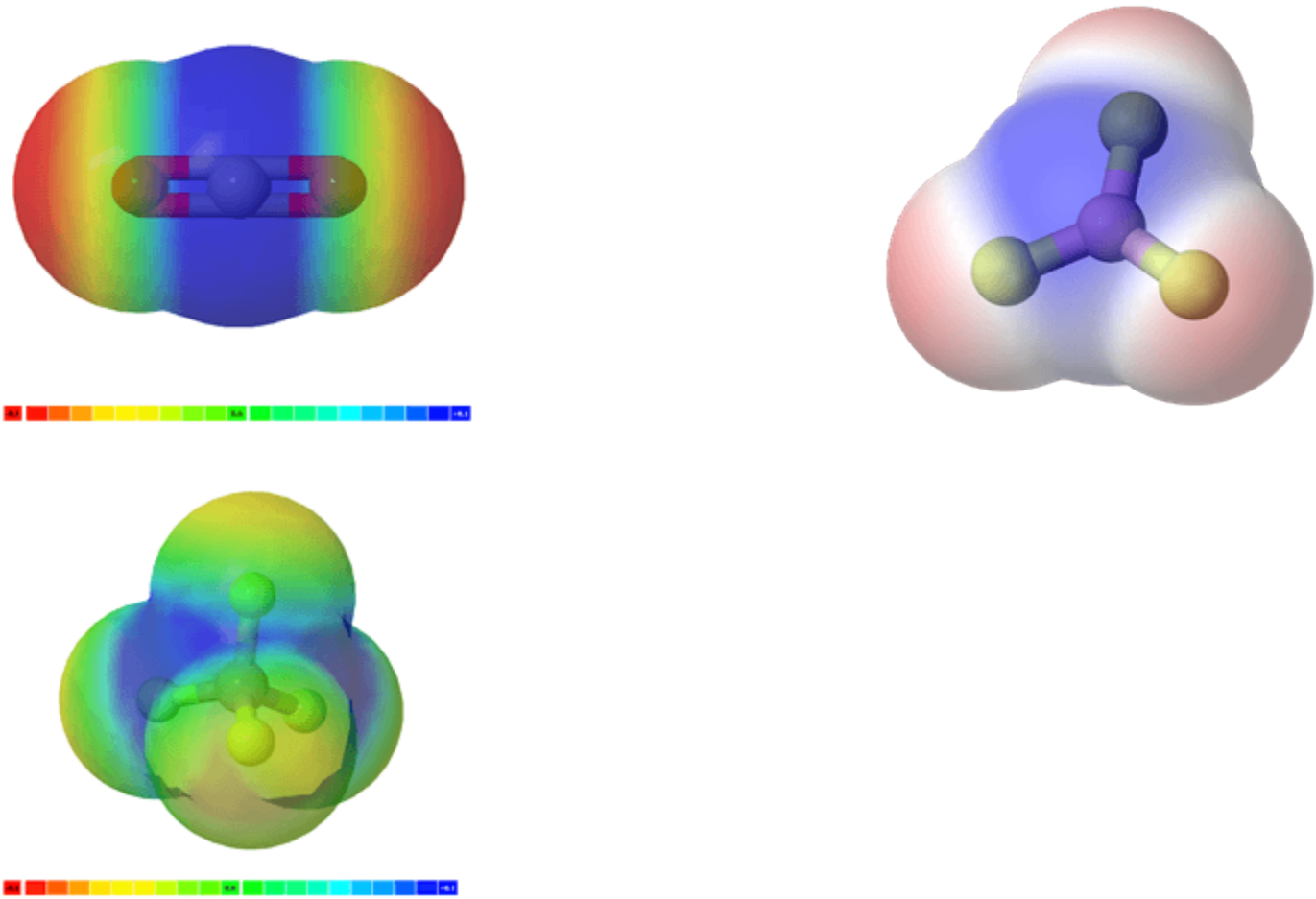


Figure 3.3.8.5: Examples of symmetric molecules with all polar bonds but the molecule as a whole is nonpolar. The examples are (from left to right) carbon dioxide CO_2, boron trifluoride BF_3, and carbon tetrafluoride CF_4.

If there is more than one polar bond in a molecule, the molecule may be polar or may be nonpolar, depending on the symmetry of the molecule: a) Polarity vector of individual bonds cancel out in symmetric molecules making the molecule nonpolar. For example, symmetric molecules, like CO_2, BF_3, and CCl_4, are nonpolar, although each bond in them is polar. Fig. 3.8.5 illustrates the symmetric molecules that have polar bonds, but the polarity of bonds cancels each other, making the molecule nonpolar. b) If a molecule has polar bonds and it is not symmetric, the polarity vectors do not cancel out, and the molecule is polar. Examples of polar molecules include $CHCl_3$, NH_3, and H_2O, as illustrated in Fig. 3.8.6.

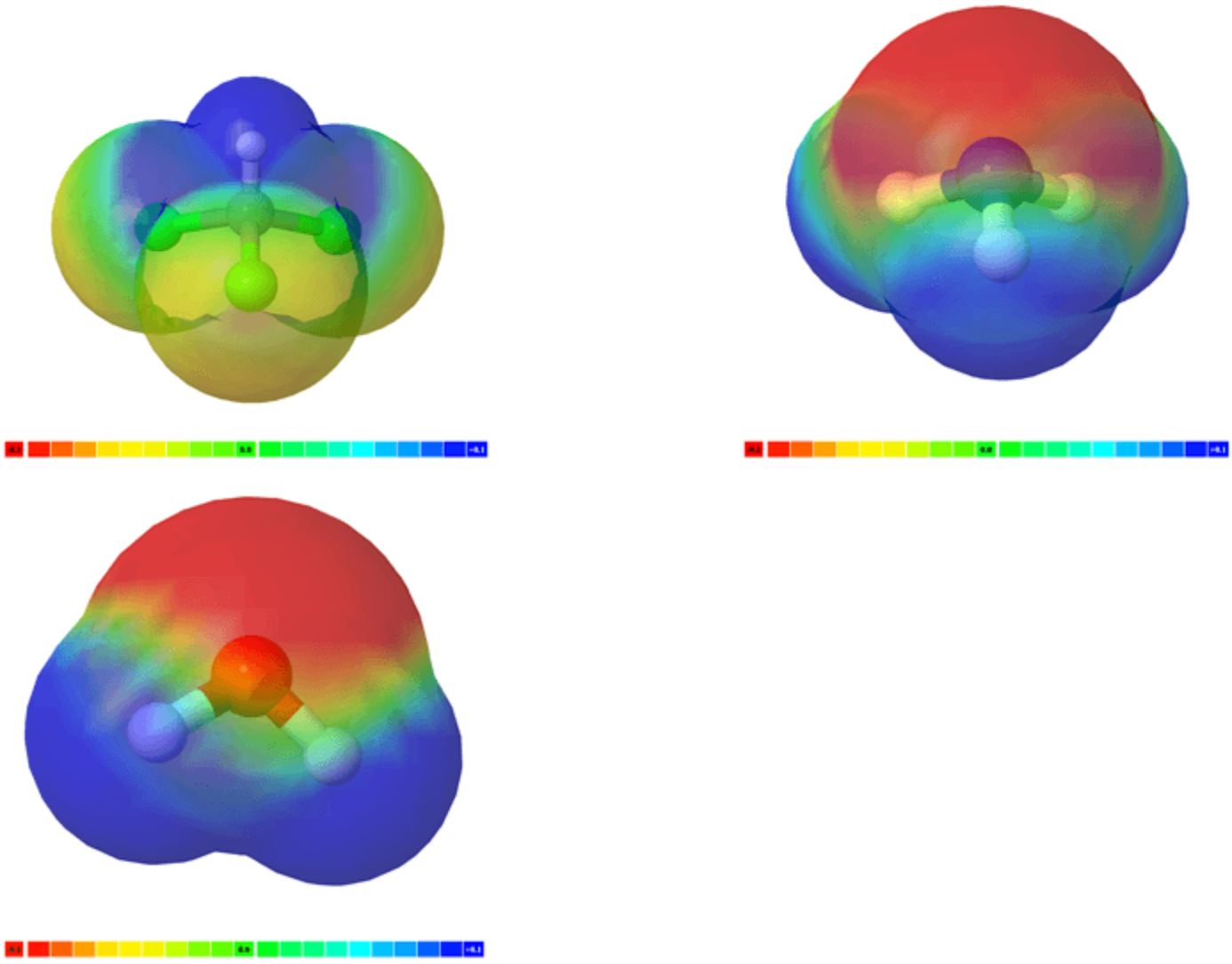


Figure 3.3.8.6: Non-symmetric molecules with polar bonds are polar molecules with a net δ+ end in red color and δ- ends in blue color. The examples are (from left to right) chloroform ($CHCl_3$), ammonia (NH_3), and water (H_2O).

This page titled 3.3.8: Polarity of molecules is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.3.9: Intramolecular forces and intermolecular forces

There are electrostatic interaction between charges or partial charges, i.e., the same charges repel each other, and opposite charges attract each other, as illustrated in Fig. 3.9.1.There are two types of electrostatic forces in compounds or molecules, intramolecular forces that exist between the bonded atoms of a compound or a molecule, and intermolecular forces that exist between molecules as described below.



Figure 3.3.9.1: Like charges repel each other and opposite charges attract each other.

Intramolecular forces

Intramolecular forces are the chemical bonds holding the atoms together in the molecules. The three major types of chemical bonds are the metallic bond, the ionic bond, and the covalent bond.

Metallic bond

Metals exist as a collection of many atoms as +ions arranged in a well-defined 3D arrangement called crystal lattice with some of the outermost electrons roaming around in the whole piece of the metal, forming a sea of electrons around the metal atoms, as illustrated in Fig. 3.9.2. The attraction between +ions and the sea of free moving electrons is the metallic bond that holds the atoms together in a piece of metal. The metallic bond is usually the strongest type of chemical bond.

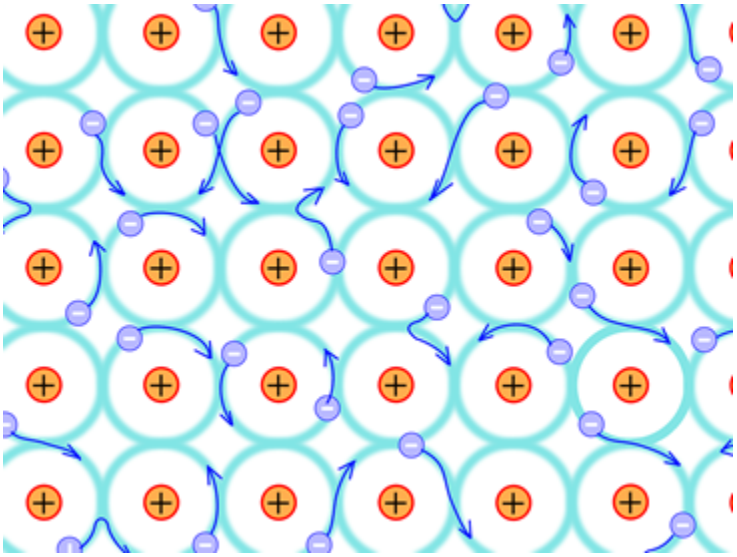


Figure 3.3.9.2: A model of metallic bond. Source: https://commons.wikimedia.org/wiki/F...bond_model.svg, Public domain

Ionic bond

When the electronegativity difference between bonded atoms is large, i.e., more than 1.9 in most cases, the bonding electrons completely transfer from a more electropositive atom to a more electronegative atom creating a cation and an anion, respectively. There is the electrostatic interaction between cation and anion, i.e., the same charges attract each other, and opposite charges repel each other, as illustrated in Fig. 3.9.1. The cations and anions orient themselves in a 3D crystal lattice in such a way that attractive interactions maximize and the repulsive interactions minimize, as illustrated in Fig. 3.9.3. Ionic bonds are usually

weaker than metallic bonds but stronger there the other types of bonds.

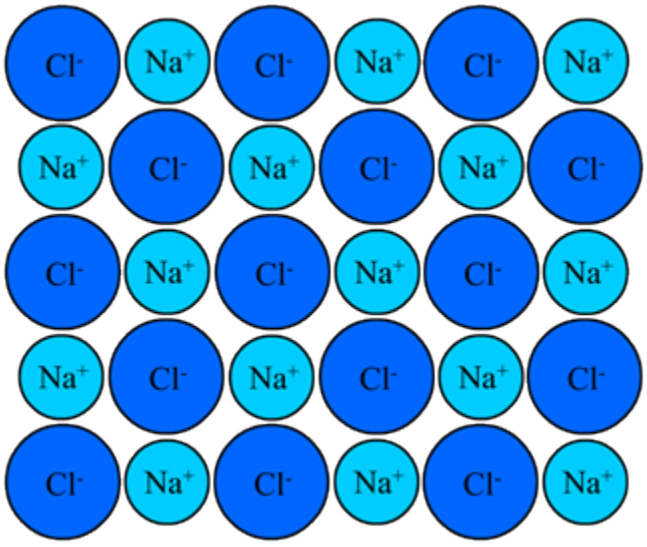


Figure 3.3.9.3: Illustration of an ionic bond: the ions in ionic crystals orient in such a way to optimize the attractive forces between opposite charges. Part of one layer of a crystal lattice is shown in this illustration, the 3D crystal lattice is illustrated in Fig. 3.1.6. Source: Eyal Bairey / Public domain

Covalent bond

When the electronegativity difference between bonded atoms is moderate to zero, i.e., usually less than 1.9, the bonding electrons are shared between the bonded atoms, as illustrated in Fig. 3.9.4. The attractive force between the bonding electrons and the nuclei is the covalent bond that holds the atoms together in the molecules. The covalent bond is usually weaker than the metallic and the ionic bonds but much stronger than the intermolecular forces.

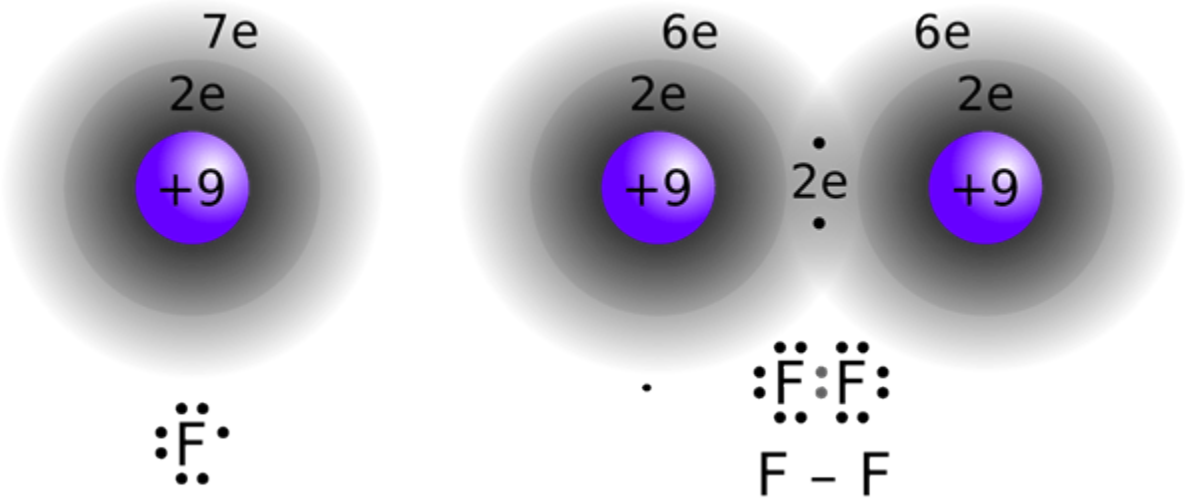


Figure 3.3.9.4: Illustration of a covalent bond in fluorine molecule: Jacek FH, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0>, via Wikimedia Commons

Criteria to predict the type of chemical bond

Metals tend to have lower electronegativity and nonmetals have higher electronegativity. When the electronegativity difference between the bonded atoms is large, usually more than 1.9, the bond is ionic. Generally, a bond between a metal and a nonmetal is ionic. When the electronegativity difference is low, usually less than 1.9, the bond is either metallic or covalent. Nonmetals tend to make a covalent bond with each other. Nonmetals also have higher electronegativities. So, when the average electronegativity of the bonded atom is high and the electronegativity difference between them is low, they tend to make a covalent bond. Metals

tend to make the metallic bond with each other. Metals also tend to have lower electronegativity values. So, when the average electronegativity of the bonded atom is low and the electronegativity difference between them is also low, they tend to make a metallic bond. Fig. 3.9.5 illustrates the criteria to predict the type of chemical bond based on the electronegativity difference. Keep in mind that there is no sharp boundary between metallic, ionic, and covalent bonds based on the electronegativity differences or the average electronegativity values.

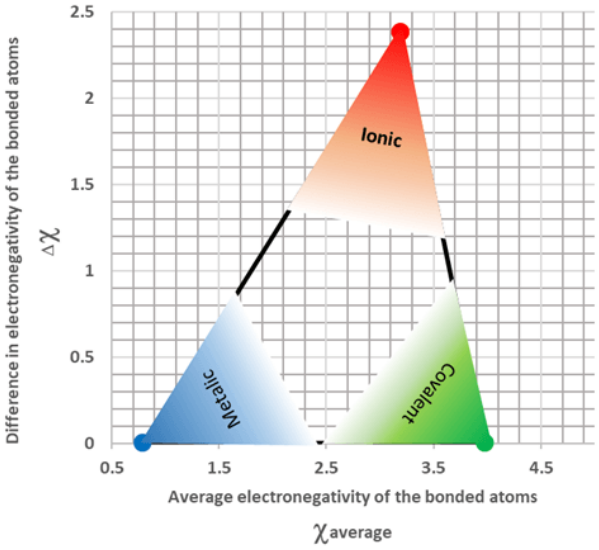


Figure 3.3.9.5: van Arkel-Ketelaar Triangle: plot of the difference in electronegativity ($_{\Delta}\chi$) versus the average electronegativity in a bond ($\chi_{average}$). The top corner is mostly ionic, the lower left is metallic, and the lower right is covalent.

Intermolecular forces

Intermolecular forces are the electrostatic interactions between molecules. The intermolecular forces are usually much weaker than the intramolecular forces, but still, they play important role in determining the properties of the compounds. The major intermolecular forces include dipole-dipole interaction, hydrogen bonding, and London dispersion forces.

Dipole-dipole interactions

Polar molecules have permanent dipoles, one end of the molecule is partial positive (δ+) and the other is partial negative (δ-). The polar molecules have electrostatic interactions with each other through their δ+ and δ- ends called dipole-dipole interactions, though these interactions are weaker than ionic bonds. The polar molecules orient in a way to maximize the attractive forces between the opposite charges and minimize the repulsive forces between the same charges, as illustrated in Fig. 3.9.6.

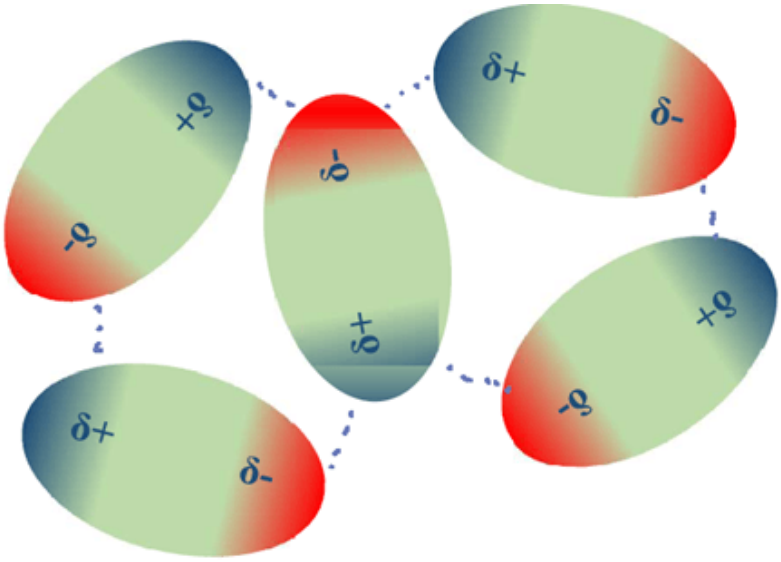


Figure 3.3.9.6: Illustration of dipole-dipole interactions by dotted liens in polar molecules.

Hydrogen bonds

Hydrogen bonding is a dipole-dipole interaction when the dipole is a hydrogen bond to O, N, or F, e.g. in water molecules as illustrated in Fig. 3.9.7. Although hydrogen bond is a dipole-dipole interaction, it is distinguished from the usual dipole-dipole interactions because of the following special features.

1. The electronegativity difference between H and O, N, or F is usually more than other polar bonds.
2. The charge density on hydrogen is higher than the δ+ ends of the rest of the dipoles because of the smaller size of hydrogen.
3. The δ+ Hydrogen can penetrate in less accessible spaces to interact with the δ- O, N, or F of the other molecule because of its small size.

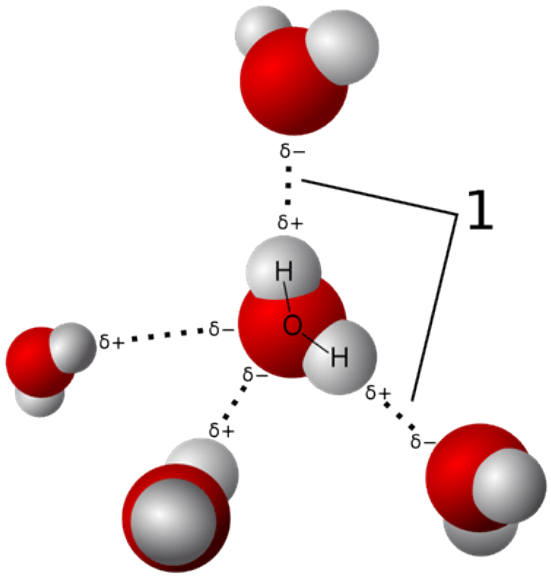


Figure 3.3.9.7: Illustration of hydrogen bonds by dotted lines in water molecules. Red balls are oxygen and white balls are hydrogen atoms. Source: User Qwerter at Czech Wikipedia: Qwerter. Transferred from cs.Wikipedia to Commons by sevela.p. Translated to english by by Michal Maňas (User:snek01). Vectorized by Magasjukur2 / Public domain

A hydrogen bond is usually stronger than the usual dipole-dipole interactions. Hydrogen bonding is the most common and essential intermolecular interaction in biomolecules. For example, two strands of DNA molecules are held together through hydrogen bonding, as illustrated in Fig. 3.9.8. Proteins also acquire structural features needed for their functions mainly through

hydrogen bonding.

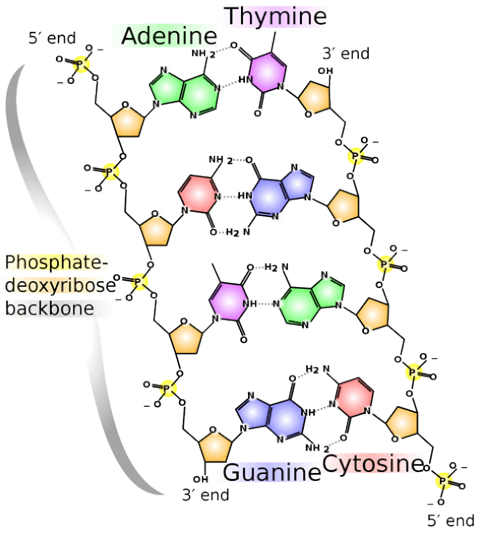


Figure 3.3.9.8: Chemical structure of DNA, with a colored label identifying the four bases as well as the phosphate and deoxyribose components of the backbone. Dotted lines are the hydrogen bonds holding the two strands of DNA together. Source: Madprime (talk · contribs) / CC0

London dispersion forces

It may appear that the nonpolar molecules should not have intermolecular interactions. Practically, there are intermolecular interactions called London dispersion forces, in all the molecules, including the nonpolar molecules. The electron cloud around atoms is not all the time symmetrical around the nuclei. It temporarily sways to one side or the other, generating a **transient dipole**. The transient dipole induces a dipole in the neighboring. A transient dipole-induced dipole interaction, called **London dispersion force** or **wander Wall's force**, is established between the neighboring molecules as illustrated in Fig. 3.9.9. Although London dispersion forces are transient, they keep re-appearing randomly distributed in space and time. London dispersion forces are not unique to nonpolar molecules, they are present in all types of molecules, but these are the only intramolecular forces present in the nonpolar molecules.

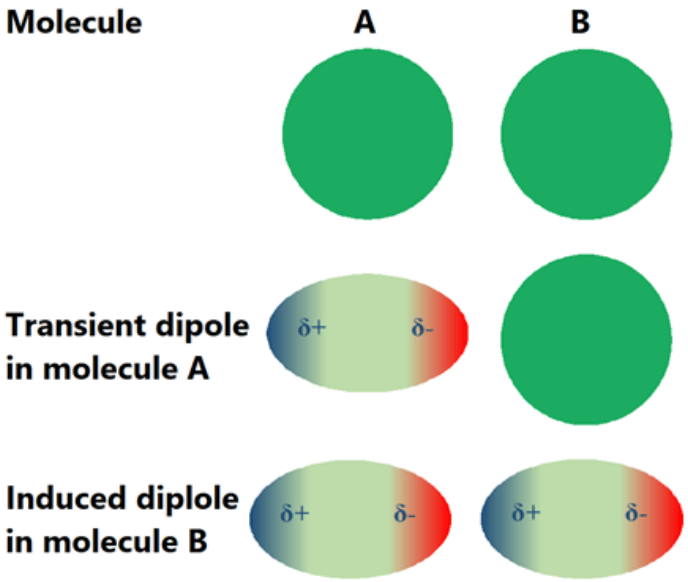


Figure 3.3.9.9: Illustration of transient dipole-induced dipole, i.e., London dispersion forces.

This page titled 3.3.9: Intramolecular forces and intermolecular forces is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 3.4: STOICHIOMETRY –THE QUANTIFICATION OF CHEMICAL REACTIONS

3.4.1: Stoichiometry

3.4.2: The mole

3.4.3: Chemical reaction

3.4.4: Patterns of chemical reactions

3.4.5: Oxidation-reduction reactions

3.4.6: Energetics of chemical reactions

3.4.7: Stoichiometric calculations

This page titled Chapter 3.4: Stoichiometry –the quantification of chemical reactions is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.4.1: Stoichiometry

What is Stoichiometry

> Stoichiometry (stoi·chi·om·e·try /ˌstɔɪkiˈɒmɪtri/) is the study of the quantities of substances and energy consumed and produced in chemical reactions.

The basis of the stoichiometric calculations is the law of conservation of mass which states that the mass is neither created nor destroyed in a chemical reaction. Another form of the law states that atoms are neither created nor destroyed in a chemical reaction. It is the basis of stoichiometric calculations that are described in this chapter.

It can be concluded from the law of mass action that atoms of each element and their masses are the same in reactants and products. A balanced chemical equation shows atoms of each element and the total mass of reactants equal to that in the product, as illustrated in Fig. 4.1.1. The number of atoms and molecules is related to their quantity in moles through Avogadro's number. The mole, in turn, is related to the mass of the substance through molar mass in grams. These relationships are described in the next sections.

$2H_2 + O_2 \rightarrow 2H_2O$	
Reactants	Products
Moles of H atom = 4 mol, and O atoms = 2 mol	Moles of H atoms = 4 mol, and O atoms = 2 mol
Mass of 2 mol H_2 = 4.032 g, and 1 mol O_2 = 32.00 g Total: 36.03 g	Mass of 2 mol H_2O = 36.03 g Total: 36.03 g

Figure 3.4.1.1: Stoichiometry of water formation from hydrogen and oxygen.

Atomic mass

Atomic mass is the weighted average of the masses of the isotopes present in a natural sample of the element, as explained with an example calculation in section 2.3. The mass of a single atom or molecule is expressed in the atomic mass unit (amu),which is equal to $\frac{1}{2}$th of the mass of $^{12}_{6}\mathrm{C}$ isotope of carbon that is unbound.

Atomic mass is listed in a periodic table as a number below the symbol and name of the element, as illustrated in Fig. 4.1.2. Atomic mass is listed as a number without a unit because it is the mass of an atom in amu and it is also the mass of one mole (molar mass) of the atom in grams. Molar mass is often used in stoichiometry calculation as explained in the next sections.

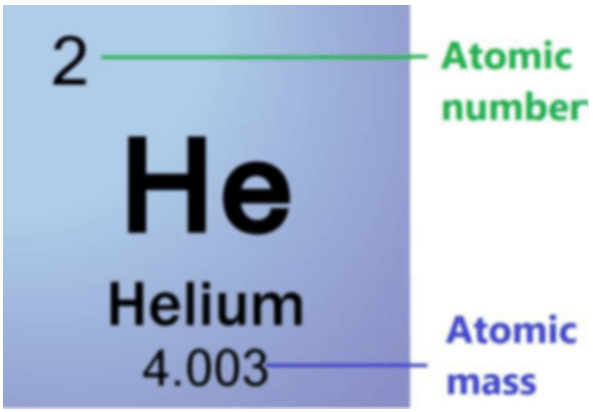


Figure 3.4.1.2: Copy and Paste Caption here. (Copyright; author via source)Atomic number and atomic mass in a periodic table for Helium

This page titled 3.4.1: Stoichiometry is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.4.2: The mole

Avogadro’s number and mole

The Avogadro's number

The Avogadro's number is equal to 6.02214076 x 10^{23} exact.

Just like the dozen is equal to 12, the Avogadro's number 6.02214076 x 10^{32} is exact by definition, but usually, 6.022 x 10^{23} is used in calculations with 4 significant figures.

Mole (mol)

The mole is a SI unit of the amount of a substance that is equal to 6.02214076 x 10^{23} particles of the substance.

The particle of a substance are usually atoms, ions, or molecules. For example, 6.02214076 x 10^{23} atoms of $^{12}_{6}\mathrm{C}$ isotope is one mole of $^{12}_{6}\mathrm{C}$. The number 6.02214076 x 10^{32} is exact by definition, but usually, 6.022 x 10^{23} is used with 4 significant figures.

One mole of a substance is equal to one Avogadro’s number of atoms, molecules, or formula units of the substance. i.e.,

$$1 \text{ Avogadro's number of particles } = 1 \text{ mol} = 6.022 \times 10^{23} \text{ particles}$$

, where particles are atoms, molecules, or formula units in chemistry.

The equality between Avogadro's number and mole gives two conversion factor:

$$\frac{6.022 \times 10^{23} \text{ particles}}{1 \text{ mol}} \quad \text{and} \quad \frac{1 \text{ mol}}{6.022 \times 10^{23} \text{ particles}}$$

, where the first factor is used to convert a number of moles to a number of particles and the second for a number of particles to number moles conversions, as explained in the following examples.

Example 3.4.2.1

How many aspirin ($\mathrm{C_9H_8O_4}$ molecules are in 0.0139 mol of aspirin?

Solution

Step 1. Write the given quantity and the desired quantity.

Given: 0.0139 mol ($\mathrm{C_9H_8O_4}$, Desired: ? molecules of ($\mathrm{C_9H_8O_4}$

Step 2. Write the two conversion factors from equality between the given and the desired quantity.

$$\frac{6.022 \times 10^{23} \text{ particles}}{1 \text{ mol}} \quad \text{and} \quad \frac{1 \text{ mol}}{6.022 \times 10^{23} \text{ particles}}$$

Step 3. Multiply the given quantity with the conversion factor that cancels the given unit and leaves the desired unit in the answer.

$$0.0139 \cancel{\text{ mol } C_9H_8O_4} \times \frac{6.022 \times 10^{23} \text{ molecules } C_9H_8O_4}{1 \cancel{\text{ mol } C_9H_8O_4}} = 8.37 \times 10^{21} \text{ molecules } C_9H_8O_4$$

Example 3.4.2.2

How many moles of aspirin (($C_9H_8O_4$) are in 9.50 x 10^{25} molecules of aspirin?

Solution

Step 1. Write the given quantity and the desired quantity.

Given: 9.50 x 10^{25} molecules of aspirin, Desired: ? mol of aspirin

Step 2. Write the two conversion factors from equality between the given and the desired quantity.

$$\frac{6.022 \times 10^{23} \text{ particles}}{1 \text{ mol}} \quad \text{and} \quad \frac{1 \text{ mol}}{6.022 \times 10^{23} \text{ particles}}$$

Step 3. Multiply the given quantity with the conversion factor that cancels the given unit and leaves the desired unit in the answer.

$$9.50 \times 10^{25} \cancel{\text{ molecules } C_9H_8O_4} \times \frac{1 \text{ mol } C_9H_8O_4}{6.022 \times 10^{23} \cancel{\text{ molecules } C_9H_8O_4}} = 158 \text{ mol } C_9H_8O_4$$

Moles of elements in a mole of a compound

Moles of an element in a mole of a compound is equivalent to atoms of the element in a molecule or formula unit of the compound. For example in 1 mole of glucose ($C_6H_{12}O_6$) there are 6 moles of carbon, 12 moles of hydrogen, and 6 moles of oxygen. Each of these equalities between a mole of a substance and the moles of the element in it gives two conversion factors for the calculations.

Example 3.4.2.3

How many moles of hydrogen are in 3.0 moles of glucose ($C_6H_{12}O_6$)?

Solution

Step 1. Write the given quantity and the desired quantity.

Given: 3.0 moles of ($C_6H_{12}O_6$), Desired: ? mol of H

Step 2. Write the two conversion factors from equality between the given and the desired quantity.

$$\frac{12 \ mol \ H}{1 \ mol \ C_6H_{12}O_6} \quad \text{and} \quad \frac{12 \ mol \ C_6H_{12}O_6}{1 \ mol \ H}$$

Step 3. Multiply the given quantity with the conversion factor that cancels the given unit and leaves the desired unit in the answer.

$$3.0 \cancel{\text{ mol } C_6H_{12}O_6} \times \frac{12 \text{ mol H}}{1 \cancel{\text{ mol } C_6H_{12}O_6}} = 36 \ mol \ H$$

Example 3.4.2.4

Calculate the moles of oxygen are in 0.50 mole of $\mathrm{Ca(NO_3)_2}$?

Solution

Step 1. Write the given quantity and the desired quantity.

Given: 0.50 moles of $\mathrm{Ca(NO_3)_2}$, Desired: ? mol of O

Step 2. Write the two conversion factors from equality between the given and the desired quantity. Note that $\mathrm{NO_3^-}$ is polyatomic ion that has three oxygen atoms in it. There are two $\mathrm{NO_3^-}$ in the formula unit as shown by subscript 2 outside the bracket enclosing the polyatomic anion (\ce{NO3^-}\). So, the equality is:

$$1 \text{ mole } \mathrm{Ca(NO_3)_2} = 6 \text{ mole O},$$

and the two conversion factors from the equality are:

$$\frac{1\ mol\ \mathrm{Ca(NO_3)_2}}{6\ mol\ O} \quad \text{and} \quad \frac{6\ mol\ O}{1\ mol\ \mathrm{Ca(NO_3)_2}}$$

Step 3. Multiply the given quantity with the conversion factor that cancels the given unit and leaves the desired unit in the answer.

$$0.50\ \cancel{mol\ \mathrm{Ca(NO_3)_2}} \times \frac{6\ mol\ O}{1\ \cancel{mol\ \mathrm{Ca(NO_3)_2}}} = 3.0\ mol\ O$$

Molar mass

The mass of one mole of a substance, in $\frac{g}{mol}$, is called the molar mass.

Recall that one mole = 1 Avogadro's number, i.e., 6.022×10^{23} atoms, molecules, or formula units of a substance. Fig. 4.2.1 helps in visualizing the molar masses of aluminum, copper, and carbon. The molar mass of an element or a compound is a reasonable quantity to be measured on an analytical balance commonly available in laboratories, while the mass of an individual atom or molecule is too small to be easily measured. That is why the mole is commonly used in stoichiometric calculations.

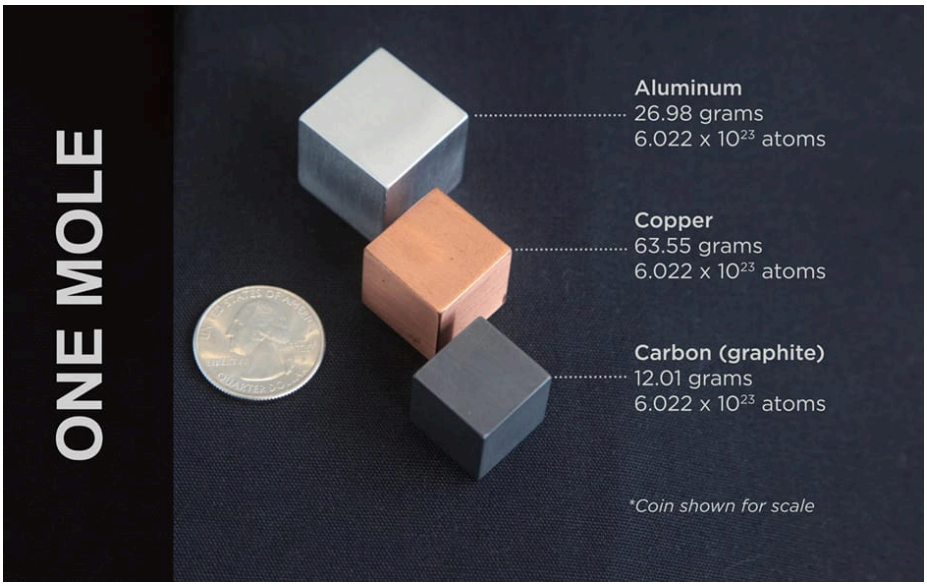


Figure 3.4.2.1: A mole of aluminum, copper,

and carbon visualized. Source: photo: R. Press/NIST; graphic design: N. Hanacek/NIST

The molar mass of atoms of an element

The atomic mass listed in the periodic table is the molar mass of atoms of the element in $\frac{g}{mol}$. For example, the atomic mass of H listed in the periodic table is 1.008, so the molar mass of H is 1.008 $\frac{g}{mol}$. Similarly, the atomic mass of O listed in the periodic table is 15.999, and the molar mass of O is 15.999 $\frac{g}{mol}$.

The molar mass of molecules of an element

The molar mass of molecules of an element is the sum of atomic mass of atoms in the molecule expressed in $\frac{g}{mol}$. For example, the molar mass of H_2 is 1.008 $\frac{g}{mol}$ + 1.008 $\frac{g}{mol}$ = 2.016 $\frac{g}{mol}$.

The molar mass of a compound

The molar mass of a compound is the sum of the atomic masses of all the atoms in the molecular formula or formula unit of the compound. For example, the molar mass of water (H_2O) is the sum of the molar mass of two hydrogen atoms + the molar mass of one oxygen atom, i.e., 2x1.008 $\frac{g}{mol}$ H + 15.999 $\frac{g}{mol}$ O = 18.02 $\frac{g}{mol}$ H_2O. In other words, to calculate the molar mass of a compound, take the atomic masses of the constituent elements from a periodic table, multiply them with the number of atoms of the element in the formula of the compound, and then add these numbers, as explained in the following examples. Note that the unit $\frac{g}{mol}$ can also be written as: $g.mol^{-1}$.

Example 3.4.2.5

Calculate the molar mass of NaOH?

Solution

Step 1. Find the atomic masses of the constituent elements from the periodic table.

Na = 22.990 $g.mol^{-1}$, O = 15.999 $g.mol^{-1}$, H = 1.008 $g.mol^{-1}$.

Step 2. Multiply the atomic masses with the number of atoms in the fourmula.

1x22.990 $g.mol^{-1}$ Na, 1x15.999 $g.mol^{-1}$O, 1x1.008 $g.mol^{-1}$ H.

Step 3. Add all the numbers from step 2.

1x22.990 $g.mol^{-1}$ Na + 1x15.999 $g.mol^{-1}$O + 1x1.008 $g.mol^{-1}$ H = 39.997 $g.mol^{-1}$ NaOH

Example 3.4.2.6

Calculate the molar mass of $Ca(NO_3)_2$?

Solution

Step 1. Find the atomic masses of the constituent elements from the periodic table.

Ca = 40.078 $g.mol^{-1}$, N = 14.007 $g.mol^{-1}$, O = 15.999 $g.mol^{-1}$.

Step 2. Multiply the atomic masses with the number of atoms in the formula. (Note that there are two NO_3 units in $(NO_3)_2$?, To get the total number of atoms, multiply the subscript outside the bracket with the subscript to the element symbol inside the bracket to get the total number of atoms. That is, there are 1x2 = 2 N and 3x2 = 6 O in this compound.

1x40.078 g.mol^{-1} Ca, 2x14.007 g.mol^{-1} N, 6x15.999 g.mol^{-1} O.

Step 3. Add all the numbers from step 2.

1x40.078 g.mol^{-1} Ca + 2x14.007 g.mol^{-1}N + 6x15.999 g.mol^{-1} O = 164.008 g.mol^{-1} $Ca(NO_3)_2$.

Conversion from grams to moles and moles to grams of a substance

The molar mass in $\frac{g}{mol}$ is a conversion factor converting the amount of a substance in moles to the mass of the substance in grams. Reciprocal of the molar mass in $\frac{mol}{g}$ is a conversion factor converting the mass of a substance in grams to the amount of the substance in moles. The conversions are explained in the following examples.

Example 3.4.2.7

How many grams of water are in 2.50 moles of water?

Solution

Step 1. Write the given quantity and the desired quantity.

Given: 2.50 moles H_2O, Desired: ? g H_2O

Step 2. Calculate the molar mass of the substance.

2x1.008 g.mol^{-1} H + 15.999 g.mol^{-1} O = 18.02 g.mol^{-1} H_2O.

Step 3. Write the molar mass of the substance, and it's reciprocal as the two conversion factors.

$$\frac{18.02\ g\ H_2O}{1\ mol\ H_2O} \quad \text{and} \quad \frac{1\ mol\ H_2O}{18.02\ g\ H_2O}$$

Step 4. Multiply the given quantity with the conversion factor that cancels the given unit and leaves the desired unit in the answer.

$$2.50\ \cancel{mol\ H_2O} \times \frac{18.02\ g\ H_2O}{1\ \cancel{mol\ H_2O}} = 45.1\ g\ H_2O$$

Example 3.4.2.8

How many moles are present in 2.50 g of aspirin? (Aspirin $C_9H_8O_4$, molar mass 180.2 g.mol^{-1})

Solution

Step 1. Write the given quantity and the desired quantity.

Given: 2.50 g $C_9H_8O_4$, Molar mass of $C_9H_8O_4$ = 180 g.mol^{-1}, Desired: ? g H_2O

Step 2. Calculate the molar mass of the substance.: given 180.2 g.mol^{-1}

Step 3. Write the molar mass of the substance, and it's reciprocal as the two conversion factors.

$$\frac{180\ g\ C_9H_8O_4}{1\ mol\ C_9H_8O_4} \quad \text{and} \quad \frac{1\ mol\ C_9H_8O_4}{180\ g\ C_9H_8O_4}$$

Step 4. Multiply the given quantity with the conversion factor that cancels the given unit and leaves the desired unit in the answer.

$$2.5 \ \cancel{g \ C_9H_8O_4} \times \frac{1 \ mol \ C_9H_8O_4}{180 \ \cancel{g \ C_9H_8O_4}} = 0.0193 \ mol \ C_9H_8O_4$$

Example 3.4.2.9

How many moles of NaOH are in 10.0 g of NaOH?

Solution

Step 1. Write the given quantity and the desired quantity.

Given: 10.0 g NaOH, Desired: ?mol NaOH

Step 2. Calculate the molar mass of the substance.

22.99 g.mol^{-1} Na + 16.00 g.mol^{-1} O + 1.01 g.mol^{-1} H = 40.00 g.mol^{-1} NaOH.

Step 3. Write the molar mass of the substance and it's reciprocal as the two conversion factors.

$$\frac{40.00 \ g \ NaOH}{1 \ mol \ NaOH} \quad \text{and} \quad \frac{1 \ mol \ NaOH}{40.00 \ g \ NaOH}$$

Step 4. Multiply the given quantity with the conversion factor that cancels the given unit and leaves the desired unit in the answer.

$$10.0 \ \cancel{g \ NaOH} \times \frac{1 \ mol \ NaOH}{40.00 \ \cancel{g \ NaOH}} = 0.250 \ \text{mol NaOH}$$

This page titled 3.4.2: The mole is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.4.3: Chemical reaction

What is a chemical reaction?

> A chemical reaction is a combination, separation, or rearrangement of atoms in a substance.

Atoms are healed together in a substance by chemical bonds. The combination makes bonds, separation breaks bonds, and rearrangement breaks some of the old bonds and makes some new bonds in the substances during a chemical reaction. It results in new substances with a different composition of elements than the starting substances. The starting substances are called reactants and the new substances formed are called products.

When a chemical bond is broken energy is required and when a chemical bond is formed energy is released. The amount of energy depended on the chemical bond, but for the same bond, the energy that needs to break is the same energy released to make the bond. Therefore, energy is released in some chemical reactions and absorbed in others.

Examples of chemical reactions are photosynthesis which converts carbon dioxide and water into glucose in the green leaves of plants using energy from sunlight. Digestion of food is a chemical reaction that releases the energy needed for the functioning of living things. The burning of a candle is a chemical reaction that converts the organic compounds in the fule to carbon dioxide, water, and heat energy. Rusting iron is another chemical reaction that converts the element iron to compound iron oxide.

Indications of a chemical reaction

A chemical reaction is usually accompanied by some physical changes that can be observed. The changes include color change, flame, heat, light, the evolution of a gas, formation of a precipitate, etc., as illustrated in Fig. 4.3.1

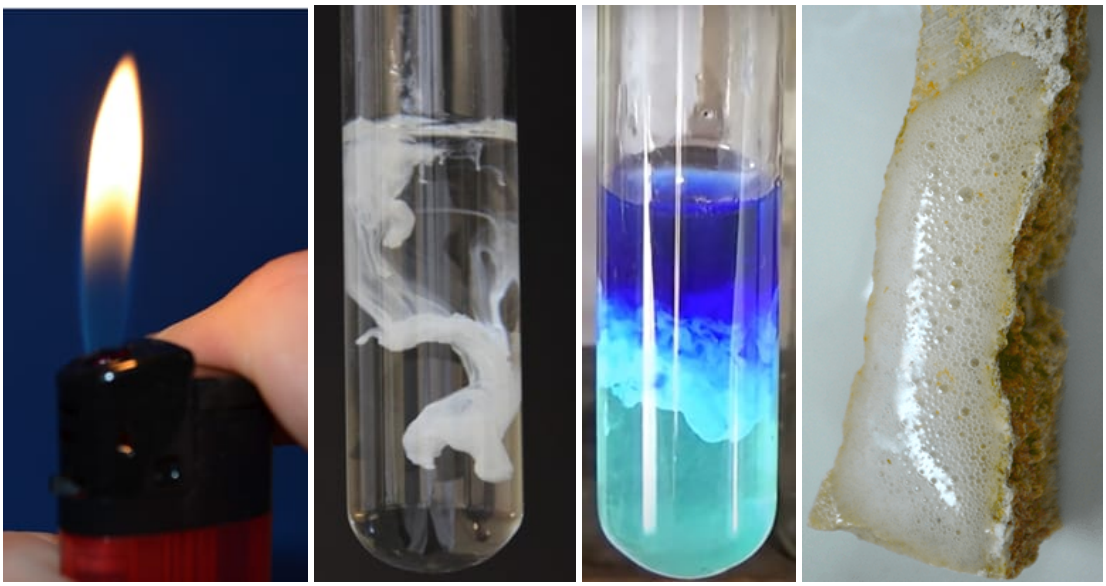


Figure 3.4.3.1: Examples of indications of chemical reactions: from left a) butane burning with flame and light, b) Silver chloride formation with precipitation, c) copper ammonia complex formation with a color change to greenish to dark blue, and d) Reaction of limestone with HCl with carbon dioxide gas formation. Source: a) Renate9Ø / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.Ø) , b) Luisbrudna / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.Ø), c) https://www.youtube.com/watch?v=7TIIOj-SHkw&t=1Ø2s, d) Alessandro e Damiano / CC BY (https://creativecommons.org/licenses/by/4.Ø)

Representing a chemical reaction- a chemical equation

Chemical equaitron

A chemical equation represents a chemical reaction.

1. The formulas of the reactants are written on the left side separated by a + sign.
2. Formulas of the products are written on the right-side separated by a + sign.
3. An arrow pointing in the direction of products separates the reactants from the products.

Fig. 4.3.2 illustrates the photosynthesis reaction between carbon dioxide (CO_2 and water (H_2O producing glucose ($C_6H_{12}O_6$ and oxygen (O_2:

$$6\,CO_2 + 6\,H_2O \xrightarrow{\text{sun light}} C_6H_{12}O_6 + 6\,O_2$$

Writing a chemical equation

1. First correct formulas of reactants and products are written, separated by plus signs and arrows.
2. Then the number of species is adjusted by adding a number, called a coefficient, at the beginning of a formula. The coefficients are needed to make atoms of each element equal on both sides of the equation because atoms are neither created nor destroyed in a chemical reaction.
3. Subscripts within the formulas can not be changed as they represent the composition of the substance that is constant.
4. The coefficient is not written if it is one.

For example, Fig. 4.3.2 shows coefficients 6 for carbon dioxide, water, and oxygen, but no coefficient for glucosemeans the coefficient is actually one for glucose in a balanced chemical equation for photosynthesis reaction.

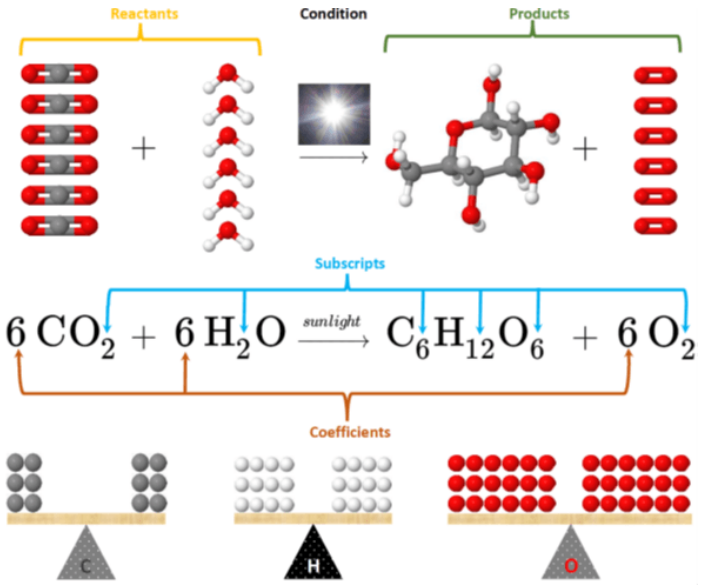


Figure 3.4.3.2: Model and components of a balanced chemical equation with the example of photosynthesis reaction.

Symbols in a chemical equation

Sometimes the physical state of the substances is shown by symbols in small brackets next to the formula.

1. The symbols for the physical state are: (s) for solid, (l) for liquid, (g) for gas, and (aq) for a substance dissolved in water. The symbol (aq) stands for aqueous, i.e., water.

2. If a gas evolves, it can be shown by an up-arrow (↑), and if a precipitate forms, it is shown by a down-arrow (↓).

For example, combustion of butane ($ceC4H10$ in lighter shown in Fig. 4.3.1. can be represented as:

$$2\,C_4H_{10}\,(g) + 13\,O_2\,(g) \longrightarrow 8\,CO_2\,(g) + 10\,H_{26}O\,(g)$$

, where all the reactants and products are in the gas phase. Reaction of NaCl and $AgNO_3$ in the water that results in a precipitate, as shown in Fig. 4.3.1 is:

$$NaCl\,(aq) + AgNO_3\,(aq) \longrightarrow AgCl\,(s)\downarrow + NaNO_3\,(aq)$$

, where NaCl, $AgNO_3$, and $NaNO_3$ are in the water, and AgCl precipitates out as white solid. The reaction of limestone with HCl, shown in Fig. 4.3.1 is:

$$CaCO_3\,(s) + 2\,HCl\,(aq) \longrightarrow CaCl_2\,(aq) + H_2O\,(l) + CO_2\,(g)\uparrow$$

Reaction conditions, catalysts, or heat can be written above or below the arrow. Heat can also be represented by Δ symbol, e.g.,

$$CaCO_3\,(s) + CaO\,(s) \xrightarrow{\Delta} CO_2(g)$$

A two way arrows $\rightleftharpoons$ or $\leftrightarrows$ represents a two ways reaction, e.g.,

$$2\,NO_2\,(g) \rightleftharpoons N_2O_4\,(g)$$

Balancing a chemical equation

The first step is wright correct formulas of reactants and products separated by plus signs and an arrow. For example, the initial equation for burning hydrogen burns in oxygen and producing water would be:

$$H_2 + O_2 \longrightarrow H_2O$$

Note that hydrogen is written as H_2 (not H) and oxygen as O_2 (not O) because these reactants usually exist as molecules, not as atoms.

The next step is to add coefficients to balance atoms of each element on the two sides of the equation. For example, in the above equation hydrogen is balanced but oxygen is not. Balance oxygen by changing the coefficient of water from 1 to 2:

$$H_2 + O_2 \longrightarrow 2\,H_2O$$

Caution

Subscripts in the formulae can not be changed, as they are constant. For example, if O_2 is changed to O in the above equation to balance the oxygen, it is incorrect as O_2 is molecular oxygen which is a different chemical than atomic oxygen O.

The coefficient is a multiplier of each subscript in the formula, i.e., in $2H_2O$ there are 2x2 = 4 hydrogen and 2x1 =2 oxygen. Now look for the other elements again: note that hydrogen atoms have changed to 4 on the right side. To balance hydrogen, change the coefficient of H_2 from 1 to 2:

$$2\,H_2 + O_2 \longrightarrow 2\,H_2O$$

Check again: Now, the atoms of each element are the same on both sides, i.e., the equation is balanced, as illustrated in Fig. 4.3.3.

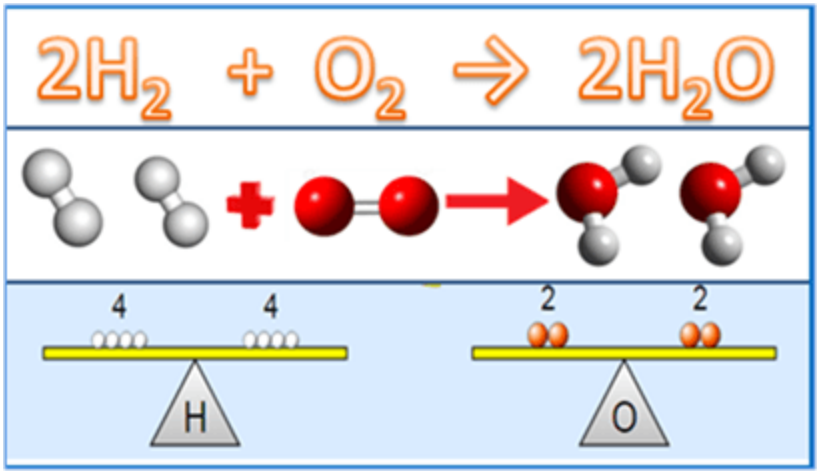


Figure 3.4.3.3: Illustration of a balanced chemical equation of water formation from hydrogen and oxygen. Atoms of each type are balanced. Source: Kvr.lohith / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Rules to balance a chemical equation

The method described above to balance a chemical equation is called the hit and trial method. There are no hard and fast rules in this method. The general guidelines are:

1. if an element occurs only in one reactant and one product, balance it first,
2. if there is a free element in the reactants or products, balanced it last,
3. if there is a polyatomic ion in reactants and products, balance the number of polyatomic ions as a unit, i.e., add coefficient to but do not change the subscripts in the polyatomic ion or of any formula.
4. in some cases, a fractional coefficient is needed to balance the equation, in that case first use the fractional coefficient and when the equation is balanced, remove the fraction by multiplying all the coefficients with a common multiplier,
5. if the set of coefficients in the balanced equation is not in the simplest whole-number ratio, the equation is still balanced, but it is recommended to convert the set of coefficients to the simplest whole-number ratio.

The following examples explain the guidelines.



Figure 3.4.3.4: Propane torch being used to solder copper pipes for residential water mains. Source: Kevin L Neff / CC BY (https://creativecommons.org/licenses/by/2.0)

Example 3.4.3.1

Combustion of propane (C_3H_8) is used in gas welding, as shown in Fig. 4.3.4. The propane reacts with oxygen and produces carbon dioxide and water. Write and balance the chemical equation of the reaction?

Solution

Start with writing the correct formulae of reactants and products in a chemical equation form:

$$C_3H_8 + O_2 \longrightarrow CO_2 + H_2O$$

According to rule 1, look at carbon or hydrogen first. Both are not balanced. Balance the carbon by changing the coefficient of CO_2 from 1 to 3:

$$C_3H_8 + O_2 \longrightarrow 3\,CO_2 + H_2O$$

Balance the hydrogen by changing the coefficient of H_2O from 1 to 4:

$$C_3H_8 + O_2 \longrightarrow 3\,CO_2 + 4\,H_2O$$

Now, look at oxygen (rule 2) and balance it by changing the coefficient of O_2 from 1 to 5:

$$C_3H_8 + 5\,O_2 \longrightarrow 3\,CO_2 + 4\,H_2O$$

The last equation is balanced, as illustrated in Fig. 4.3.5.

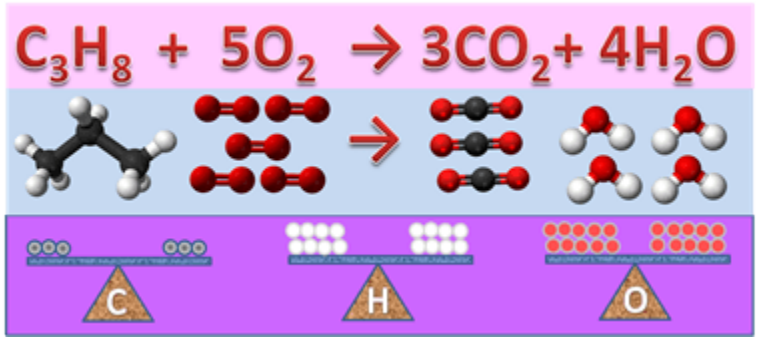


Figure 3.4.3.5: Illustration of a balanced chemical equation of propane combustion reaction. Source: Kvr.lohith / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Example 3.4.3.2

The reaction between Aluminium metal powder and iron oxide, called thermite reaction, is highly exothermic that melts the ion. The thermite reaction is used for railroad welding as shown in Fig. 4.3.6. Wright and balance the reaction equation.

Answer available at the end of this section.

Solution

Start with writing the correct formulae of reactants and products in a chemical equation form:

$$Al + Fe_2O_3 \longrightarrow Fe + Al_2O_3$$

Rule 1 does not apply as all elements are in one reactant and one product. According to rule 2, look at either one of the compounds. Balance aluminum by changing the coefficient in the eactant from 1 to 2:

$$2\,Al + Fe_2O_3 \longrightarrow Fe + Al_2O_3$$

Now balance iron by change its coefficient from 1 to 2 in the product:

$$2\,\mathrm{Al} + \mathrm{Fe_2O_3} \longrightarrow 2\,\mathrm{Fe} + \mathrm{Al_2O_3}$$

All atoms are balanced at this stage and coefficients are in the simplest whole-number ratio. So the equation is balanced.



Figure 3.4.3.1: Thermite reaction proceeding. Shortly after this, the liquid iron flows into the mold around the rail gap. Source: http://en.Wikipedia.org/wiki/Image:V...ewelding-1.jpg, Public domaine.

Example 3.4.3.3

Methanol (CH_4O) reacts with oxygen and produces carbon dioxide and water. Write and balance the chemical reaction equation?

Solution

The starting equation with correct formulas is:

$$\mathrm{CH_4O} + \mathrm{O_2} \longrightarrow \mathrm{CO_2} + \mathrm{H_2O}$$

According to rule 1, look at carbon and hydrogen first. Carbon is already balanced. There are 4 hydrogens on the left but only two on the right side, so balance them by changing the coefficient of water from 1 to 2:

$$\mathrm{CH_4O} + \mathrm{O_2} \longrightarrow \mathrm{CO_2} + 2\,\mathrm{H_2O}$$

Now, look at oxygen (rule 2). There are three oxygen atoms on the right (one in methanol and two in free element oxygen) but four oxygen on the right, it is not balanced. Changing the coefficient of CH_4O from 1 to 2 balances the oxygen, but carbon and hydrogen go out of balance. Changing the coefficient of O_2 from 1 to 2 does not balance the oxygen. A fractional coefficient 3/2 for O_2 works at this stage (rule 4):

$$\mathrm{CH_4O} + \tfrac{3}{2}\,\mathrm{O_2} \longrightarrow \mathrm{CO_2} + 2\,\mathrm{H_2O}$$

All elements are balanced, i.e., the equation is balanced, but, according to rule 5, it is better to remove the fraction by multiplying each coefficient in the equation with 2:

$$2\,\mathrm{CH_4O} + 3\,\mathrm{O_2} \longrightarrow 2\,\mathrm{CO_2} + 4\,\mathrm{H_2O}$$

The set of coefficients in the balanced equation is already in the simplest whole-number ratio, so no further step is needed in this case.

Example 3.4.3.4

Aluminum reacts sulfuric acid to produce aluminum sulfate and hydrogen gas. Wright a balanced chemical equation for the reaction.

Solution

The starting equation with correct formulas of reactants and products is:

$$\mathrm{Al} + \mathrm{H_2SO_4} \longrightarrow \mathrm{Al_2(SO_4)_3} + \mathrm{H_2}$$

All elements occur in one reactant and one product, so rule 1 do not apply. According to rule 2, leave aluminum till the end and, according to rule 3, balance the polyatomic ion (SO_4^{2-}) as a unit by changing the coefficient of sulfuric acid from 1 to 3:

$$\mathrm{Al} + 3\,\mathrm{H_2SO_4} \longrightarrow \mathrm{Al_2(SO_4)_3} + \mathrm{H_2}$$

Now balance the hydrogen by changing it coefficient from 1 to 3:

$$\mathrm{Al} + 3\,\mathrm{H_2SO_4} \longrightarrow \mathrm{Al_2(SO_4)_3} + 3\,\mathrm{H_2}$$

Finally balance aluminum by changing its coefficient from 1 to 2:

$$2\,\mathrm{Al} + 3\,\mathrm{H_2SO_4} \longrightarrow \mathrm{Al_2(SO_4)_3} + 3\,\mathrm{H_2}$$

Double-check at the end –All elements are balanced, and all the coefficients are in the simplest whole-number ratio. No further step is needed.

Answers

Example 3.4.3.2

This page titled 3.4.3: Chemical reaction is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.4.4: Patterns of chemical reactions

General types of chemical reactions

There are several ways to classify chemical reactions. The general types of chemical reactions fall in the categories of combination, decomposition, replacement, and combustion reactions, as illustrated in Fig. 4.4.1.

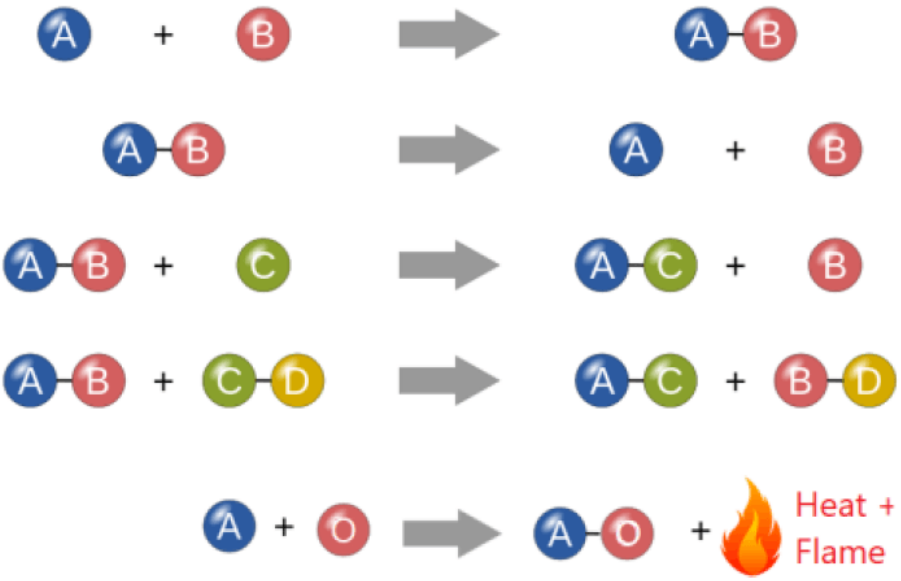


Figure 3.4.4.1: Illustration of four basic patterns of chemical reactions: combination, decomposition, single & double replacement, and combustion reactions. Source: https://www.hiclipart.com/free-trans...dunqu/download

Combination reactions

A compound is synthesized or formed from two or more substances, e.g.:

$$\mathrm{C + O_2 \longrightarrow CO_2}$$

$$\mathrm{2\,H_2 + O_2 \longrightarrow 2\,H_2O}$$

$$\mathrm{2\,Mg + O_2 \longrightarrow 2\,MgO}$$

$$\mathrm{2\,Na + Cl_2 \longrightarrow 2\,NaCl}$$

$$\mathrm{CaO + CO_2 \longrightarrow CaCO_3}$$

Fig. 4.4.2 shows an example of a hydrogen with oxygen combination reaction that is being developed for use as a fuel in the future.



Figure 3.4.4.2: Berlin public bus using hydrogen as a fuel in an internal combustion engine (ICE). Hydrogen is stored in ten pressure cylinders of 50 kg of hydrogen at 350 bar. Reaction: \{\ce{2H2 + O2 -> 2H2O}\) + heat. Source: StralsundByzantion / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Decomposition reactions

The decomposition reactions are the reverse of the combination reaction, i.e., one compound splits apart into two or more substances, usually by heating, e.g.:

$$\mathrm{H_2CO_3} \longrightarrow \mathrm{H_2O + CO_2}$$

$$\mathrm{CaCO_3} \xrightarrow{\Delta} \mathrm{CaO + CO_2}$$

$$\mathrm{2\,KClO_3} \xrightarrow{\Delta} \mathrm{2\,KCl + 3\,O_2}$$

$$\mathrm{2\,H_2O\,(l)} \xrightarrow{\text{Electrolyisis}} \mathrm{2\,H_2\,(g) + O_2\,(g)}$$

Fig. 4.4.3 illustrates the last reaction, i.e., decomposition of water through electrolysis.

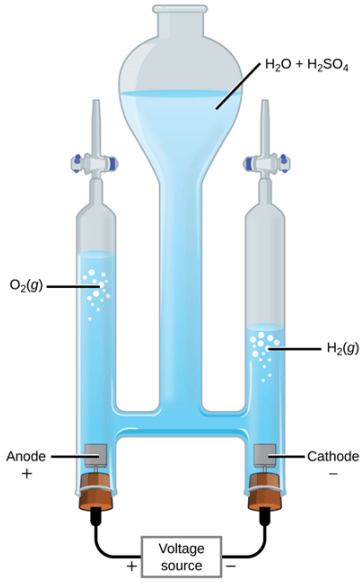


Figure 3.4.4.3: Decomposition of water through electrolysis $\mathrm{2\,H_2O\,(l)} \xrightarrow{\text{Electrolyisis}} \mathrm{2\,H_2\,(g) + O_2\,(g)}$. Source: OpenStax / CC BY (https://creativecommons.org/licenses/by/4.0)

Replacement or substitution reactions

There are two sub-classes of this category of reactions, i.e., single replacement and double replacement reaction.

Single replacement reactions involve one substance replacing a part of another, e.g.:

$$\mathrm{Zn\,(s) + CuCl_2\,(aq) \longrightarrow ZnCl_2\,(aq) + Cu\,(s)}$$

$$\mathrm{2\,Al\,(s) + 6\,HCl\,(aq) \longrightarrow AlCl_3\,(aq) + 3\,H_2\,(g)}$$

Fig. 4.4.4 shows an example of a single replacement reaction of magnesium resulting in hydrogen gas formation.



Figure 3.4.4.4: Single replacement reaction of magnesium with hydrochloric acid producing hydrogen gas: $\mathrm{Mg\,(s) + 2\,HCl\,(aq) \longrightarrow MgCl\,(aq) + H_2\,(g)}$. Source: taken form NCSSM online youtube video, 05/18/20, https://www.youtube.com/watch?v=OBdgeJFzSec

Double replacement reactions or metathesis involve the mutual exchange of partners between two substances, e.g. the following precipitation reactions:

$$\mathrm{NaCl\,(aq) + AgNO_3\,(aq) \longrightarrow AgCl(s)\,(v) + NaNO_3\,(aq)}$$

$$\mathrm{Na_2CO_3\,(aq) + CaCl_2\,(aq) \longrightarrow CaCO_3(s)\,(v) + 2\,NaCl\,(aq)}$$

Combustion reactions

Combustion is a reaction of a substance with oxygen, often with the formation of flame and release of much heat, e.g.:

$$\mathrm{C_8H_{16} + 12\,O_2 \longrightarrow 8\,CO_2 + 8\,H_2O + Heat}$$

$$\mathrm{C + O_2 \longrightarrow CO_2 + Heat}$$

$$\mathrm{2\,H_2 + O_2 \longrightarrow 2\,H_2O + Heat}$$

$$\mathrm{2\,Mg + O_2 \longrightarrow 2\,MgO + Heat}$$

Fig. 4.4.5 shows the above reaction, i.e., combustion of Mg in air.



Figure 3.4.4.5: An image of Ignited Magnesium burning in the air in normal conditions. Yannickcosta1, CC BY SA3, https://commons.wikimedia.org/wiki/File:Magburn1.jpg

Usually, combustion is considered as the reaction of a substance containing carbon and hydrogen with oxygen resulting in carbon dioxide, water, flame, and heat, e.g., burning methane on a kitchen stove:

$$\mathrm{CH_4 + 3\,O_2 \longrightarrow CO_2 + 2\,H_2O + Heat}$$

Classification of chemical reaction

The chemical reactions are generally classified based on what exchanges during the reaction. These include;

1. the transfer of electrons in oxidation-reduction reactions,
2. transfer of protons in acid-base reactions, and
3. a part of reactants mutually exchanges in precipitation reactions, as described below.

Oxidation-reduction reactions

The oxidation-reduction or **redox** reaction involves the exchange of electrons. For example, reactions between a metal and nonmetal involve the transfer of electrons from the metal to the nonmetal forming an ionic bond, as shown in Fig. 4.4.6.

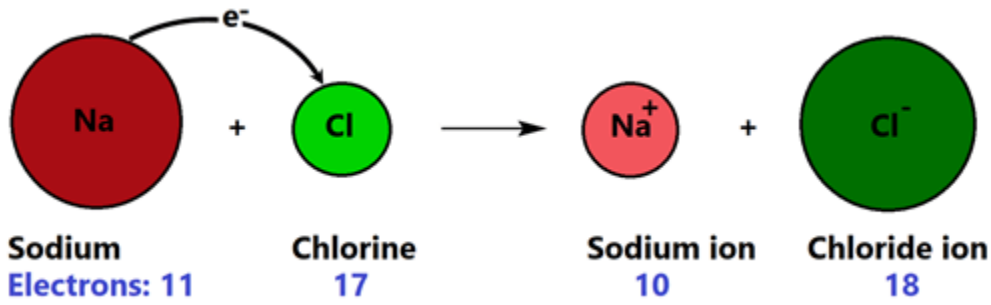


Figure 3.4.4.6: Illustration of an electron transfer in an oxidation-reduction reaction: $\mathrm{Na + Cl \longrightarrow Na^+ + Cl^-}$

Acid-base reactions

The acid-base reactions involve the transfer of protons from an acid to a base, as shown in Fig. 4.4.7.



Figure 3.4.4.7: Proton transfer from HCl to NH_3 is an example of an acid-base reaction: $\mathrm{HCl} + \mathrm{NH_3} \longrightarrow \mathrm{NH_4Cl}$. Hydrochloric acid (in beaker) reacts with ammonia fumes to produce ammonium chloride (white smoke). Source:
Walkerma, Public domain

Precipitation reactions

These are double displacement reactions in water that results in the precipitation of one of the products, as shown in Fig. 4.4.8.



Figure 3.4.4.8: Lead (II) iodide precipitates when potassium iodide is mixed with lead (II) nitrate: $\mathrm{Pb(NO_3)_2\,(aq) + 2\,NaI\,(aq) \longrightarrow 2\,NaNO_3\,(aq) + PbI_2(s)\,(v)}$. Source: PRHaney / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

The precipitation reactions and the acid-base reactions are described in the later chapters. The oxidation-reduction reactions are discussed in the following section.

This page titled 3.4.4: Patterns of chemical reactions is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.4.5: Oxidation-reduction reactions

What are Oxidation and Reduction?

The oxidation-reduction is defined in three ways:

1. **O**xidation **i**s the **l**oss of electrons and **r**eduction **i**s the **g**ain of electrons. The word **OIL RIG** helps in remembering this definition.
2. The addition of oxygen is oxidation, and the removal of oxygen is a reduction.
3. The removal of hydrogen is oxidation, and the addition of hydrogen is a reduction.

The oxidation-reduction or in short **redox** reaction is one of the most common types of chemical reactions happening in and around us. For example, rusting of metals, photosynthesis, digestion of food, and combustion of fuels are redox reactions.



Figure 3.4.5.1: Green patina on the statue of liberty is a result of the oxidation of copper. Source: https://www.hiclipart.com/free-trans...hybox/download

Oxidation-half and reduction-half reactions

The oxidation-reduction reactions happen in a pair because one thing loses electrons and is oxidized; the other thing gains the electrons and is reduced. That is why it is commonly called redox reaction, where red- represents reduction and –ox represents oxidation.

The green patina on the statue of liberty, shown in Fig. 4.5.1, is a result of the oxidation of copper:

$$2\,\mathrm{Cu\,(s)} + \mathrm{O_2\,(g)} \longrightarrow \mathrm{CuO\,(s)}$$

The transfer of electrons becomes apparent when the reaction is split into the oxidation-half and reduction-half. The oxidation-half is:

$$\mathrm{Cu} \longrightarrow \mathrm{Cu^{2+}} + 2\,\mathrm{e^-}$$

copper lost electrtons, so copper is oxides. The reduction half is:

$$\mathrm{O_2} + 4\,\mathrm{e^-} \longrightarrow 2\,\mathrm{O^{2-}}$$

Oxygen gained electrons, so oxygen is reduced.

Oxidizing agent and reducing agent

- The substance that oxidizes another substance is an **oxidizing agent**. Oxygen is an oxidizing agent in the above reaction, as it oxidizes copper.
- The substance that reduces another substance is a **reducing agent**. Copper is a reducing agent in the above reaction, as it reduces oxygen.

The electrons lost in an oxidation-half must be equal to the electrons gained by the accompanying reduction-half. Multiplying the oxidation-half with 2 makes the electrons lost equal to the electrons gained in the reduction half in the above reaction. Then adding the oxidation and the reduction half gives the overall reaction:

$$2\,\mathrm{Cu} \longrightarrow 2\,\mathrm{Cu^{2+}} + \cancel{4e^-}$$

$$\mathrm{O_2} + \cancel{4e^-} \longrightarrow 2\,\mathrm{O^{2-}}$$

$$\text{Overall reaction: } 2\,\mathrm{Cu\,(s)} + \mathrm{O_2\,(g)} \longrightarrow \mathrm{CuO\,(s)}$$

Manipulating a chemical equation

The chemical equations can be manipulated like algebraic equations, i.e., they can be multiplied or divided by a constant, added, and subtracted, as demonstrated in the example of the copper redox-half reactions above. Note that electrons on one side of the equation canceled the electrons on the other side of the equation during the addition operation.

A silver color zinc strip dipped in copper nitrate solution becomes coated with a layer of reddish color copper, as shown in Fig. 4.5.2. The molecular equation of the reaction, that shows formula units of compounds in the reactants and products, is:

$$\mathrm{Zn\,(s)} + \mathrm{Cu(NO_3)_2\,(aq)} \longrightarrow \mathrm{Cu\,(s)} + \mathrm{Zn(NO_3)_2\,(aq)}$$

The complete ionic equation for the reaction is obtained by showing the dissolved ionic compounds as ions, e.g.:

$$\mathrm{Zn\,(s)} + \mathrm{Cu^{2+}\,(aq)} + 2\,\mathrm{NO_3{}^-\,(aq)} \longrightarrow \mathrm{Cu\,(s)} + \mathrm{Zn^{2+}\,(aq)} + 2\,\mathrm{NO_3{}^-\,(aq)}$$

The oxidation-half of the reaction is:

$$\mathrm{Zn\,(s)} \longrightarrow \mathrm{Zn^{2+}\,(aq)} + 2\,\mathrm{e^-}$$

The reduction-half of the reaction is:

$$\mathrm{Cu^{2+}\,(aq)} + 2\,\mathrm{e\text{-}} \longrightarrow \mathrm{Cu\,(s)}$$

The net ionic equation that is the addition of the oxidation-half and the reduction-half is:

$$\mathrm{Zn\,(s)} + \mathrm{Cu^{2+}\,(aq)} \longrightarrow \mathrm{Cu\,(s)} + \mathrm{Zn^{2+}\,(aq)}$$

Note that NO_3^- was on both sides of the complete ionic equation and has been canceled out in the net ionic equation:

$$\mathrm{Zn\,(s)} + \mathrm{Cu^{2+}\,(aq)} + \cancel{2NO3^-(aq)} \longrightarrow \mathrm{Cu\,(s)} + \mathrm{Zn^{2+}\,(aq)} + \cancel{2NO3^-(aq)}$$

Ions that do not participate in the chemical reaction are called spectator ions, and they appear on both sides of the molecular equation, like NO_3^- in this case.

Biological oxidation and reduction reactions

In the redox reactions involving metal species, the transfer of electrons is usually evident through the oxidation-half and the

reduction-half reactions, as in the above examples. In the organic and biochemical redox reactions, the transfer of electrons is usually not so obvious, but the transfer of hydrogen or oxygen is usually apparent. For example, the metabolism of methanol $\mathrm{H_3C{-}OH}$ starts with oxidation through the loss of hydrogen:

$$\mathrm{H_3C{-}OH \longrightarrow H_2C{=}O + 2\,H}$$

Metanol is oxidized to formaldehyde $\mathrm{H_2C{=}O}$. Formaldehyde oxidizes further by gaining oxygen:

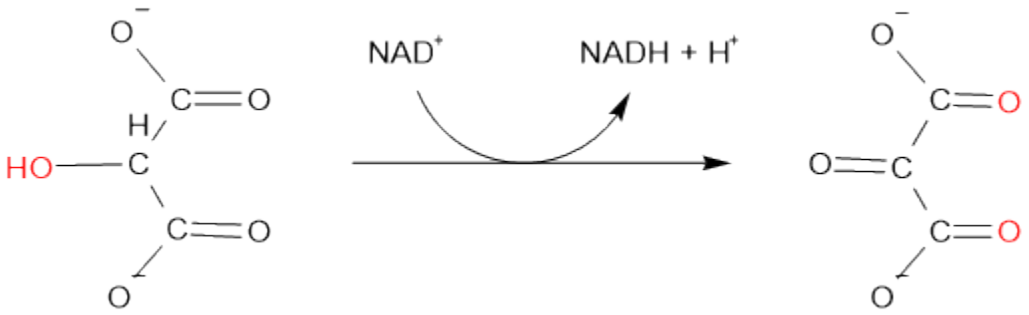
$$\mathrm{H_2C{=}O \longrightarrow HCOOH}$$

Finally, the formic acid (HCOOH) is oxidized by gaining oxygen and forming carbon dioxide and water:

$$\mathrm{2\,HCOOH \longrightarrow 2\,O{=}C{=}O + 2\,H_2O}$$

The increase in the C-O bond from a single bond in methanol to two double bonds (four C-O bonds) in carbon dioxide is a clear indication of the oxidation of the carbon.

Removal of hydrogen is also oxidation, e.g., hydroxymalonate is oxidized to oxomalonate by an enzyme hydroxymalonate dehydrogenase:



The reverse of these, i.e., removal of oxygen and addition of hydrogen is reduction.

This page titled 3.4.5: Oxidation-reduction reactions is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.4.6: Energetics of chemical reactions

Energetics deals with the energies involved in chemical reactions. There are two types of energies: the first is related to thermodynamics which is the energy released or absorbed when the reactants convert to the products, and the second is related to the kinetics of the reaction, i.e., the energy that the reacting molecules must possess to surpass the energy barrier for conversion to the products.

Reaction coordinate diagram

The progress of the reaction versus energy graph is shown in Fig. 4.7.1.

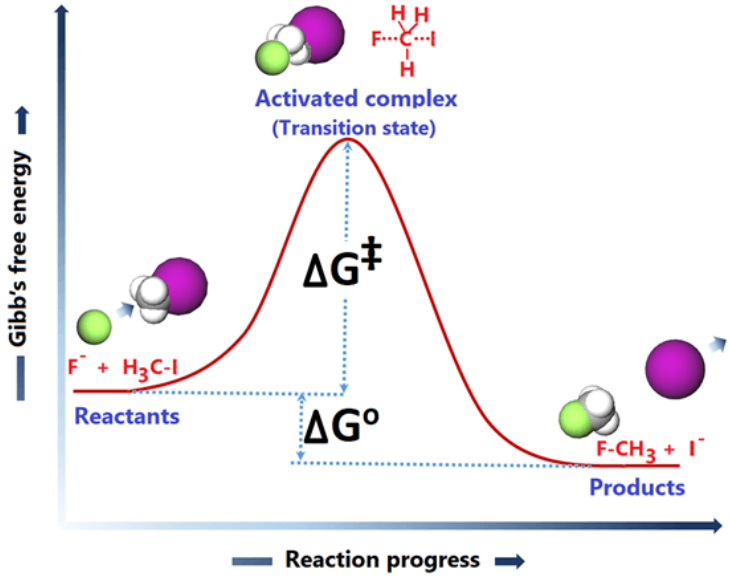


Figure 3.4.6.1: A reaction coordinate diagram for the reaction: $F^- + H_3C{-}I \longrightarrow F{-}CH_3 + I^-$. The green sphere is F^-, purple is I^-, gray is C, and white is H. $\Delta G^{\ddagger}$ is the activation energy, and ΔG^{o} is the overall energy released or absorbed in a chemical reaction.

The starting point shows the energy level of reactants. Remember that every substance has internal energy which is manifested as temperature. The molecules are moving around due to the internal energy and collied with other molecules. The collision between the reactants is usually the first requirement for the reaction to proceed.

The F^- ion in this example collides with the target H_3C-I from the side opposite to the C-I bond. The collision at an angle other than 180° to the C-I bond is less effective or ineffective in initiating the C-I bond breaking. So, the proper orientation of the reacting molecules at the time of the collision is the second requirement.

After a properly oriented collision, the C-I bond starts breaking, the F-C bond starts making, and the energy of the system rises to a maximum level where the bond breaking and making are about halfway. The specie at the maxima of the energy curve is called **activated complex** or a **transition state**. The difference in energy between the reactants and the transition state is called activation energy ($\Delta G^{\ddagger}$). The reactants must have enough kinetic energy to surpass the activation energy barrier, which is the third requirement for the reaction to happen. A summary of the basic requirements of a chemical reaction to happen is the following.

Requirement of a chemical reaction

1. The reacting molecules collide with each other –the more frequent the collision, the faster the reaction.
2. The molecules should have proper orientation at the time of the collision –this factor varies from reaction to reaction.

3. The colliding molecules must have energy, in addition to the potential chemical energy, equal to or higher than the energy of activation –the lower the activation energy, the faster the reaction.

The activated complex forms if the basic requirements are fulfilled. The activated complex rolls down the hill on the energy scale and settles at the energy level of the products where the old bonds are entirely broken, and the new bonds are fully formed. The graph of the reaction progress versus the energy, as in Fig. 4.7.1 is called the **reaction coordinate diagram**.

Exothermic and endothermic reactions

The energy absorbed or released in the form of heat at constant pressure conditions is called enthalpy change (ΔH), which is almost the same as the internal energy absorbed or released (ΔG) in most of the reactions.

A chemical reaction that releases heat is exothermic. A chemical reaction that absorbs heat is an endothermic reaction.

Bond-forming is always exothermic, and bond-breaking is the opposite, i.e., endothermic, as illustrated in Fig. 4.7.2.

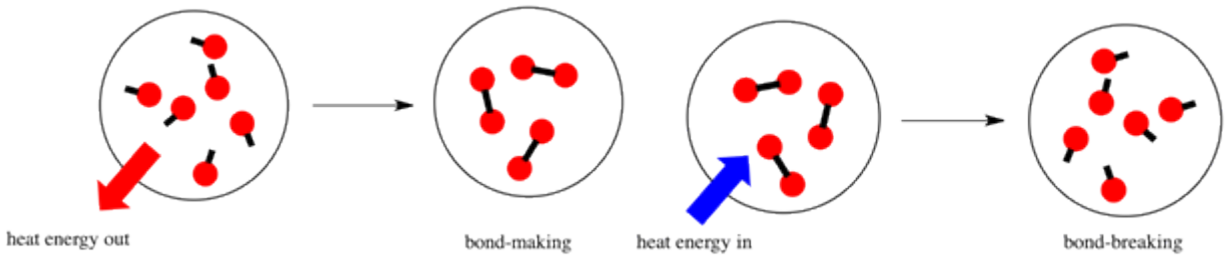


Figure 3.4.6.2: Bond forming releases heat, i.e., exothermic, and breaking the same bond absorbed the same energy, i.e., endothermic. Source: https://www.hiclipart.com/free-trans...alsen/download

In a chemical reaction, some bonds break, and some bonds form. If the bond-breaking absorbs less heat than the heat released in the new bond making, the reaction is exothermic. The products are lower in energy than the reactants, and the ΔH is negative for exothermic reactions. The opposite is true for an endothermic reaction, i.e., the bond-breaking absorbs more heat than the bond-forming; the products are higher in energy than the reactants, and ΔH is positive, as illustrated in Fig. 4.8.3.

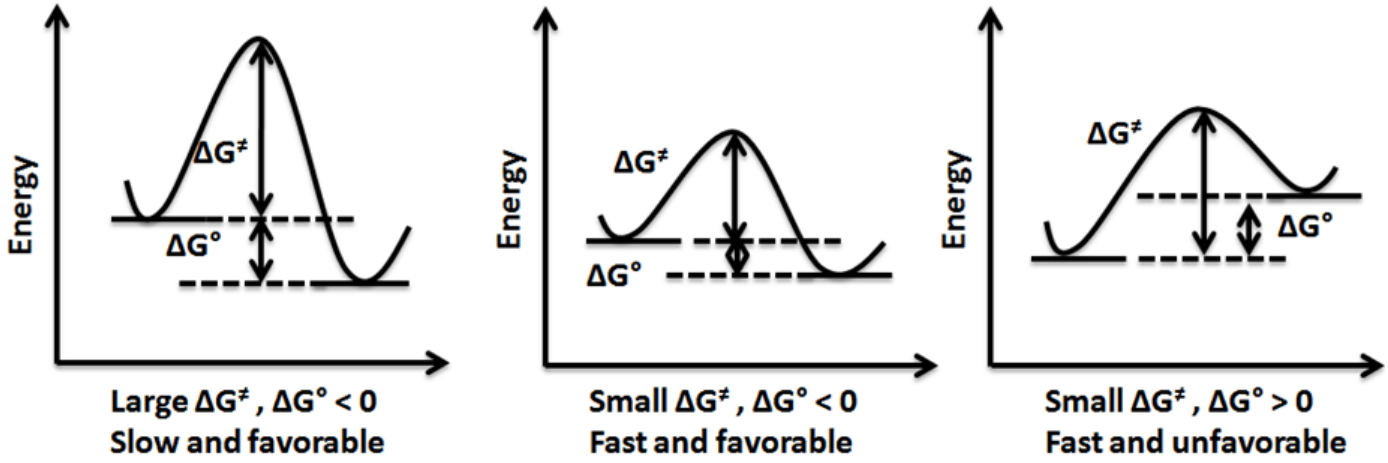


Figure 3.4.6.3: Reaction Coordinate Diagrams showing exothermic with products lower in energy than the reactants (left and middle) and endothermic with products higher in energy than the reactant (right) reaction. It also shows that the reaction on the left is the slowest due to the highest activation energy ($\Delta G^{\ddagger}$), and the reaction in the middle is the fastest due to the lowest activation energy ($\Delta G^{\ddagger}$). Source: AimNature / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

An example of an exothermic reaction is the combustion of methane that releases heat used for cooking.

$$CH_4 + 2\,O_2 \longrightarrow CO_2 + 2\,H_2O \quad \Delta H^o = -891\,kJ$$

An example of an endothermic reaction is photosynthesis.

$$6\,CO_2 + 6\,H_2O \longrightarrow C_6H_{12}O_6 + 6\,O_2 \quad \Delta H^o = +2800\,kJ$$

Note that the sign of ΔH is –ve for exothermic and +ve for an endothermic reaction. Sometimes, the energy is shown as a product for exothermic reaction, and as reactant for edothermic reaction. For example, the above two creations may be written as:

$$CH_4 + 2\,O_2 \longrightarrow CO_2 + 2\,H_2O + 891\,kJ$$

$$6\,CO_2 + 6\,H_2O + 2800\,kJ \longrightarrow C_6H_{12}O_6 + 6\,O_2$$



Figure 3.4.6.4: Cold Hot Pack used for medication. Source: Mamun2a, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

The instant hot or cold packs used in hospitals are based on the heat of dissolution of salts in water, as shown in Fig. 4.7.4. The salt and the water are in separate pockets in the pack. When the seal is ruptured, the salt dissolves in water and releases or absorbs heat. The dissolution of $CaCl_2$ salt in water is an exothermic process that is the source of heat in a hot pack:

$$CaCl_2\,(s) \longrightarrow CaCl_2\,(aq) \quad \Delta H^o = -82\,kJ$$

The dissolution of NH_4NO_3 salt in water is an endothermic process that is utilized to absorb heat in a cold-pack.

$$NH_4NO_3\,(s) \longrightarrow NH_4NO_3\,(aq) \quad \Delta H^o = -26\,kJ$$

Factors that affect the rate of a chemical reaction

The rate of a chemical reaction, i.e., the amount of a reactant consumed or the amount of a product formed per unit time, is dependent on the energy of activation of the reaction —the higher the activation energy, the slower the rate of a reaction and vice versa. A reaction requires collision between the reactant molecules, with proper orientation, and sufficient energy to surpass the activation energy barrier.

> Any factor that increases the rate of collisions, enhances the proper orientation or increases the kinetic energy of the molecules causes an increase in the rate of reaction. The factors include concentration, temperature, and catalysts.

Effect of concentration of reactants

> The higher the concentration, the more frequent the collisions, and the faster the reaction, as illustrated in Fig. 4.7.5.

In the breathing process, oxygen (O_2) binds with hemoglobin (Hb) in the lungs.

$Hb(aq) + O_2(g) \rightarrow HbO_2(aq)$

Patients having breathing problems are given breathing masks with a higher concentration of oxygen than in the atmosphere to increase the rate of oxygen binding with the hemoglobin.

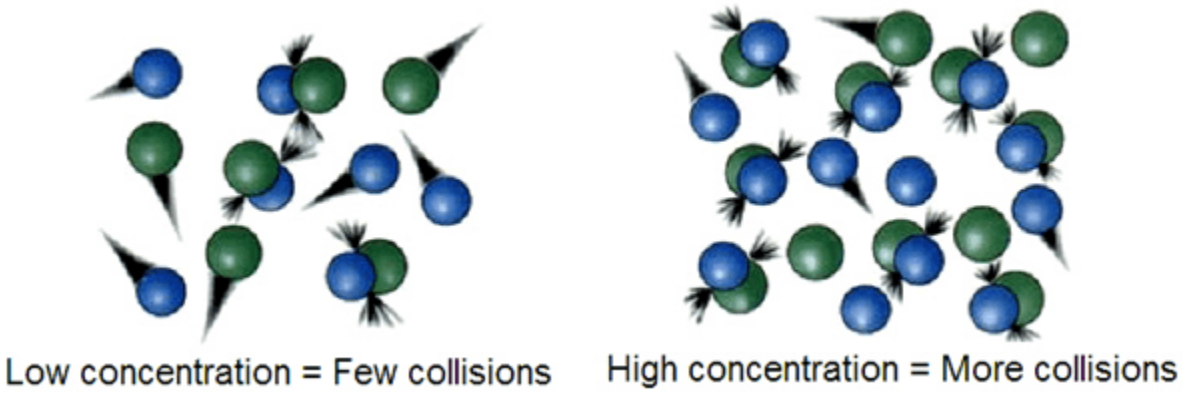


Figure 3.4.6.5: Illustration of higher the concentration, the more frequent the collisions leading to faster reaction. Source: Sadi Carnot / Public domain

Effect of temperature

The kinetic energy of molecules at a given temperature follows Boltzman distribution, as illustrated in Fig. 4.8.6. Increased temperature increases the average kinetic energy of the molecules which increases the fraction of molecules with more than the activation energy.

> An increase In temperature increases the rate of chemical reactions—generally, a 10 ^{o}C increase in temperature doubles the rate of a chemical reaction.

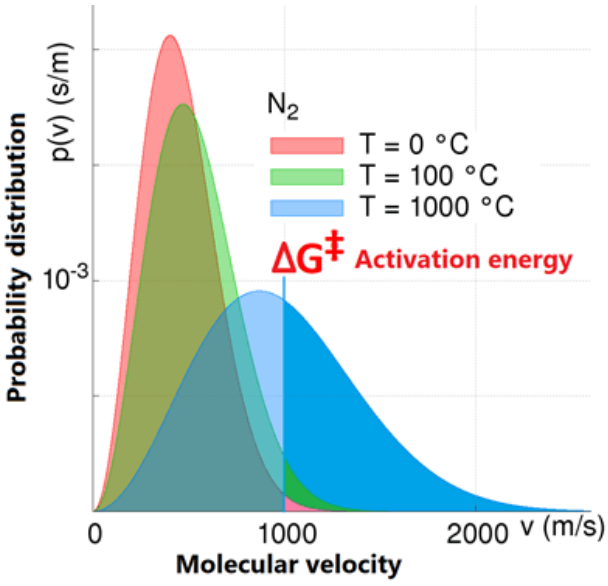


Figure 3.4.6.6: Illustrating increased temperature increases the average kinetic energy of the molecules that increases the fraction of molecules with more than the activation energy barrier ($\Delta G^{\ddagger}$). Source: modified from “TDF / CC0”

Some practical examples of the use of this principle are the following. Storing foods in refrigerators at lower temperatures decreases the rate of reactions resulting in longer life of the foods. Cooking foods in a pressure cooker increases the temperature

resulting in faster cooking than in an open pan. In some cardiac surgeries, the body temperature is lowered to 28 ^{o}C to decrease the rate of metabolism which decreases the oxygen demand so that the heart may be stopped temporarily for the surgery.

Effect of catalysts and enzymes

The catalysts and enzymes increase the rate of reaction by decreasing the energy of activation of the reaction through an alternate route, as illustrated in Fig. 4.7.7.

The catalysts and the enzymes do not consume in the reaction -they regenerate and repeat the action. Enzymes are the catalysts in biochemical reactions. Enzymes also increase the rate of reaction by binding with the reactants and properly orienting them for the reaction.

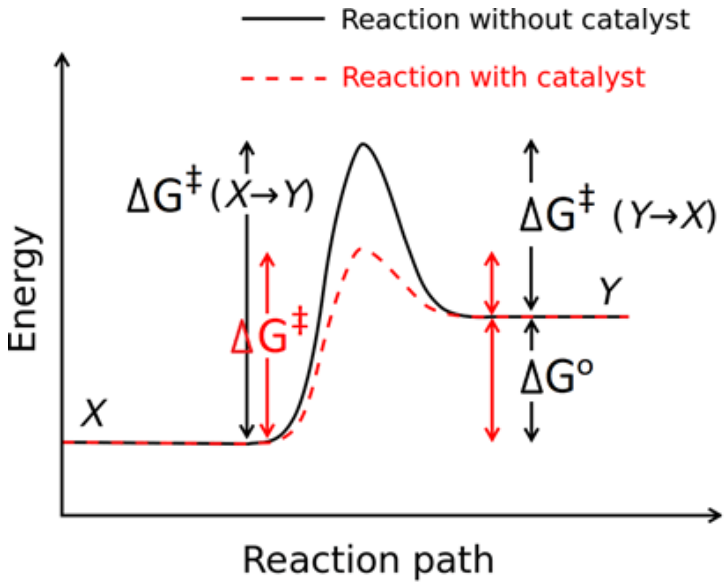


Figure 3.4.6.1: Effect of catalysts –When the catalyst is added, the activation energy $\Delta G^{\ddagger}$ decreases. However, the energy absorbed or released, i.e., ΔG^{o} remains constant. Source: Bkell, Public domain.

This page titled 3.4.6: Energetics of chemical reactions is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.4.7: Stoichiometric calculations

Conversion factors from a chemical equation

Stoichiometry pronounced as “stɔɪkiˈɒmɪtri” is the calculation of the amount of reactants and products in a chemical reaction. It is based on the fact that a balanced chemical equation is also a set of mole-to-mole equalities between the reactants and the products. Each equality gives two conversion factors that allow calculating the mole of one substance from the given mole of any other substance in the equation.

Fig. 4.6.1 lists the chemical equation for photosynthesis reaction, the mole-to-mole equalities from the equation, and the two conversion factors from each of the equality, as an example. The conversion factors are used to calculate the unknown quantity in the mole from the known quantity in the mole of any other reactant or product in the same chemical equation, as explained in the following examples.

Equation	$6CO_2+6H_2O \rightarrow C_6H_{12}O_6 + 6O_2$		
Equalities	6 mol CO_2 = 6 mol H_2O	6 mol CO_2 = 1 mol $C_6H_{12}O_6$	6 mol CO_2 = 6 mol O_2
Conversion factors	$\frac{6\ mol\ CO_2}{6\ mol\ H_2O}$ and $\frac{6\ mol\ H_2O}{6\ mol\ CO_2}$	$\frac{6\ mol\ CO_2}{1\ mol\ C_6H_{12}O_6}$ and $\frac{1\ mol\ C_6H_{12}O_6}{6\ mol\ CO_2}$	$\frac{6\ mol\ CO_2}{6\ mol\ O_2}$ and $\frac{6\ mol\ O_2}{6\ mol\ CO_2}$
Equalities	6 mol H_2O = 1 mol $C_6H_{12}O_6$	6 mol H_2O = 6 mol O_2	1 mol $C_6H_{12}O_6$ = 6 mol O_2
Conversion factors	$\frac{6\ mol\ H_2O}{1\ mol\ C_6H_{12}O_6}$ and $\frac{1\ mol\ C_6H_{12}O_6}{6\ mol\ H_2O}$	$\frac{6\ mol\ H_2O}{6\ mol\ O_2}$ and $\frac{1\ mol\ C_6H_{12}O_6}{6\ mol\ O_2}$	$\frac{1\ mol\ C_6H_{12}O_6}{6\ mol\ O_2}$ and $\frac{6\ mol\ O_2}{1\ mol\ C_6H_{12}O_6}$

Figure 3.4.7.1: Balanced chemical equation of the photosynthesis reaction, mol-to-mol equalities, and conversion factors derived from the equation.

Mole-to-mole conversion from a chemical equation

Calculation of the mole of the desired substance from the given mole of any reactant or product needs one conversion factor. The steps are:

1. write the given quantity and the desired quantity,
2. write the balanced chemical equation,
3. write the equality between the given and the desired substances,
4. right the conversion factor that has the given substance in the denominator and the desired substance in the numerator,
5. multiply the given quantity with the conversion factor. Double-check to make sure that it cancels the given substance and leaves the desired substance in the answer.

Example 3.4.7.1

Calculate the moles of glucose produced from 3.0 moles of carbon dioxide in the photosynthesis reaction?

Solution

i. Given: 3.0 mole CO_2, Desired: ? moles of $C_6H_{12}O_6$

ii. Chemical equation: $6\,CO_2 + 6\,H_2O \longrightarrow C_6H_{12}O_6 + 6\,O_2$

iii. The desired equality: 6 mol CO_2 = 1 mol $C_6H_{12}O_6$,

iv. Desired conversion factor: $\frac{1\text{ mol } C_6H_{12}O_6}{6\text{ mol } CO_2}$

v. Calculations:

$$3.0\ \cancel{\text{mol CO}_2} \times \frac{1\text{ mol } C_6H_{12}O_6}{6\ \cancel{\text{mol CO}_2}} = 0.50\text{ mol } C_6H_{12}O_6$$

Example 3.4.7.2

Magnesium reacts with HCl by this reaction: $Mg\,(s) + 2\,HCl\,(aq) \longrightarrow MgCl_2\,(aq) + H_2(g)$. Calculate the moles of Mg needed to produce 3 moles of H_2?

Solution

i. Given: 3 mol of H_2. Desired: ? Moles of Mg.

ii. The chemical equation is given.

iii. The desired equality: 1 mol Mg = 1 mol H_2.

iv. The desired conversion factor: $\frac{1\ mol\ Mg}{1\ mol\ H_2}$

v. Calculation:

$$3\ \cancel{\text{mol H}_2} \times \frac{1\text{ mol Mg}}{1\ \cancel{\text{mol H}_2}} = 3\ \text{mol}\ \ Mg$$

Mole-to-mass conversion from a chemical equation

Calculations described in the previous two examples show calculating moles of the desired substance from the moles of the given substance using a conversion factor from a chemical equation. The molar mass is a conversion factor from a mole-to-gram of the substance. So, add a molar mass of the desired substance as a second conversion factor, as explained in the following examples. Double-check that all the units cancel out, leaving the mass unit of the desired substance in the final answer.

Example 3.4.7.3

Calculate grams of AgCl precipitate formed from 2.0 moles of $CaCl_2$ consumed in the following reaction:

$$2\,AgNO_3 + CaCl_2 \longrightarrow Ca(NO_3)_2 + 2\,AgCl(s)\,(v)$$

Solution

i. Given: 2.0 moles of $CaCl_2$. Desired: ? g AgCl

ii. Molar mass of AgCl = 1x107.87 g Ag.mol^{-1} + 1x35.45 g C.mol^{-1} = 143.3 g AgCl.mol^{-1}

iii. The chemical equation is given.

iv. The desired equality: 1 mol $CaCl_2$ = 2 mol AgCl, and 1 mol AgCl =143.3 g AgCl.

v. The desired conversion factors:

$$\frac{2 \text{ mol AgCl}}{1 \text{ mol CaCl}_2} \quad \text{and} \quad \frac{143.3 \text{ g AgCl}}{1 \text{ mol AgCl}}$$

vi. Calculation:

$$2.0 \cancel{\text{ mol CaCl}_2} \times \frac{2 \cancel{\text{ mol AgCl}}}{1 \cancel{\text{ mol CaCl}_2}} \times \frac{143.3 \text{ g AgCl}}{1 \cancel{\text{ mol AgCl}}} = 573.3 \text{ g AgCl}$$

Example 3.4.7.4

How many grams of carbon dioxide are needed to react with 2 moles of water in the photosynthesis reaction?

Solution

i. Given: 2.0 moles of H_2O Desired: ? g CO_2.

ii. Molar mass of CO_2 = 1x12.011 g C.mol^{-1} + 2x15.999 g O.mol^{-1} = 44.009 g CO_2.mol^{-1}

iii. The chemical equation: : $6CO_2 + 6\ H_2O \rightarrow C_6H_{12}O_6 + 6O_2$.

iv. The desired equality: 6 mol CO_2 = 6 mol H_2O, and 1 mol CO_2 = = 44.009 g CO_2.

v. The desired conversion factors:

$$\frac{6 \text{ mol CO}_2}{6 \text{ mol H}_2}\text{O} \quad \text{and} \quad \frac{44.009 \text{ g CO}_2}{1 \text{ mol CO}_2}$$

vi. Calculation:

$$2.0 \cancel{\text{ mol H}_2\text{O}} \times \frac{6 \cancel{\text{ mol CO}_2}}{6 \cancel{\text{ mol H}_2\text{O}}} \times \frac{44.009 \text{ g CO}_2}{1 \cancel{\text{ mol CO}_2}} = 88 \text{ g CO}_2$$

Mass-to-mass conversion from a chemical equation

A chemical equation gives a mole-to-mole conversion factor. If the given substance is in grams and the desired substance is also in grams, then two additional conversion factors based on the molar masses are needed. That is, the following conversions are needed:

> Mass of given substance⇒mole of the given substance⇒mole of the desired substance⇒grams of the desired substance.

The reciprocal molar mass of the given substance is the first conversion factor, the mole to mole conversion factor from the chemical equation is the second conversion factor, and the molar mass of the desired substance is the third conversion factor needed. Make sure that each conversion factor cancels the denominator unit of its multiplier to the right, and the desired unit is left in the answer. The following examples explain these calculations.

Example 3.4.7.5

How many grams of Mg are needed to produce 1.01 g of H_2 gas in this reaction:

$$\text{Mg(s)} + 2\,\text{HCl(aq)} \longrightarrow \text{MgCl}_2\text{(aq)} + \text{H}_2\text{(g)}$$

Solution

i. Given: 1.01 g H_2 Desired: ? g Mg.

ii. Molar mass of H_2 = 2 x1.008 g H.mol^{-1} = 2.016 g H_2.mol^{-1}, and molar mass = 24.305 g Mg.mol^{-1}.

iii. The chemical equation is given in the problem.

iv. The desired equalities : 1 mol Mg = 1 mol H_2, 1 mol H_2 = 2.016g H_2, 1 mol Mg = 24.305 g Mg

v. Calculate by multiplying the given quantity consecutively with the three desired conversion factors from the equalities:

$$1.01 \cancel{\text{ g H}_2} \times \frac{1 \cancel{\text{ mol H}_2}}{2.016 \cancel{\text{ g H}_2}} \times \frac{1 \cancel{\text{ mol Mg}}}{1 \cancel{\text{ mol H}_2}} \times \frac{24.305 \text{ g Mg}}{1 \cancel{\text{ mol Mg}}} = 11.9 \text{ g Mg}$$

Example 3.4.7.6

How many grams of glucose are produced if 22.0 g of carbon dioxide is consumed in the photosynthesis reaction?

Solution

i. Given: 22.0 g CO_2 Desired: ? g $C_6H_{12}O_6$.

ii. Molar masses: of CO_2 = 1x12.011 g C.mol^{-1} + 2x15.999 g O.mol^{-1} = 44.009 g CO_2.mol^{-1}, and molar mass of $C_6H_{12}O_6$= 6x12.011 g C.mol^{-1} + 12x1.008 g H.mol^{-1} + 6x15.999 g O.mol^{-1} = 180.156 g $C_6H_{12}O_6$.mol^{-1} .

iii. The chemical equation: $6\,CO_2 + 6\,H_2O \longrightarrow C_6H_{12}O_6 + 6\,O_2$

iv. The desired equalities: 6 mol CO_2 = 1 mol $C_6H_{12}O_6$, 1 mol CO_2 = 44.009 g CO_2, 1 mol $C_6H_{12}O_6$ = 180.156 g $C_6H_{12}O_6$.

v. Calculate by multiplying the given quantity consecutively with the three desired conversion factors from the equalities:

$$22.0 \cancel{\text{ g CO}_2} \times \frac{1 \cancel{\text{ mol CO}_2}}{44.009 \cancel{\text{ g CO}_2}} \times \frac{1 \cancel{\text{ mol C}_6\text{H}_{12}\text{O}_6}}{6 \cancel{\text{ mol CO}_2}} \times \frac{180.156 \text{ g C}_6\text{H}_{12}\text{O}_6}{1 \cancel{\text{ mol C}_6\text{H}_{12}\text{O}_6}} = 15.0 \text{ g C}_6\text{H}_{12}\text{O}_6$$

Fig. 4.6.2 illustrates what each conversion factor does in the case of the above example number 4.6.5

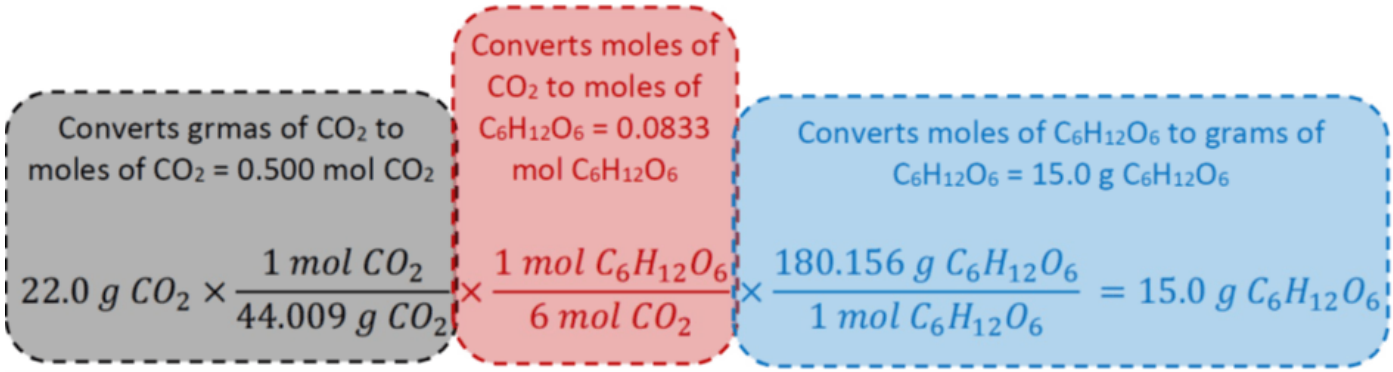


Figure 3.4.7.2: Illustration of the operation of conversion factors in a mass-to-mass calculation based on a balanced chemical equation of example 4..6.6 for question: "calculate grams of $C_6H_{12}O_6$ produced from 22.0 g CO_2 in photosynthesis reaction $6\,CO_2 + 6\,H_2O \longrightarrow C_6H_{12}O_6 + 6\,O_2$".

This page titled 3.4.7: Stoichiometric calculations is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 3.5: SOLUTIONS

3.5.1: Introduction to solution
3.5.2: Solubility
3.5.3: Electrolytes
3.5.4: Concentration of solutions
3.5.5: Osmosis

This page titled Chapter 3.5: Solutions is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.5.1: Introduction to solution

Solutions are all around us, e.g., air, seawater, body fluids, metal alloys are solutions. Fig. 5.1.1 illustrates that air is a mixture of nitrogen, oxygen, carbon dioxide, and some other gases; Fig. 5.1.2 illustrates that seawater is a mixture of water, chloride, sodium, sulfate, magnesium, and some other ions, and Fig. 5.1.3. illustrates that about 60% of the human body is composed of solutions called body fluids.

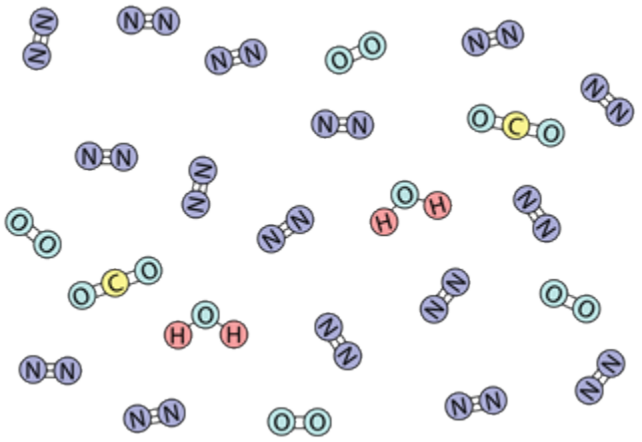


Figure 3.5.1.1: Air is a solution of nitrogen (N_2), oxygen (O_2), carbon dioxide (CO_2), and some other gases . Source: RoRo / CC0

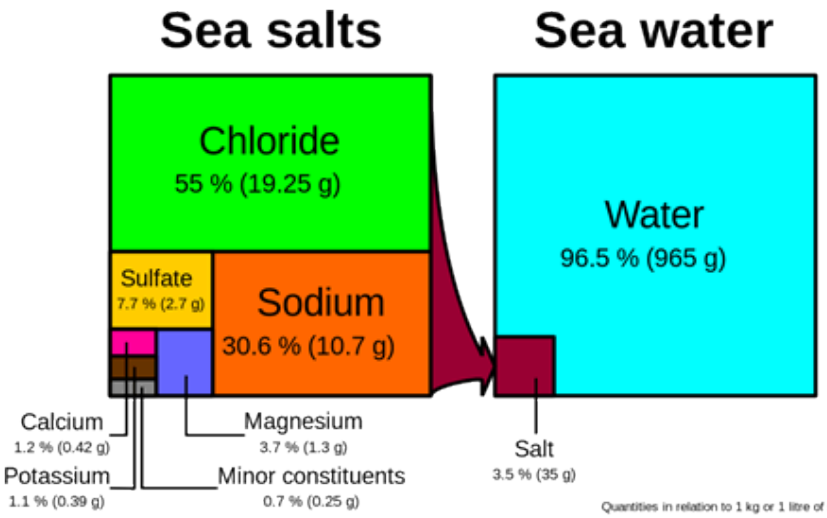
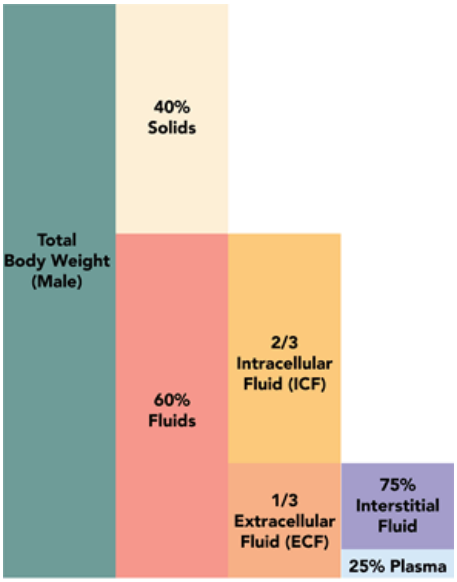


Figure 3.5.1.2: Seawater is a solution of water and several ions, expressed in wt/wt%. Source: derivative work: Tcncv (talk)Sea_salt-e_hg.svg: Hannes Grobe, Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany; SVG version by Stefan Majewsky / CC BY-SA



(https://creativecommons.org/licenses/by-sa/2.5)
Figure 3.5.1.3: This schematic shows the relative volumes of the different fluid compartments in an adult male. Source: Alan Sved and David Walsh / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

What is a solution?

- A solution is a homogeneous mixture of two or more pure substances.
- The substance that is in a large amount in the solution is called the **solvent**.
- The substance that is in smaller amounts in a solution is called the **solute**.

For example, the air is a solution in which nitrogen is the solvent, and water is the solvent in seawater and body fluids. Oxygen, carbon dioxide, and water vapors are solutes in the air; and sodium, chloride, sulfate, magnesium, and some other ions are solutes in seawater.

Types of solution

The solutions are generally classified in two ways: i) based on the physical state of the solution and the solute, and ii) based on the particle size of the solute.

Types of solution based on the physical state of the solution and the solute

The solutions can be classified based on the physical state of the solution, solvent, and solute. For example, the air is gas in a gas solution; carbonated water is a gas in a liquid solution; vinegar is a liquid in a liquid solution; metal alloys are solid in solid solutions. Table 5.1.1 lists the major types of solutions, solvents, and primary solutes in them.

Table 1: Examples of main types of solutions and solvent and major solute in them.

Type	Example	solvent	Primary solute
Gas in gas	Air	Nitrogen	Oxygen
Gas in liquid	Carbonated water	Water	Carbon dioxide
Liquid in liquid	Vinegar	Water	Acetic acid
Solid in liquid	Seawater	Water	Sodium chloride
Solid in solid	Brass	Copper	Zinc

Types of solutions based on the particle size of the solute

Solution

A solution is a homogeneous mixture comprising smaller component/s called solute/s of small molecules or ions comparable in size to the molecules of a larger component called the solvent.

For example, NaCl dissolved in water is a solution. The solute is almost uniformly distributed in the solvent, making a homogeneous mixture. The solute does not separate by filtration or by a semipermeable membrane but can be separated by some other physical process. For example, the distillation process separates a solid in a liquid or a liquid into a liquid solution. The solution is transparent, though it may be colored. A light passing through a solution is not visible, as shown in Fig. 5.1.4.

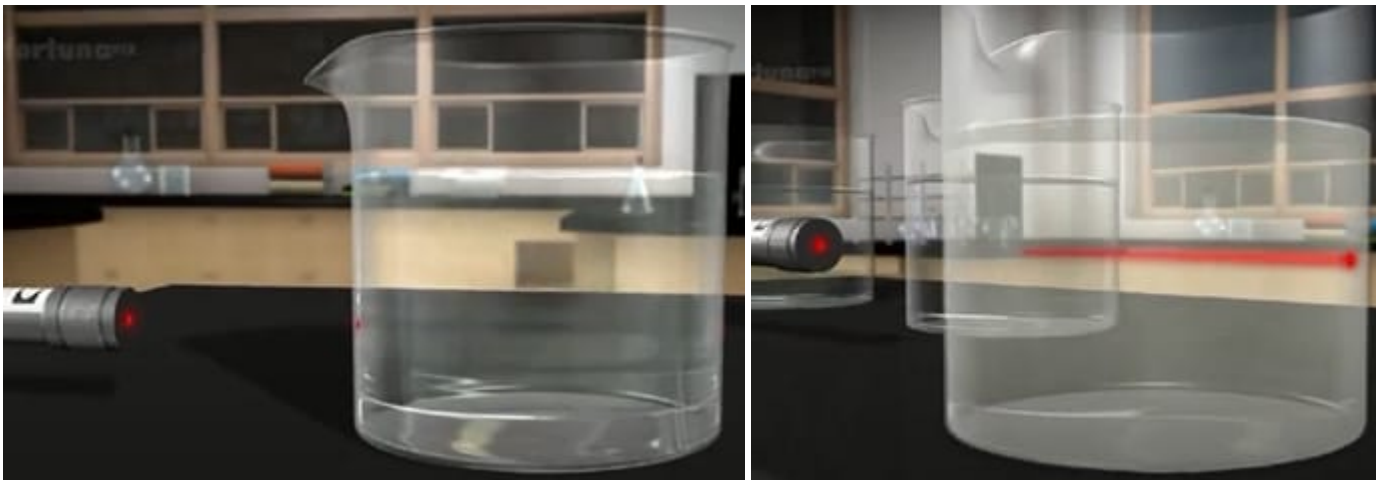


Figure 3.5.1.4: Demonstration of Tyndall effect: a red-laser bean passing through a solution is not visible (left), but visible when passing through a colloid (right). Source: https://youtu.be/8Xcpq6e8pBY, Creative Commons Attribution license (reuse allowed)

Suspension

A suspension is a heterogeneous mixture of solvent and solute particles of larger than 10,000 Å.

For example, muddy water is a suspension. If the suspension is allowed to stand, the suspended particles settle down and separate. The suspended particles can be filtered out. Some medicines, e.g., milk of magnesia, are suspensions. It is instructed to shake just before administering medicine to re-suspend the settled suspension.

Colloid

A colloid falls between a solution and a suspension. The colloidal particles are larger molecules like proteins or groups of molecules or ions.

Unlike a suspension, the colloids usually do not settle if allowed to stand. The colloidal particles can not be filtered but can be separated by a semipermeable membrane. When a light beam passes through a colloid, it scatters by the colloid particles, called the Tyndall effect, and becomes visible, as shown in Fig. 5.1.4.

Examples of colloids include:

1. fog and clouds that are liquid water droplets dispersed in air;
2. smoke that is solid carbon particles dispersed in air;
3. whipped cream that is air dispersed in a liquid;
4. styrofoam is a gas dispersed in a solid; and

5. ager medium that is liquid dispersed in a solid medium.

Water –a universal solvent

Water (H_2O) is an essential substance for life. It covers more than 70% of the earth's surface (Fig. 5.1.5), and it comprises more than 60% of the human body (5.1.3). In addition to being the most abundant solvent, water is a universal solvent because it is a polar molecule with a partial negative charge on oxygen and a partial positive charge on hydrogen atoms as shown in Fig. 5.1.6. The polarity of water molecules allows them to interact with other water molecules as well as with other polar compounds through dipole-dipole interactions and with other ions through ion-dipole interactions. These interactions help to dissolve a lot of polar and ionic compounds that are in and around us.

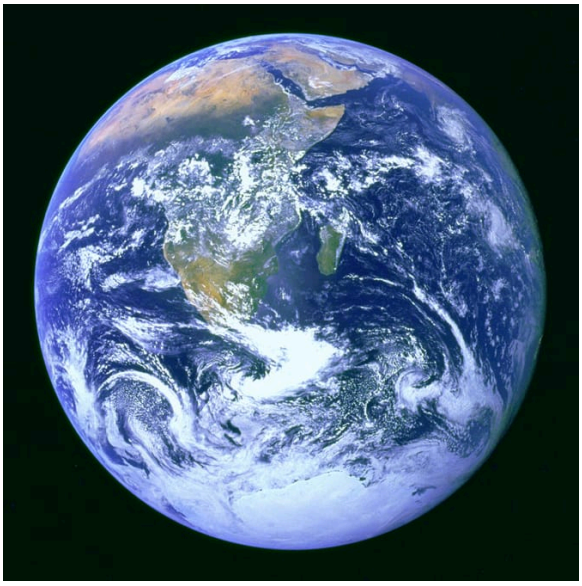


Figure 3.5.1.5: the image of the earth showing saltwater in the sea, fresh water on the Antarctic ice sheet, and clouds. Source: by NASA mission AS17. Public domain.

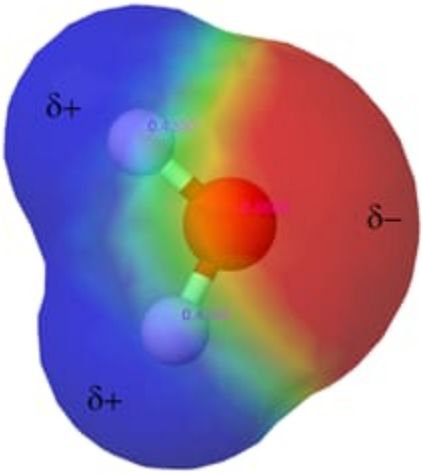


Figure 3.5.1.6: A model of water molecule showing oxygen in red and hydrogen in white color balls overplayed with an electrostatic potential map showing the partial positive (+d) region in red and the partial negative (-d) region in blue, drawn using a free web app **https://molview.org**

How does water dissolves polar and ionic compounds?

Water molecules establish electrostatic interaction, called hydrogen bonding, through the partial +ve end of one molecule with a partial –ve end of a neighboring molecule. These interactions impart unique properties to water, like its relatively higher boiling point and melting point compared to other substances of similar molecular weight. Other polar substances have similar interactions, e.g., ethanol has hydrogen bonding similar to water as illustrated in Fig. 5.1.7.

Dissolution of polar substance in water

When water mixes with other polar substances, like ethanol, some of the hydrogen bonding between water molecules replace with similar hydrogen bonding with ethanol molecules. Since the electrostatic potential energy is similar, the natural tendency to go towards more dispersion drives the dispersion of ethanol molecules uniformly in water resulting in the solution.

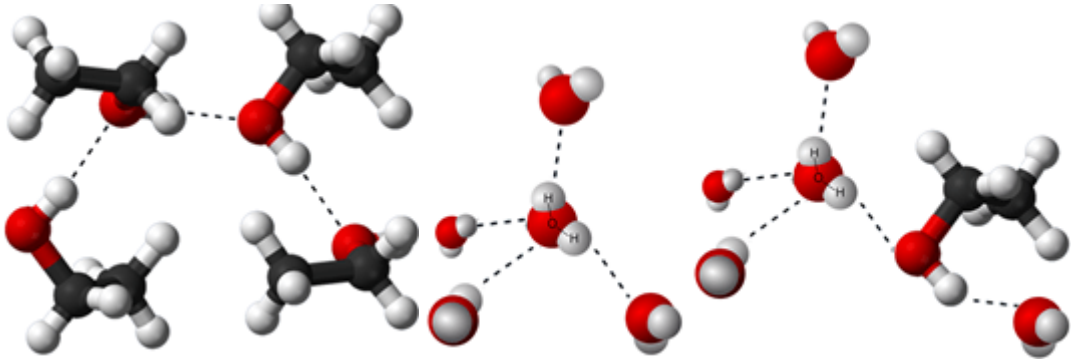


Figure 3.5.1.7: Models showing hydrogen bonding in ethanol (left), water (middle), and ethanol water solution (right). Source: modified from Benjah-bmm27, and snek01. Public domain

Ionic compounds are held together by electrostatic forces between opposite ions, i.e., ionic bonds. When an ionic compound is added to water, water molecules surround the cation and establish ion-dipole interaction by orienting their partial -ve end to the cation. Similarly, water molecules establish ion-dipole interaction with anions by orienting their partial +ev end towards the anion, as illustrated in Fig. 5.1.8

Dissolution of ionic compounds in water

The ion-dipole interactions, along with nature's tendency to disperse the particles, are usually strong enough to overcome the ionic bonds, dissociate the compounds into ions, and disperse them almost uniformly in the water.

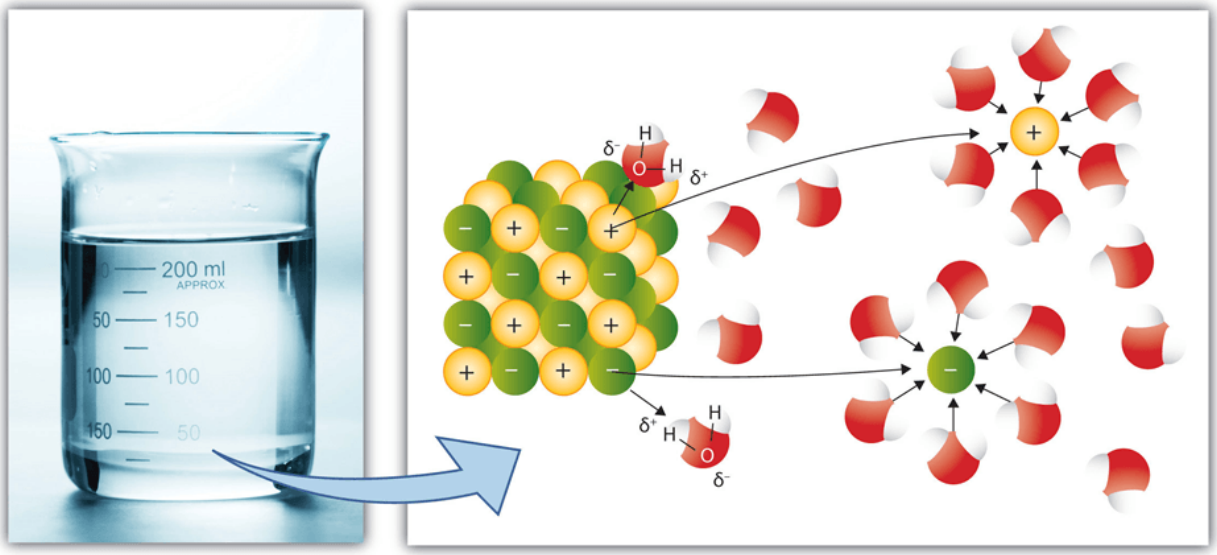


Figure 3.5.1.8 : Dissolution of sodium chloride in water through dissociation into ions and dispersion in water driven by ion-dipole interactions and nature's tendency to disperse particles. Source: Andy Schmitz / CC BY (https://creativecommons.org/licenses/by/3.0)

The separation of the cations from the anions of the ionic compound is called dissociation.

> The formation of a layer of water molecules around ions, driven by ion-dipole interactions, is called hydration.

Non-polar substances, like vegetable oil or gasoline, do not dissolve in water. The molecules in non-polar substances have only London dispersion forces. They easily dissolve in non-polar solvents like hexane or carbon tetrachloride that have similar London dispersion forces among their molecules.

> The fact that ionic and polar substances dissolve in polar solvents and non-polar substances dissolve in non-polar solvents of similar intermolecular interactions is called "**like dissolves like**."

This page titled 3.5.1: Introduction to solution is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.5.2: Solubility

Solubility and its related terminologies

The majority of solutes do not dissolve in water or other solvents in all proportions.

What is solubility

The maximum proportion of the solute that can dissolve in a given amount of the solvent, usually expressed in grams of solute in 100 grams of solvent, is called the **solubility** of the solute in the solvent.

1. Substances that make a solution when mixed in any proportion are called **miscible**. For example, ethanol and water are miscible.
2. Some substances make a solution when mixed in some proportion but not in all proportions; these are called **partially miscible**. For example, n-butanol can mix in water up to 7.3 g n-butanol/100 ml water.
3. Substances that almost do not dissolve in each other are called **immiscible**. For example, n-Hexane is immiscible in water.
4. The solution that has not yet reached its solubility limits and can dissolve more solute added is called an **unsaturated solution**.
5. The solution that has reached its solubility limits and can not dissolve if more solute is added to it is called a **saturated solution**.

When a solute is added to a saturated solution, it does dissolve, but, at the same time, the dissociated components recombine to form the crystals of the solute, i.e., recrystallize, at the same rate, so that there is no net dissolution, as illustrated in Fig. 5.2.1.

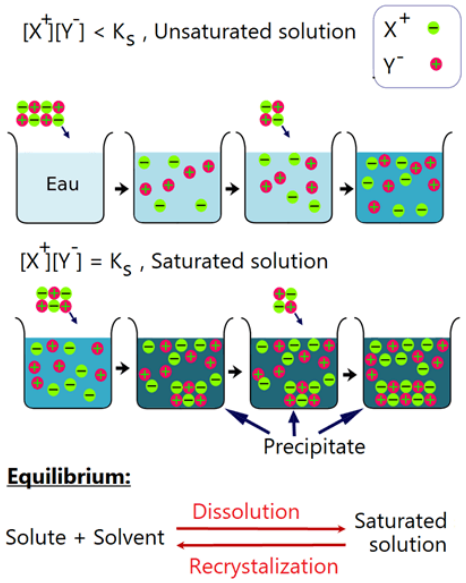


Figure 3.5.2.1: Illustration of unsaturated and saturated solution and the dynamic equilibrium between dissolution and recrystallization in the saturated solution. Source: No machine-readable author provided. Romary assumed (based on copyright claims). / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/)

Effect of temperature on solubility

The solubility of most of the solids and liquids increases as the temperature increases.

For example, when a saturated solution of sugar in water is heated, it can dissolve more sugar. Fig 5.2.2 show the solubility vs temperature curves for some compounds.

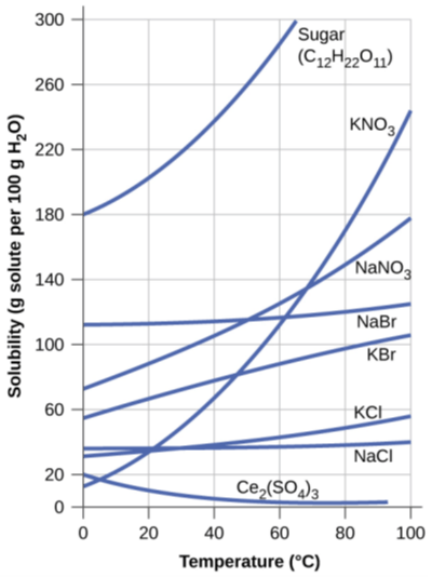


Figure 3.5.2.2: Solubility of some ionic and polar molecular compounds in water as a function of temperature. Download for free at https://openstax.org/details/books/chemistry.

When a hot saturated solution is cooled, the concentration of solute in the solution becomes above the solubility limits, making a **supersaturated solution**.

The supersaturated solution is unstable and ultimately crystallizes out the excess solute leaving any impurities in the solution, as illustrated in Fig. 5.2.3. This process is called re-crystallization, which is used to purify the solutes.



Figure 3.5.2.3: Salicylic acid recrystallized in water. Source: https://www.youtube.com/watch?v=QwqvzfPneUo

Medical issues related with solubility

The crystallization of excess solute from a supersaturated solution is responsible for some medical problems like gout and kidney stone. Gout is the crystallization of uric acid in the cartilage, tendons, and soft tissues when the concentration of the uric acid in blood plasms exceeds its solubility limit of ~7 mg/100 mL at 37 ^{o}C. It causes redness, swelling, and pain in the affected area, as illustrated in Fig. 5.2.4. Kidney stones are solid materials formed in the urinary tubes, as illustrated in Fig. 5.2.5. Kidney stones are the result of the crystallization of excess calcium phosphate, calcium oxalate, or uric acid in the urine.



Figure 3.5.2.4: Illustration of gout as a result of the recrystallization of uric acid in the cartilage, tendons, and soft, causes swelling, joint stiffness, and aching around the joint in the foot. Source: www.scientificanimations.com/ / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

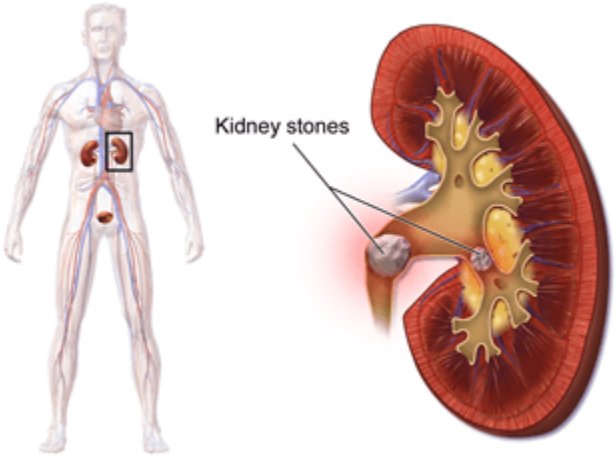


Figure 3.5.2.5: Kidney Stones illustrated. Source: Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". WikiJournal of Medicine 1 (2). DOI:10.15347/wjm/2014.010. ISSN 2002-4436/ CC BY (https://creativecommons.org/licenses/by/3.0)

Opposite to the solids and liquids, the solubility of gases generally decreases with an increase in temperature.

That is why the carbonated water releases dissolved gas when heated, causing pressure increase, which, in turn, causes the bursting of the soda can

Effect of pressure on solubility

The pressure has almost no effect on the solubility of solids and liquids but has a strong effect on the solubility of gases.

Henry's law

The solubility of gases in liquids is directly proportional to the pressure of the gas above the liquid.

An increase in pressure causes a decrease in the gas volume that increases the gas concentration. More frequent collision of the gas molecules with the gas-liquid boundary in a concentrated solution causes an increase in the rate of dissolution of the gas in the liquid, as illustrated in Fig 5.2.6. The opposite happens when the gas pressure decreases. For example, carbon dioxide starts bubbling out when a soda can is open because the gas escapes resulting in a decrease in the gas pressure above the liquid and a decrease in the solubility of the gas in water.

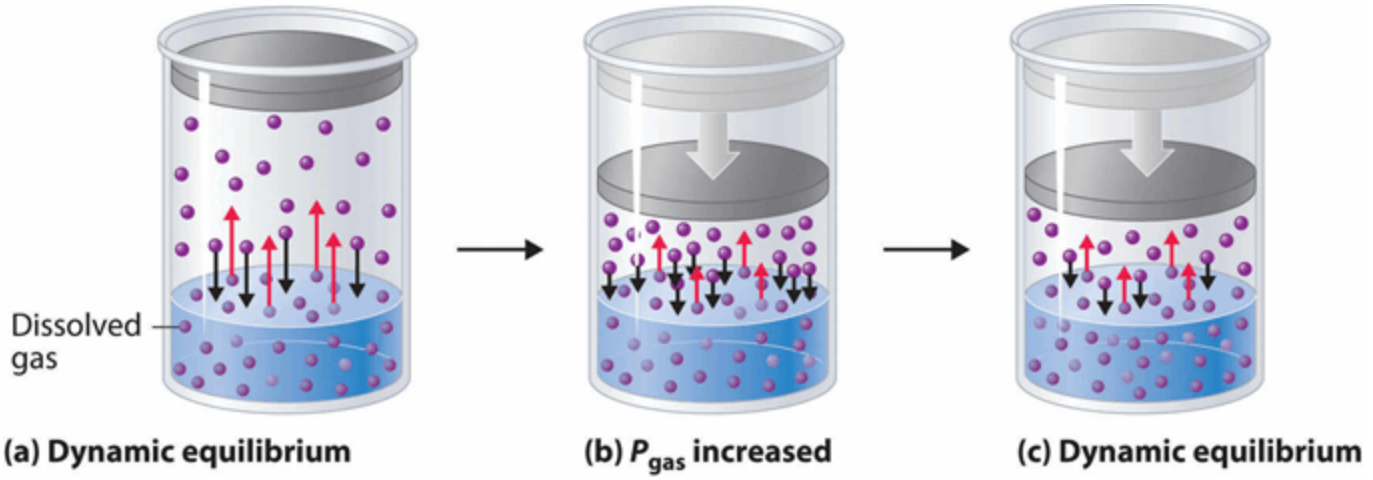


Figure 3.5.2.6: The Solubility of a gas increases as the partial pressure increases at a constant temperature. Source: Abozenadah, H., Bishop, A., Bittner, S. and Flatt, P.M. (2017) Preparatory Chemistry. CC BY-NC-SA. Available at: https://wou.edu/chemistry/courses/on...ory-chemistry/

Solubility guidelines for dissolution of ionic compounds in water

> If the solubility of a compound is less than 0.01 mol/L, it is considered **insoluble**.

The solubility of ionic compounds in water depends on the nature of the compound. For example, lead(II)iodide (PbI_2) and silver chloride (AgCl) are insoluble in water because the solubility of PbI_2 is 0.0016 mol/L of the solution and the solubility of AgCl is about 1.3 x 10^{-5} mol/L of solution. Potassium iodide (KI) and $Pb(NO_3)_2$ are soluble in water. When aqueous solutions of KI and $Pb(NO_3)_2$ are mixed, the concentration of PbI_2 in the mixture goes above its solubility limits, and it precipitates out, as illustrated in Fig. 5.2.7.



Figure 3.5.2.7: Precipitation reaction: $\mathrm{Pb(NO_3)_2\,(aq) + 2\,KI\,(aq) \longrightarrow PbI_2(s)\,(v) + 2\,KNO_3\,(aq)}$. source: PRHaney [CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)]

There are no fail-proof guidelines for predicting the solubility of ionic compounds in water. However, the following guideline can predict the solubility of most ionic compounds.

Soluble ions

1. Salts of alkali metals (Li^{1+}, Na^{1+}, K^{1+}, Rb^{1+}, Cs^{1+}) and ammonia (NH_4^{1+}) are soluble. No exceptions.
2. Salts of nitrate (NO_3^{1-}), acetate (CH_3COO^{1-}), and perchlorate (ClO_4^{1-}) are soluble. No exceptions.
3. Salts of chloride (Cl^{1-}), bromide (Br^{1-}), and Iodide (I^{1-}) are soluble, except when the cation is Pb^{2+}, Hg_2^{2-}, or Ag^{1+}. (Remember the acronym “LMS” based on the first letter of the element name, or phrase ‘Let Me See” to recall Lead, Mercury, and Silver.)
4. Sulfates (SO_4^{2-}) are soluble except when the cation Pb^{2+}, Hg_2^{2-}, Ag^{1+}, or a heavy alkaline earth metal ions: calcium (Ca^{2+}), barium (Ba^{2+}), or strontium (Sr^{2+}). ((recall “Let Me See” for Lead, Mercury, and Silver. Remember the acronym “CBS” based on the first letter of the element name, or the phrase “Come By Soon” to recall calcium, barium, and strontium.)

Insoluble ions

1. Hydroxide (OH^{1-}) and sulfides (S^{2-}) are insoluble except when the cation is an alkali metal, ammonia, or a heavy alkaline earth metal ions: Ca^{2+}, Ba^{2+}, and Sr^{2+}. (Recall the phrase “Come By Soon” to recall calcium, barium, and strontium.)
2. Carbonates (CO_3^{2-}), phosphates (PO_4^{3-}), and oxide (O^{2-}) are insoluble except when the cation is an alkali metal, ammonia
3. If there is a conflict between the two guidelines, then the guideline listed first has priority. For example, the salts of insoluble ions become soluble when the cation is an alkali metal, ammonia (rule#1).

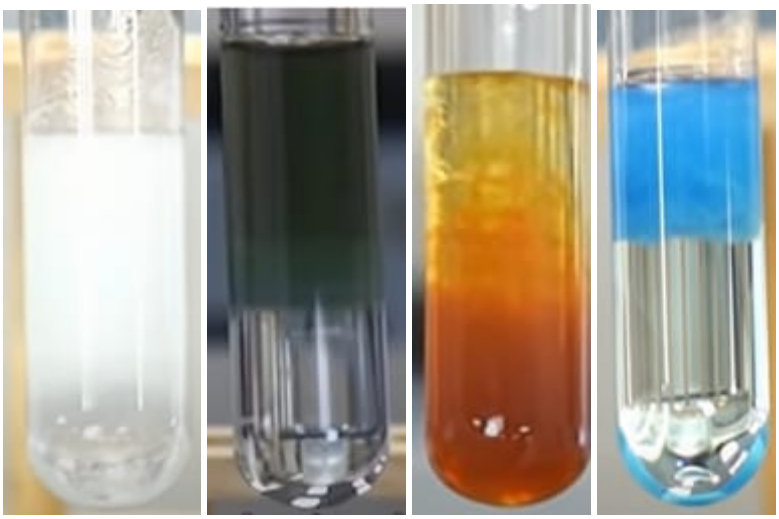


Figure 3.5.2.8: The precipitates of some insoluble ionic compounds formed by mixing the aqueous solution of appropriate soluble ionic compounds. The precipitates are from the left: white Calcium sulfate ($CaSO_4$), black Iron(II) hydroxide ($Fe(OH)_2$, brown Iron(III) hydroxide ($Fe(OH)_3$), and blue Copper(II) hydroxide ($Cu(OH)_2$). Source: https://youtu.be/jltLlzZ6FqU

Fig. 5.2.8 shows precipitates of some insoluble ionic compounds formed by mixing aqueous solutions of appropriate soluble ionic compounds. The precipitation can be predicted, as illustrated in Fig. 5.2.9. List the ions of the soluble ionic compounds and then cross-combine the cations of one with the anion of the other to make the potential products. If any of the potential products is an insoluble ionic compound, it precipitates out.

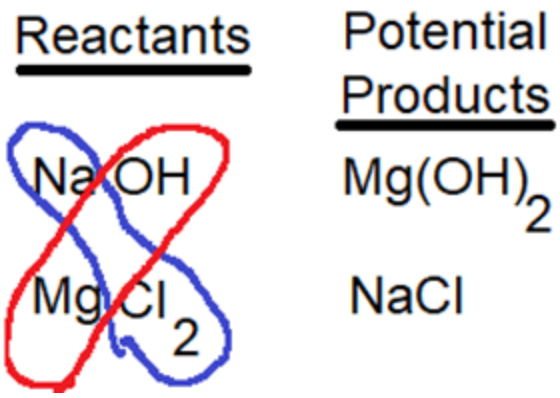


Figure 3.5.2.9: Cross-combine the cation-anion in the reactants. If any of the cross-combination is an insoluble salt, it will precipitate out, e.g: $\mathrm{NaOH(aq) + MgCl_2(aq) \rightarrow Mg(OH)_2(\ s) \downarrow + NaCl(aq)}$.

This page titled 3.5.2: Solubility is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.5.3: Electrolytes

What is an electrolyte?

Electric current is defined as the movement of electric charges. The substances through which an electric current can flow are called electrical conductors, and the others are electrical nonconductors. Metals are electrical conductors because valence electrons of metal atoms can move around in a piece of metal. Ionic compounds are composed of cations and anions, but the ions in a solid can not move around. Therefore, solid ionic compounds are electrical nonconductors. Pure water does not have a sufficient concentration of ions in it and is an electrical nonconductor. The ionic compounds dissociate into ions when dissolved in water. The solution of ionic compounds in water is an electrical conductor because the ions can move around in the solution, as illustrated in Fig. 5.3.1.

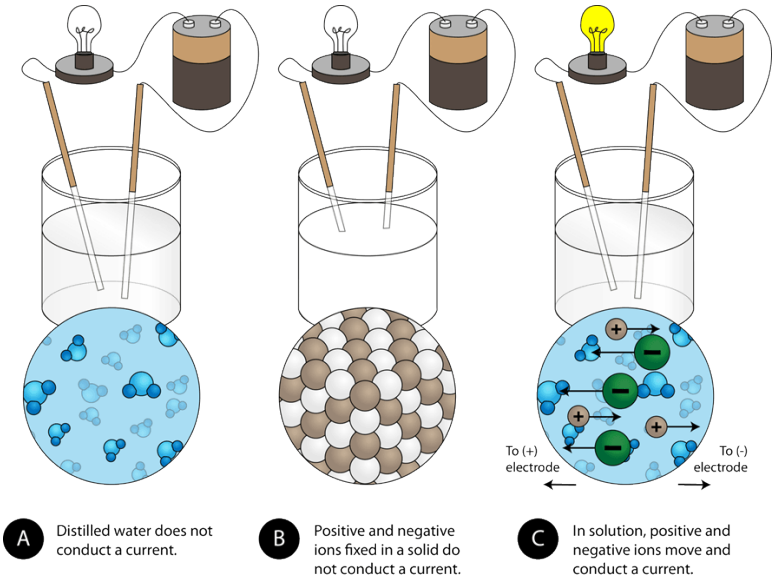


Figure 3.5.3.1: Pure water (A) and solid ionic compounds (B) are electrical noncondutors, but the solution of ionic compounds in water is an electrical conductor. Source: https://www.hiclipart.com/free-trans...cmeih/download

> Substances that produce electrically conducting solution when dissolved in water (or in another polar solvent) are called **electrolytes**.

All ionic compounds, acids, and bases produce ions in water and are classified are electrolytes.

> Substances that produce an electrically nonconducting solution when dissolved in water are called **nonelectrolytes**.

Molecular compounds other than acids and bases, such as methanol, acetone, sugar, and glucose, remain neutral molecules when dissolved in water. The molecular solutes, other than acids and bases, are nonelectrolytes.

Strong and weak electrolytes

> Substances that almost 100% dissociate into ions when dissolved in water are **strong electrolytes**.
>
> 1. All ionic compounds that are soluble in water are strong electrolytes.
> 2. Strong acids are strong electrolytes.
> 3. Strong bases are ionic compounds and strong electrolytes.

A strong electrolyte does not mean that it is necessarily highly soluble in water. It means that the portion of the solute that dissolves, it also dissociates 100% into ions in water, e.g., all ionic compounds. The solubility $Ca(OH)_2$ is only 0.16 g $Ca(OH)_2$/100 g water at 20 ^oC, but all the dissolved $Ca(OH)_2$ dissociates into Ca^{2+} and OH^- ions.

Strong bases are hydroxides of alkali metals, i.e., LiOH, NaOH, KOH, RbOH, and CsOH, and hydroxides of heavy alkaline earth metals, i.e., $Ca(OH)_2$, $Sr(OH)_2$ and $Ba(OH)_2$, which are strong electrolytes. Strong acids, i.e., HCl, HBr, HI, $HClO_4$, HNO_3, and H_2SO_4, are molecular compounds but are strong electrolytes because they dissociate almost 100% into ions when dissolved in water. For example, HCl almost wholly dissociates into ions when dissolved in water.

$$\mathrm{HCl(g) + H_2O(l) \rightarrow H_3O^{1+}(aq) + Cl^{1-}(aq)}$$

Substances that partially dissociate into ions when dissolved in water are **weak electrolytes**. Weak acids and weak bases are weak electrolytes.

Week acids and week bases, like acetic acid (CH_3COOH) and ammonia (NH_3), are soluble in water, but partially dissociate into ions. For example, if 1-mole acetic acid or 1-mole of ammonia is dissolved in 1 liter of water at room temperature, they establish the following equilibrium between dissolved molecules and dissociated ions:

$$\mathrm{CH_3COOH(l) \overset{Water}{\longrightarrow} CH_3COOH(aq) + H_2O(l) \rightleftarrows CH_3COO^{1-}(aq) + H_3O^{1+}(aq),}$$
$$\mathrm{NH_3(\ g) \overset{Water}{\longrightarrow} NH_3(aq) + H_2O(l) \rightleftarrows {NH_4}^{1+}(aq) + OH^{1-}(aq),}$$

where only about 0.4% of the dissolved molecules dissociate into ions, the remaining about 99.6% of molecules remain neutral. Weak acids and weak bases are weak electrolytes.

Fig. 5.3.2 illustrates the difference between nonelectrolytes, strong electrolytes, and weak electrolytes.

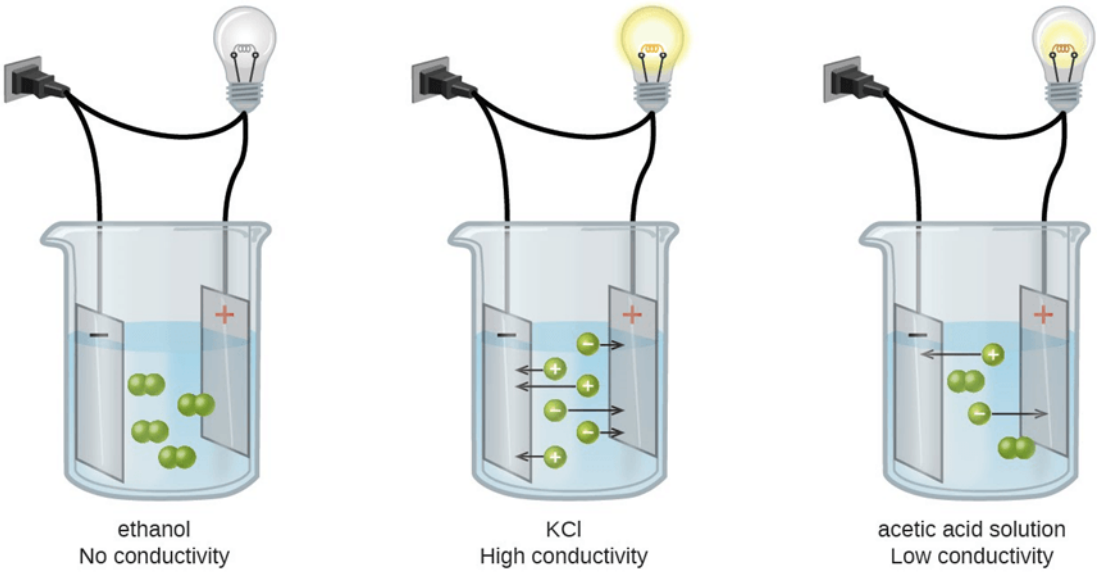


Figure 3.5.3.2: Illustration of a nonelectrolyte that does not conduct electricity, a strong electrolyte that has high electrical conductivity, and a weak electrolyte that has low electrical conductivity. Source: Karishma50/ (https;//creativecommons.org/licences/by-sa/4.0)

Equivalent

The amount of molecules and atoms is usually measured in moles. Ionic compounds are composed of ions but are overall neutral because the +ve charge is balanced by the –ve charge. Therefore, the mole ratio of cations to anions is not always one to one. For example, NaCl has a one-to-one mole ratio of Na^+ and Cl^-, but $CaCl_2$ has a one-to-two ratio of Ca^{2+} and Cl^-. A new unit, called equivalent (officially abbreviated as Equiv but commonly abbreviated as Eq), is introduced, to differentiate between a mole of ion

and a mole of charge on the ion.

The *equivalent* is the amount of a substance needed to react with or supply one mole of hydrogen ions (H^+) in an acid-base reaction, or react with or supply one mole of electrons in a redox reaction. In other words, the number of equivalents of a given ion in a solution is equal to the number of moles of that ion multiplied by its valence.

$$1 \text{ mol } Na^+ = 1 \text{ Eq } Na^+, \text{ but } \quad 1 \text{ mol } Ca^{2+} = 2 \text{ Eq } Ca^{2+}, \text{ and } 1 \text{ mol } Al^{3+} = 3 \text{ Ea } Al^{3+}$$

Similarly,

$$1 \text{ mol } Cl^- = 1 \text{ Eq } Cl^-, \text{ but } \quad 1 \text{ mol } {CO_3}^{2-} = 2 \text{ Eq } {CO_3}^{2-} \text{ and } 1 \text{ mol } {PO_4}^{3-} = 3 \text{ Ea } {PO_4}^{3-}$$

The solutions of electrolytes are overall electrically neutral, i.e., the number of equivalents of cations is equal to the number of equivalents of anions in the solution.

For example, if 1 mole of NaCl is dissolved in water, there is 1 Eq of Na^+ ions and 1 Eq of Cl^- ions in the solution. If one mole of $CaCl_2$ is dissolved in the water there is 2 Eq of Ca^{2+} ions and 2 Eq of Cl^- ions in the solution. Similarly, if 1 mole of NaCl and 1 mole of $AlCl_3$ are dissolved in the water there is 1 Eq of Na^+ ions, 3 Eq of Al^{3+} ions, and 4 Eq of Cl^- ions to balance cations in the water.

Equality gives two conversion factors. For example, the equality 1 mol Ca^{2+} = 2 Eq Ca^{2+} gives the following two conversion factors:

$$\frac{1 \text{ mol } Ca^{2+}}{2 \text{ Eq } Ca^{2+}}$$

and

$$\frac{2 \text{ Ea } Ca^{2+}}{1 \text{ mol } Ca^{2+}}$$

The first conversion factor converts the given amount in equivalents to moles, and the second converts the given amount in moles to equivalents of the ion as explained in the following examples.

Example 3.5.3.1

a) Calculate mEq of Fe^{3+} in 0.0200 mol of Fe^{3+}? b) if chloride ion is the only anion in the solution, how many mEq of Cl^- are present in the solution?

Solution

a) Given 0.0200 mol Fe^{3+}. Desired mEq Fe^{3+}.

The equality: 1 mol Fe^{3+} = 3 Eq Fe^{3+} gives the conversion factor $\frac{3 \text{ Eq } Fe^{3+}}{1 \text{ mol } Fe^{3+}}$ that converts the given amount in moles to Eq of Fe^{3+}. Another equality: 1 Eq Fe^{3+} = 1000 mEq Fe^{3+} gives the conversion factor $\frac{1000 \text{ mEq } Fe^{3+}}{1 \text{ Eq } Fe^{3+}}$, that converts the Eq of Fe^{3+} to mEq of Fe^{3+}.

Calculations:

$$0.0200 \cancel{\text{ mol } Fe^{3+}} \times \frac{3 \cancel{\text{ Eq } Fe^{3+}}}{1 \cancel{\text{ mol } Fe^{3+}}} \times \frac{1000 \text{ mEq } Fe^{3+}}{1 \cancel{\text{ Eq } Fe^{3+}}} = 60 \text{ mEq } Fe^{3+}$$

Example 3.5.3.2

An intervenous saline solution contains 145 mEq/L of Na^+. How many moles of Na^+ are in 0.500 L of the solution?

Solution

a) Given: 145 mEq Na^+ /L solution and 0.500 L solution, Desired: ? mol Na^+

Conversion factors needed: The concentration in mEq/L is the first factor for L to mEq conversion, the equality 1 mEq Na^+ = 1 mmol Na^+ givens the second conversion needed for mEq to mmol conversion, and the equality 1000 mmol Na^+ = 1 mol Na^+ given the third conversion factor needed needed for mEq to mol conversion. The three conversion factors are applied one after the other in a single row in the following calculation:

$$0.500 \cancel{\text{ L solution}} \times \frac{145 \cancel{\text{ mEq Na}}}{1 \cancel{\text{ L solution}}} \times \frac{1 \cancel{\text{ mmol Na}^+}}{1 \cancel{\text{ mEq Na}^+}} \times \frac{1 \text{ mol Na}^+}{1000 \cancel{\text{ mmol Na}^+}} = 0.0725 \text{ mol Na}^+$$

Note how the units of the numerator in one fraction are canceled by the units of the denumerator of the following fraction leaving only the desired units uncancelled that become the units of the answer number.

Electrolytes in body fluids

Fig. 5.3.3 shows the electrolytes and their concentrations commonly found in body fluids. They play an essential role in cell and body functions. For example, sodium ions regulate the water content and play a role in electrical impulse transmission in the nervous system. Potassium ions play a role in maintaining a regular heartbeat and also play a part in electrical impulse transmission. Chloride ions are there to balance the charge. Carbonate ions are involved in maintaining the pH of the blood.

Concentrations of electrolytes in body fluids are not high enough to be reported in a mole or equivalent units. The electrolytes in the body fluids are usually reported in millimoles (mmol) or milliequivalent (mEq) units, where: $1000 \text{ mmol} = 1 \text{ mol}$ and $1000 \text{ mEq} = 1 \text{ Eq}$.

The overall concentration of electrolytes in intravenous fluids given to patients is about the same as of electrolytes in the body fluids. For example, Ringer's lactate solution contains about: 130 mEq/L Na^+, 4 mEq/L K^+, 3 mEq/L Ca^{2+}, 109 mEq/L Cl^-, and 28 mEq/L $lactate^-$. Note that the overall +ve charge (130+4+3= 137 mEq/L) is equal to the overall –ve charge (109+28 = 137 mEq/L).

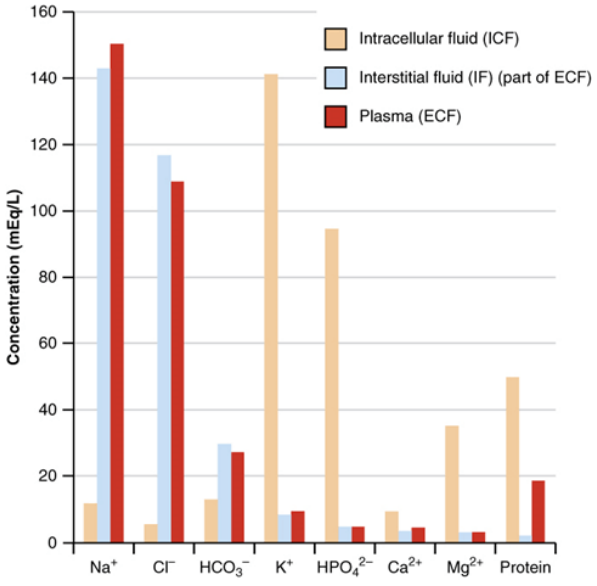


Figure 3.5.3.3: Concentrations of electrolytes in body fluids. Source: OpenStax College / CC BY (https://creativecommons.org/licenses/by/3.0)

This page titled 3.5.3: Electrolytes is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.5.4: Concentration of solutions

The concentration of a solution tells the amount of solute dissolved in a given amount of solution.

Making a solution of know concentration

A measured amount of solute is dissolved in enough solvent to make the desired volume of the solution, as illustrated in Fig. 5.4.1. The concentration of the solution can be expressed in different ways using mass, volume, or mole units, as explained in the following.

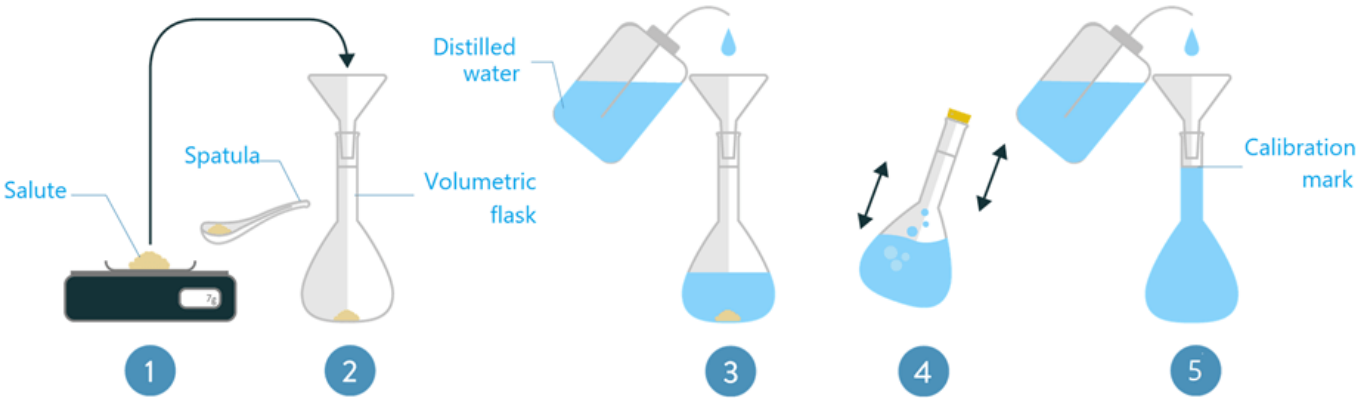


Figure 3.5.4.1: Making an aqueous solution: 1) measure the mass of the solute on an analytical balance, 2) pour the solute into a volumetric flask of know volume, 3) add distilled water (or solvent) to more than half fill the flask, 4) stopper the flask and shake it to dissolve the solute, 5) add more water to fill the flask to the calibration mark –measure from the lowest point of the water meniscus, 6) stopper the flask and shake again to make a homogeneous solution. Source: modified from https://www.hiclipart.com/free-trans...cvcrq/download

Concentration in percent (%)

What is percent (%)

Percentage (%) is a number or ratio that represent a fraction of 100.

For example, 5% means 5:100, where 5 is the part, and 100 is the total. The percentage is calculated as a hundred times of part by total, i.e.,

$$\text{Percentage } (\%) = \frac{\text{part}}{\text{Total}} \times 100.$$

Example 3.5.4.1

A 50.0 g NaCl is dissolved in water to make a 500 g solution. What is the percentage of NaCl in the solution?

Solution

Given part = 50.0 g NaCl, and total = 500 g solution. Desired: %NaCl in the solution?

Formula: $\text{Percentage } (\%) = \frac{\text{part}}{\text{Total}} \times 100.$

Calculations: $\text{Percentage } (\%) = \frac{50.0 cancel\ g}{500 cancel\ g} \times 100 = 10.0\ \%\text{NaCl}$

The units cancel in the fraction calculation part, and a % sign is added to the answer to tell that it is a fraction out of a hundred.

Mass percent (m/m)% concentration

The mass percent concentration expresses the mass units of solute in a hundred mass units of the solution.

$$\text{Mass}(\%) = \frac{\text{mass of solute } (g)}{\text{mass of solute } (g) + \text{ mass of solvent } (g)} \times 100 = \frac{\text{mass of solute } (g)}{\text{mass of solution } (g)} \times 100$$

Note that the total is solute and solvent added together, i.e., solution.

Example 3.5.4.2

What is the mass % of NaOH in a solution prepared by dissolving 10.0 g NaOH in 100 g water?

Solution

Given solute = 10.0 g, and solvent = 100 g, Desired: Mass% NaOH?

Formula: $\text{Mass } (\%) = \frac{\text{mass of solute } (g)}{\text{mass of solute } (g) + \text{ mass of solvent } (g)} \times 100$

Calculations: $\text{Mass } (\%) = \frac{10.0\ \text{gNaOH}}{(10.0\ \text{g} + 100\ \text{g})\ \text{solution}} \times 100 = 9.09\%\text{NaOH}$

Note that the mass% concentration and its reciprocal are two conversion factor: $\frac{\text{given } g \text{ solute}}{100 g \text{ solution}}$ and $\frac{100 g \text{ solution}}{\text{given g solute}}$

Example 3.5.4.3

Neosporin antibiotic is a 3.5% m/m neomycin solution. How many grams of neomycin are in 50 g of ointment?

Solution

Given: $3.5\%\ \text{neomycin } = \frac{3.5\ \text{g neomycin}}{100\ \text{g solution}}$, and Solution amount = 50 g, Desired: ? g neomycin

Calculations:

$$50\ \cancel{\text{g solution}} \times \frac{3.5\ \text{g neomycin}}{100\ \cancel{\text{g solution}}} = 1.8\ \text{g neomycin}.$$

Volume percent (v/v)% concentration

The volume percent concentration expresses the volume units of solute in a hundred volume units of the solution.

The mathematical form of the v/v % is:

$$\text{Volume } (\%) = \frac{\text{volume of solute (mL)}}{\text{volume of solution (mL)}} \times 100$$

Fig. 5.4.2 shows the volume percent concentration ranges of different classes of fragrances.

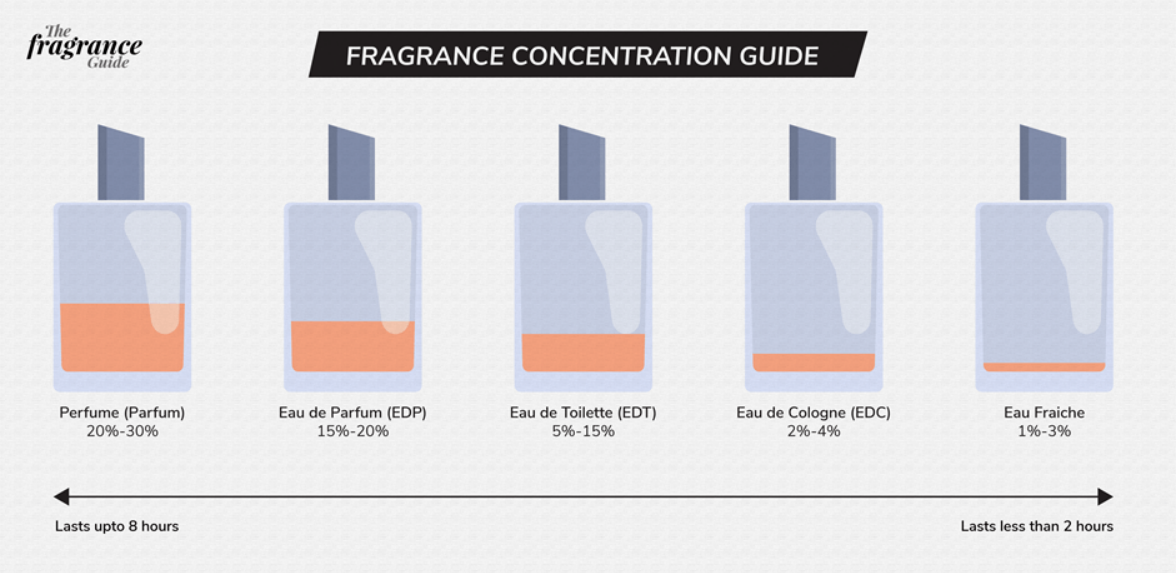


Figure 3.5.4.2: This info-graphic explains the differences between Parfums, Perfumes, Eau de Perfume, Eau de Parfum, Eau de Toilette, and Cologne. It also gives an approximate idea of how the concentration of Parfum oils vary as we move to lighter scents. Source: Nicole Smith / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Example 3.5.4.4

What is the volume % of rose extract in a solution prepared by dissolving 14.0 mL rose extract in a solvent to make 200 mL of solution?

Solution

Given: Solute = 14.0 g, and solution = 200 g, Desired: Volume% rose solution?

Formula:

$$\text{Volume } (\%) = \frac{\text{volume of solute } (mL)}{\text{volume of solution } (mL)} \times 100$$

Calculations:

$$\text{Volume } (\%) = \frac{14 \text{ mL rose extract}}{200 \text{ mL solution}} \times 100 = 7.0\% \text{ rose solution}$$

The two conversion factors for v/v % concentration are:

$$\frac{\text{given mL solute}}{100 \text{ mL solution}} \quad \text{, and} \quad \frac{100 \text{ mL solution}}{\text{given mL solute}}$$

Example 3.5.4.5

What is the volume of bromine (Br_2) in 250 mL of 4.8% v/v of Br_2 solution in carbon tetrachloride?

Solution

Given: Concentration 4.8% v/v bromine $= \frac{4.8\ mL \text{ bromine}}{100\ mL \text{ solution}}$, volume of solution = 250 mL, Desired: Volume of solute, i.e., ? mL bromine.

Calculations:

$$250 \text{ mL solution} \times \frac{4.8 \text{ mL bromine}}{100 \text{ mL solution.}} = 12 \text{ mL bromine.}$$

Mass/volume percent (m/v)% concentration

The mass/volume percent concentration expresses the mass units of solute in a hundred volume units of solution.

Mathematical form of m/v % is:

$$\text{Mass/volume } (\%) = \frac{\text{mass of solute } (g)}{\text{volume of solution } (mL)} \times 100$$

Example 3.5.4.6

What is the mass/volume % of glucose solution prepared by dissolving 50 g glucose in enough water to make 1000 mL of solution?

Solution

Given: Solute = 50.0 g, and Solution = 1000 mL, Desired: Mass/volume % glucose solution?

Formula:

$$\text{Mass/volume } (\%) = \frac{\text{mass of solute } (g)}{\text{volume of solution } (mL)} \times 100$$

Calculations:

$$\frac{\text{Mass}}{\text{volume}}(\%) = \frac{50 \text{ g glucose}}{1000 \text{ mL solution}} \times 100 = 5.0\% \text{ glucose solution by } \frac{\text{m}}{\text{v}}.$$

The two conversion factors from ms/v % concentration are:

$$\frac{\text{given } g \text{ solute}}{100 \text{ mL solution}} \quad \text{and} \quad \frac{100 \text{ mL solution}}{\text{given g solute}}$$

Example 3.5.4.7

How many grams of clindamycin antibiotics are in a 45 mL capsule of the 1.0% (m/v) clindamycin?

Solution

Given: % m/v concentration: 1.08 % m/v clindamycin $= \frac{1.0 \text{ g clindamycin}}{100 \text{ mL solution}}$, and volume of solution = 45 mL, Desired: ? g clindamycin?

Calculations:

$$45 \text{ mL solution} \times \frac{1.0 \text{ g clindamycin}}{100 \text{ mL solution}} = 4.5 \text{ g clindamycin.}$$

Parts per million (ppm) and parts per billion (ppb) concentration

Parts per million (ppm) is a number or ratio expressed as a fraction of a million (10^6).

For example, 2 ppm means $\frac{2}{1{,}000{,}000}$ or 2:1,000,000, where 2 is the part, and 1,000,000 is the total. The concentration in ppm is

calculated as a million times of part by total, i.e.:

$$\text{Concentration in ppm } = \frac{part}{\text{Total}} \times 10^6$$

Parts per billion (ppb) is a number or ratio expressed as a fraction of billion (10^9).

That is:

$$\text{Concentration in ppb } = \frac{\text{part}}{\text{Total}} \times 10^9$$

Like percentage concentration, the ppm and ppb can be mass/mass (m/m), volume/volume (v/v) or mass/volume (m/v).

Example 3.5.4.8

EPA's action limit for copper is 1.3 mg/L in drinking water. What is this limit in ppm of copper m/v in the drinking water?

Solution

Given: 1.3 mg copper in 1L solution, Desired: ? ppm m/v of Copper in water

Formula: $\text{Concentration in ppm } = \frac{\text{solute(g)}}{\text{solution(mL)}} \times 10^6$

Calculations: First, convert the given units of mass and volume into the corresponding units that the formula takes, then plug the values in the formula and calculate.

$$\text{Solute } = 1.3 \text{ mg} \times \frac{1 \text{ g}}{1000 \text{ mg}} = 0.0013 \text{ g}$$

$$\text{Solution } = 1 \text{ L} \times \frac{1000 \text{ mL}}{1 \text{ L}} = 1000 \text{ mL}$$

$$\text{Concentration in ppm } = \frac{0.0013 \text{ g}}{1000 \text{ mL}} \times 10^6 = 1.3 \text{ ppm copper } v/\text{m}$$

Example 3.5.4.9

EPA's action limit for lead is 0.015 mg/L in drinking water. What is this limit in ppb of lead m/v in the drinking water?

Solution

Given: 0.015 mg in 1L solution, Desired: ? ppb m/v of Lead in water

Formula: $\text{Concentration in ppb } = \frac{\text{solute}(g)}{\text{solution}(\text{mL})} \times 10^9$

Calculations: First, convert the given units of mass and volume into the corresponding units that the formula takes, then plug the values in the formula and calculate.

$$\text{Solute } = 0.015 \text{ mg} \times \frac{1 \text{ g}}{1000 \text{ mg}} = 0.000015 \text{ g}$$

$$\text{Solution } = 1 \text{ L} \times \frac{1000 \text{ mL}}{1 \text{ L}} = 1000 \text{ mL}$$

$$\text{Concentration in ppb } = \frac{0.000015 \text{ g}}{1000 \text{ mL}} \times 10^9 = 15 \text{ ppb lead v/m}$$

Molarity

Molarity (M) expresses the moles of solute in a liter of solution.

The most common solution concentration unit used in chemistry is molarity (M):

$$\text{Molarity } (M) = \frac{n \text{ (moles of solute)}}{V \text{ (Litters of solution)}}$$

Example 3.5.4.10

What is the molarity (M) of a solution prepared by dissolving 50.0 g NaOH in enough water to make 250 mL solution?

Solution

Given: Solute = 50.0 g NaOH, Volume of solution = 250 mL, Desired: ? M NaOH solution?

Formula: $\text{Molarity } (M) = \frac{n \text{ (moles of solute)}}{V \text{ (Litters of solution)}}$

Calculations: First, convert the given units of mass and volume into the corresponding units that the formula takes, then plug the values in the formula and calculate.

$$\text{Solute } = 50.0 \text{ g NaOH} \times \frac{1 \text{ mol NaOH}}{40.00 \text{ g NaOH}} = 1.25 \text{ mol NaOH}$$

$$\text{Solution } = 250 \text{ mL} \times \frac{1 \text{ L}}{1000 \text{ mL}} = 0.250 \text{ L}$$

$$\text{Molarity } (M) = \frac{1.25 \text{ mol NaOH}}{0.250 \text{ L solution}} = 5.00 \text{ M NaOH}$$

The two conversion factor from molarity are the following:

$$\frac{n \text{ (moles of solute)}}{V \text{ (Litters of solution)}} \quad \text{and} \quad \frac{V \text{ (Litters of solution)}}{n \text{ (moles of solute)}}$$

Example 3.5.4.11

How many litters of 0.211 M HCl solution are needed to provide 0.400 mol of HCl?

Solution

Given: amount of solute = 0.400 mol HCl, Concentration of solute = 0.211M = $\frac{0.211 \text{ mol HCl}}{1L \text{ solution}}$

Calculations:

$$\text{Liters of solution needed } = 0.400 \text{ mol HCl} \times \frac{1 \text{ L soluton}}{0.211 \text{ mol HCl}} = 1.90 \text{ L solution}$$

Example 3.5.4.12

A 2.50 L of 1.12 M NaOH solution contains how many moles of NaOH?

Solution

Given: Volume of solution = 2.50 L solution, concentration of solution = $1.12\ \text{M NaOH} = \frac{1.12\ \text{mol NaOH}}{1\ \text{L solution}}$, Desired: ? moles of NaOH?

Calculations:

$$\text{Moles of NaOH in the solution} = 2.50\ \text{L solution} \times \frac{1.12\ \text{mol NaOH}}{1\ \text{L solution}} = 2.80\ \text{mol NaOH}$$

Dilution of solutions

Dilution of a solution is the addition of a solvent to decrease the solute concentration of the solute in the solution.

The product of concentration (C) and volume (V) is the amount of solute, i.e.,

$$\text{Amount of solute} = C \text{ in } \frac{\text{amount of solute}}{\text{volume of solution}} \times V \text{ in volume of solution.}$$

The amount of solute does not change by adding solvent. Therefore, the product of concentration and volume, i.e., CV, which is the amount of solute, is a constant, i.e.,

$$C_1V_1 = C_2V_2 = \text{ amount of solute}$$

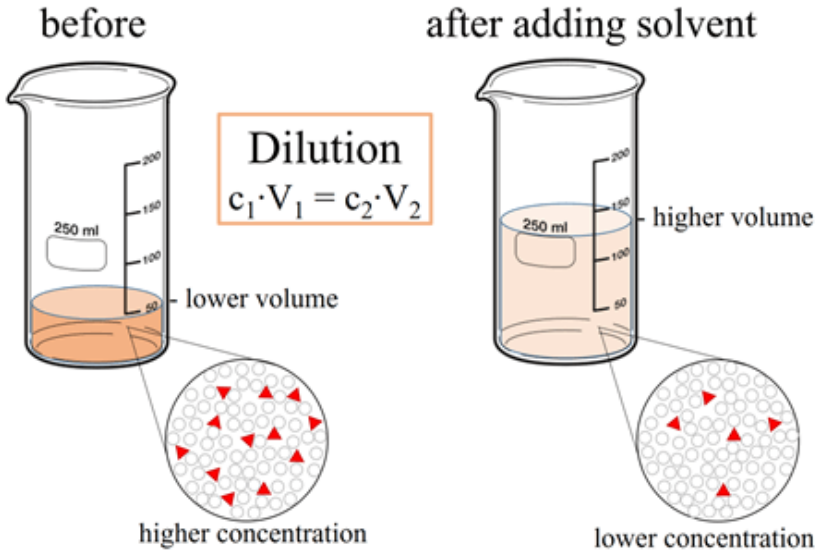


Figure 3.5.4.3: This image illustrates that when more solvent is mixed with the solution, the amount of solute remains the same, but its concentration decreases. Source: Theislikerice / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Fig. 5.4.3 shows that if the initial concentration is C_1, the initial volume is V_1, and after dilution, the final concentration is C_2, the final volume is V_2, then $C_1V_1 = C_2V_2$= amount of solute that is constant. If three of the four variables in this equation are known, the missing one can be calculated, as explained in the following example.

Caution

Keep in mind that the concentrations and volumes should be in the same units on both sides of the equation: $C_1V_1 = C_2V_2$. If they are not in the same units, convert them to the same units before plunging them in the formula.

Example 3.5.4.13

how much volume of 11.3 M HCl is needed to prepare 250 mL of 2.00 M HCl?

Solution

Given: C_1 = 11.3 M HCl, C_2 = 2.00 M HCl, V_2 = 250 mL solution, Desired V_1 = ?

Formula: $C_1V_1 = C_2V_2$, rearrange it to isolate the desired parameter: $V_1 = \frac{C_2V_2}{C_1}$

Calculations:

$$V_1 = \frac{C_2V_2}{C_1} = \frac{2.00 \text{ M HCl} \times 250 \text{ mL solution}}{11.3 \text{ M HCl}} = 44.2 \text{ mL solution}$$

Example 3.5.4.14

what is the molarity of the NaOH solution prepared by diluting 100 mL of 0.521 M NaOH solution to 500 mL?

Solution

Given C_1 = 0.521 M NaOH, V_1 = 100 mL solution, V_2 = 500 mL solution, Desired: Concentration of the final solution C_2 = ? M NaOH

Formula: $C_1V_1 = C_2V_2$, rearrange to isolate the desired parameter: $C_2 = \frac{C_1V_1}{V_2}$

Calculations:

$$C_2 = \frac{C_1V_1}{V_2} = \frac{0.521 \text{ M NaOH} \times 100 \text{ mL NaOH}}{500 \text{ mL NaOH}} = 0.104 \text{ M NaOH}$$

Example 3.5.4.15

Dopamine is administered intravenously to a patient to increase blood pressure. How many milliliters (mL) of a 4.0% (m/v) dopamine solution is needed to prepare 250 mL of a 0.030% m/v) solution?

Solution

Given: C_1 = 4.0% (m/v), C_2 = 0.030% (m/v), V_2 = 250 mL solution, Desired V_1= ?

Formula: $C_1V_1 = C_2V_2$, rearrange to isolate the desired parameter: $V_1 = \frac{C_2V_2}{C_1}$

Calculations:

$$V_1 = \frac{C_2V_2}{C_1} = \frac{0.030\% \times 250 \text{ mL solution}}{4.0\%} = 5.0 \text{ mL of solution}$$

Logarithmic dilution

> Take a unit volume of a given solution and add enough solvent to increase the volume of the solution 10 times for a logarithmic diluiton.

A logarithmic dilution is ten times dilution, i.e., proven by the following formula:

$$C_2 = \frac{C_1V_1}{V_2} = \frac{1 \text{ mL} \times C_1}{10 \text{ mL}} = 0.1 \times C_1$$

Repeating the above step with the diluted solution results in 10x10 = 100-time dilution, and repeating third-time results in 10x10x10 = 1000-time dilution. This dilution of 10 times in each step is called logarithmic dilution. Fig. 5.4.4 shows that five steps of logarithmic dilution on a 10% initial solution results in a concentration of 10 ppm in the final solution.

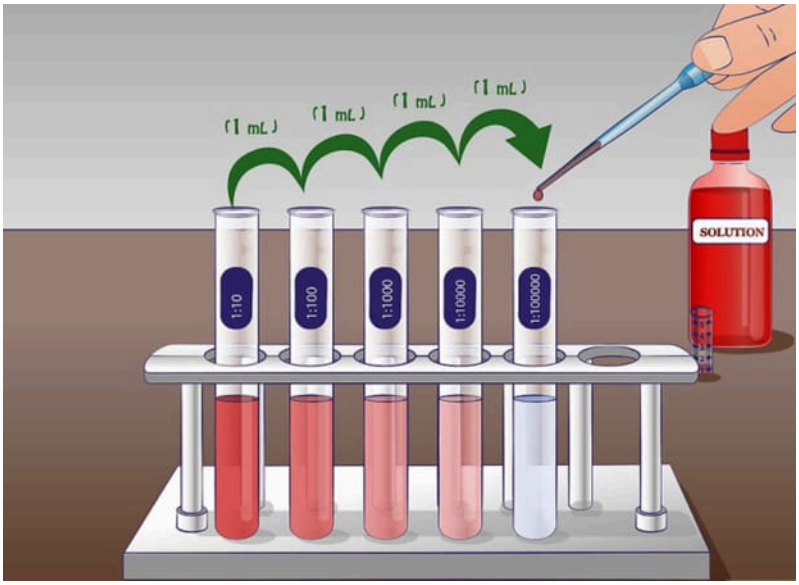


Figure 3.5.4.4: Logarithmic dilution: five steps of logarithmic dilution on a 10% initial solution results in a concentration of 10 ppm of the final solution. Source: Grasso Luigi / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

This page titled 3.5.4: Concentration of solutions is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.5.5: Osmosis

Semipermeable membranes surround living cells and organelles in the cells. The semipermeable membranes allow water and small molecules but do not allow the passage of large molecules and ions.

> Osmosis is the passage of water and small molecules across a semipermeable membrane with a net flow from a less concentrated solution to a more concentrated solution.

Osmosis helps in the absorption, retention, and flow of water, nutrients, and other molecules required in the biological systems. Fig. 5.5.1 illustrates the process of osmosis.

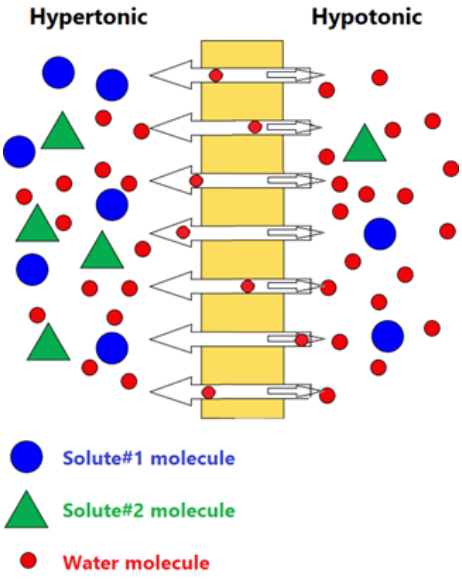


Figure 3.5.5.1: Illustration of osmosis. Osmosis is the passage of water and small molecules through a semipermeable membrane with a net flow from low solute concentration to high solute concentration. The larger molecules and ions cannot pass the semipermeable membrane.

Osmotic pressure

Osmosis causes net water flow from the less concentrated to the more concentrated solution across the semipermeable membrane. Consequently, the water level rises in the less concentrated solution compartment, as illustrated in Fig. 5.5.2. The difference in the height of water increases and applies pressure, pumping water back to the more concentrated side until the flow of water is equal on the two sides.

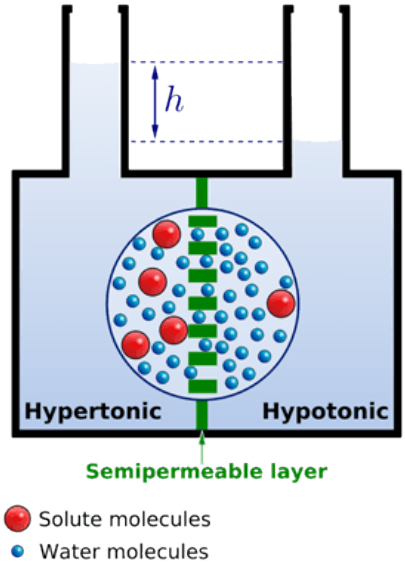


Figure 3.5.5.2: Osmotic pressure is the pressure, due to the difference in the height of water in this example, that prevents the excess flow of water from less concentrated to more concentrated solution side across the semipermeable membrane that separates the two solutions. Source: WYassineMrabetTalk✉ This W3C-unspecified vector image was created with Inkscape. / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

The pressure that prevents additional water flow to the more concentrated solution side of the semipermeable membrane is called osmotic pressure.

The osmotic pressure is proportional to the overall concentration of the solute particles. For example, 0.1 molar NaCl has osmotic pressure about twice that of 0.1 molar glucose because each mole of glucose adds one mole of solute particles, while each mole of NaCl produces two moles of particles, i.e., one-mole Na^+ and one-mole Cl^- ions in the solution.

Reverse Osmosis

Reverse osmosis is the net flow of water to the less concentrated or pure water across the semipermeable membrane by applying external pressure more than the osmotic pressure on the more concentrated solution side.

Reverse osmosis is the net flow of water to the less concentrated or pure water across the semipermeable membrane by applying external pressure more than the osmotic pressure. Note that solvent flow in reverse osmosis driven by external pressure is the opposite of regular osmosis. Reverse osmosis is used to produce drinking water from seawater sources, as illustrated in Fig. 5.5.3.

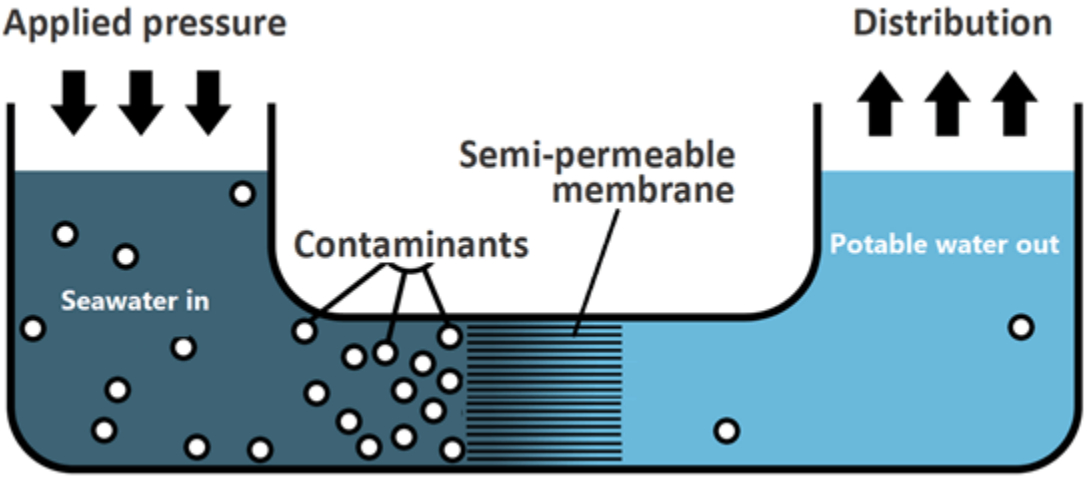


Figure 3.5.5.3: Illustration of reverse osmosis –used to produce drinking water from a seawater source. Source: Colby Fisher / CC BY-SA

(https://creativecommons.org/licenses/by-sa/3.0)

Isotonic, hypertonic, and hypotonic solutions

Cell membranes are semipermeable membranes that separate intracellular and extracellular fluids. The concentration difference across the membrane and the resulting osmotic pressure plays an essential role in cell functions. Intravenous solutions injected into patients must have the same osmotic pressure as the body fluids.

Solutions with the same solute particle concentration and osmotic pressure are called **isotonic**. If the two solutions across a semipermeable membrane do not have the same solute particle concentration, the solution with higher solute particle concentration and higher osmotic pressure is **hypertonic**, and the other has lower solute particle concentration and lower osmotic pressure is **hypotonic**.

Remember that hyper- means more and hypo-means less, concerning solute particle concentration in the case of osmosis.

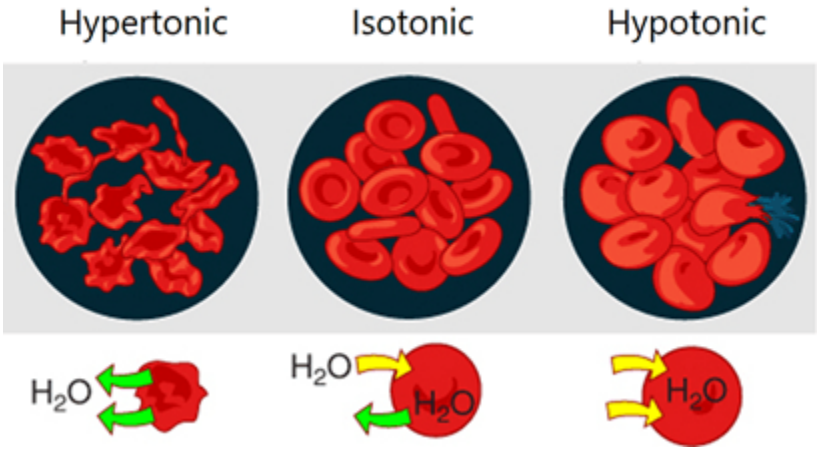


Figure 3.5.5.4: A hypertonic solution has a solute concentration higher than another solution. An isotonic solution has a solute concentration equal to another solution. A hypotonic solution has a solute concentration lower than another solution. Source: OpenStax / CC BY (https://creativecommons.org/licenses/by/3.0)

Cells placed in an external solution may retain their size, shrink, or swell depending on the relative osmotic pressure of fluid inside and outside of the cell, as illustrated in Fig. 5.5.4. For example, red blood cells placed in an isotonic solution retain their size because the flow of water into and out of the cell is the same.

- Typical isotonic solutions are 0.9% m/v NaCl solution in water or 5% m/v glucose solution in water.
- Red blood cells placed in a hypertonic solution shrink in size due to more flow of water out than into the cell –a process called **crenation**.
- Red blood cells placed in a hypotonic solution swell and burst due to more water flow into than out of the cells –a process called **hemolysis**.

A similar situation happens in plant cells that are placed in different environments concerning osmotic pressure, as illustrated In Fig. 5.5.5.

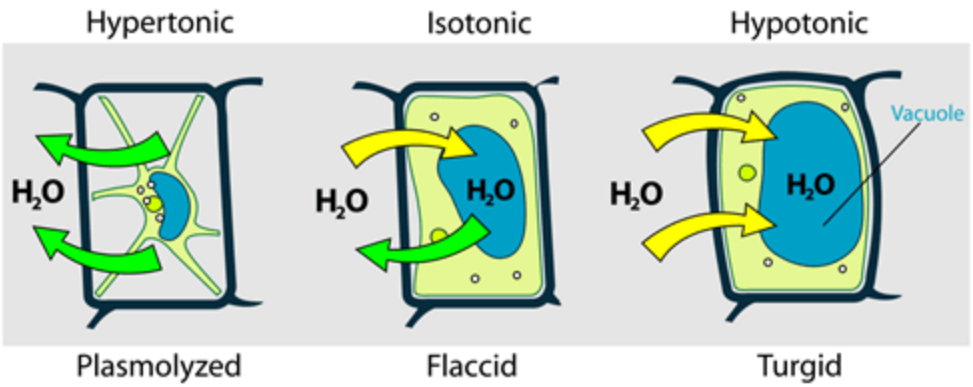


Figure 3.5.5.5: Plant cells under different environments concerning osmotic pressure. source: LadyofHats / Public domain

Dialysis

Dialysis separates colloids from water, dissolved ions, and molecules of small dimention

Dialysis is similar to osmosis with the difference that in dialysis water, small molecules and ions can pass through a dialyzing membrane leaving behind collide particles like proteins and starch molecules, as illustrated in Fig. 5.5.6.

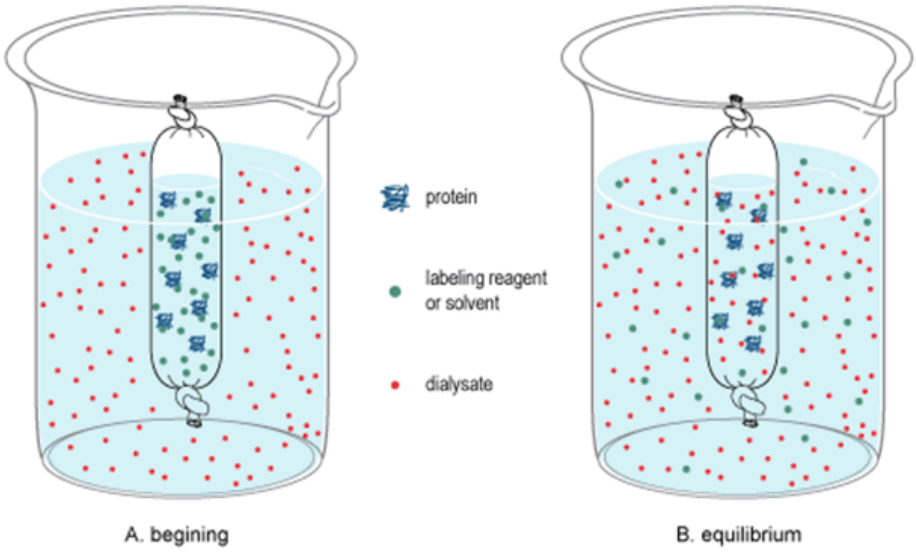


Figure 3.5.5.6: Dialysis separates colloidal particles like protein and starch from water, small molecules, and ions. Source: Potcherboy at English Wikipedia / CC BY (https://creativecommons.org/licenses/by/3.0)

Role of dialysis in the human body

Kidneys filter the blood by dialysis process. There are over a million tubular structures, called nephrons, in the kidney surrounded by dialyzing membranes. The nephrons filter the water, small molecules like glucose, amino acids, urea, and ions from the blood. Useful products and most of the water reabsorb later on, but urea and other waste products excrete through urine, as illustrated in Fig. 5.5.7.

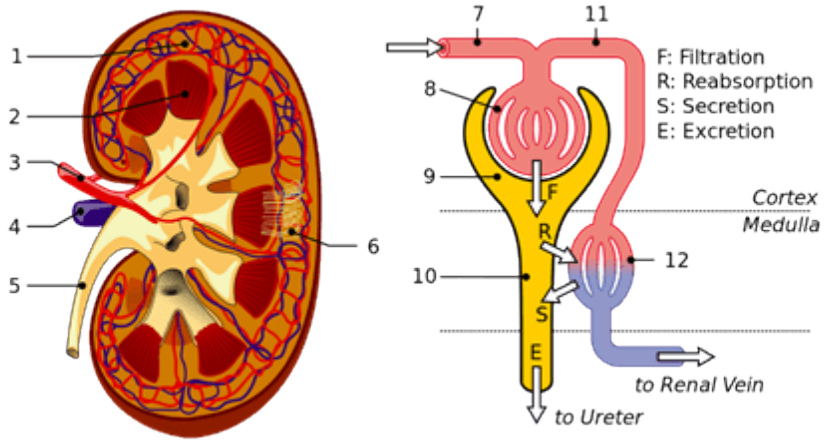


Figure 3.5.5.7: Schematics of dialysis for blood cleaning through kidney and nephron. 1: Renal cortex. 2: Medulla. 3: Renal artery. 4: Renal vein. 5: Ureter. 6: Nephrons. 7: Afferent arteriole. 8: Glomerulus. 9: Bowman's capsule. 10: Tubuli and loop of Henle. 11: Efferent arteriole. 12: Peritubular capillaries. Source: File:Physiology_of_Nephron.svg: Madhero88File:KidneyStructures_PioM.svg: Piotr Michał Jaworski; PioM EN DE PLderivative work: Daniel Sachse (Antares42) / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Hemodialysis

Hemodialysis is used to extract urea and other waste products from the blood when a person's kidney fails to remove them, as illustrated in Fig. 5.5.8. A hemodialysis system is a kind of artificial kidney in which the blood flows through long cellophane tubes placed in an isotonic solution containing NaCl, KCl, $NaHCO_3$, and glucose. Cellophane is a dialyzing membrane that does not let proteins, other large molecules, and blood pass through it, but urea excretes.

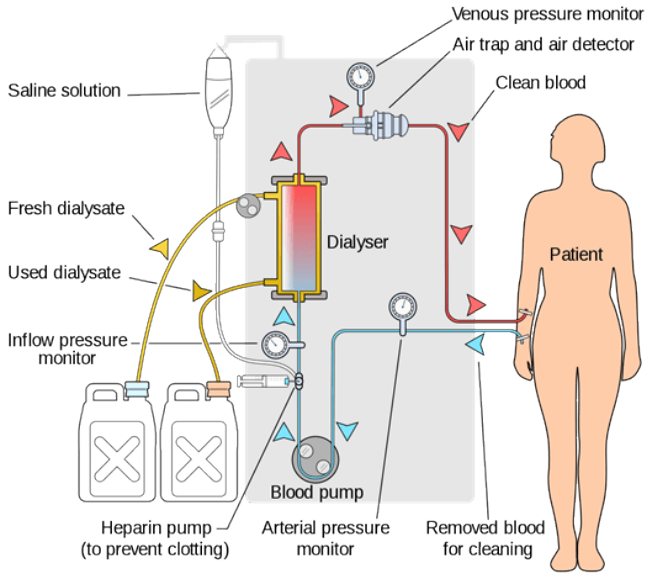


Figure 3.5.5.8: Simplified hemodialysis circuit. Source: GYassineMrabetTalk✉ This W3C-unspecified vector image was created with Inkscape. / CC BY (https://creativecommons.org/licenses/by/3.0)

This page titled 3.5.5: Osmosis is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 3.6: ACIDS AND BASES

3.6.1: What is an Acid and a Base?
3.6.2: Brønsted–Lowry acids and bases
3.6.3: Strength of acids and bases
3.6.4: Acid-base equilibrium
3.6.5: Dissociation of water
3.6.6: The pH
3.6.7: Acid-base reactions
3.6.8: pH Buffers

This page titled Chapter 3.6: Acids and bases is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.6.1: What is an Acid and a Base?

General Properties of Acids and Bases

We commonly encounter acids and bases in our foods –some foods are acidic, and others are basic (alkaline) as illustrated in Fig. 6.1.1.

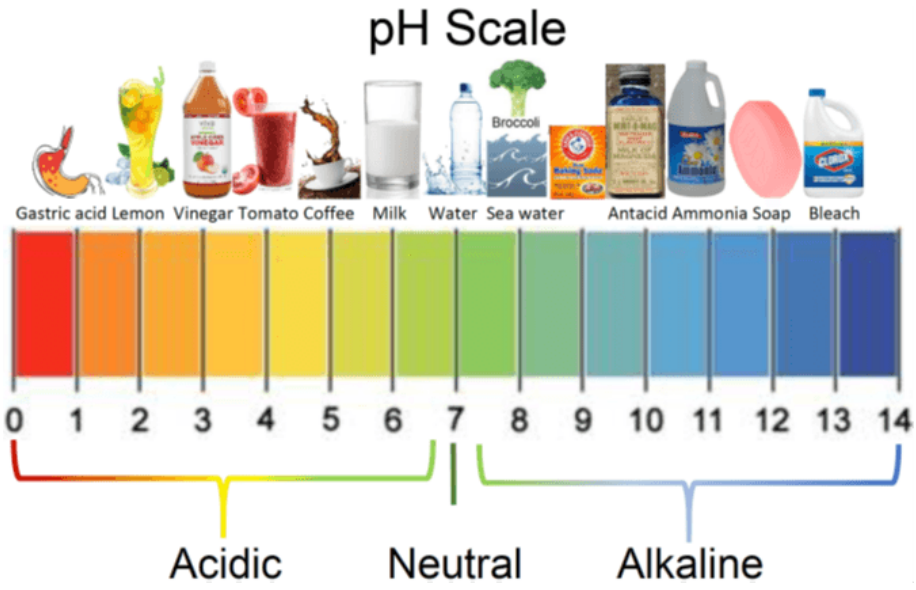


Figure 3.6.1.1: Acidic and alkaline foods along with a pH scale, Copyright; Public domain

The general properties of acids and bases are the following.

1. Acids taste sour, e.g., citrus fruits taste source because of citrus acid and ascorbic acid, i.e., vitamin C, in them. Basic (alkaline) substances, on the other hand, taste bitter.
2. Basic (alkaline) substances feel soupy, while acidic substances may sting.
3. The acids turn blue litmus paper to read but do not change the color of red litmus paper. Bases turn red litmus paper blue but do not change the color of blue litmus paper, as illustrated in Fig. 6.1.2.
4. Phenolphthalein indicator turns colorless in acid and turns pink in basic solution, as illustrated in Fig. 6.1.3.
5. Acids and bases neutralize each other. Hydrochloric acid is found in the stomach that helps digestion. Excess hydrochloric acid may cause acid burns—antacids like milk of magnesia are bases that help by neutralizing excess acid in the stomach.

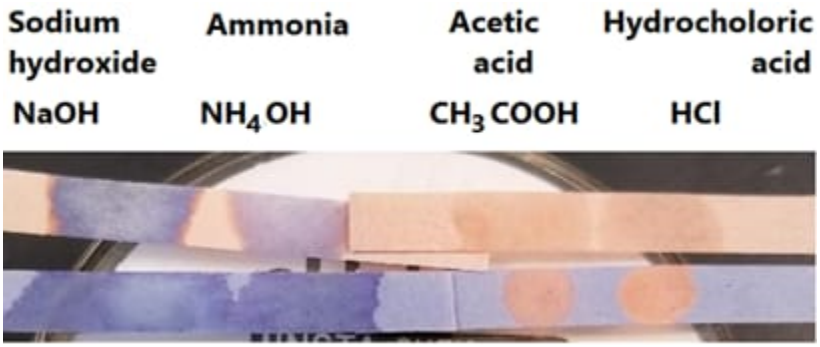


Figure 3.6.1.2: Demonstration of acids turn red litmus paper blue and bases turn blue litmus paper red - solution listed at the top of the image was spotted on the red litmus paper

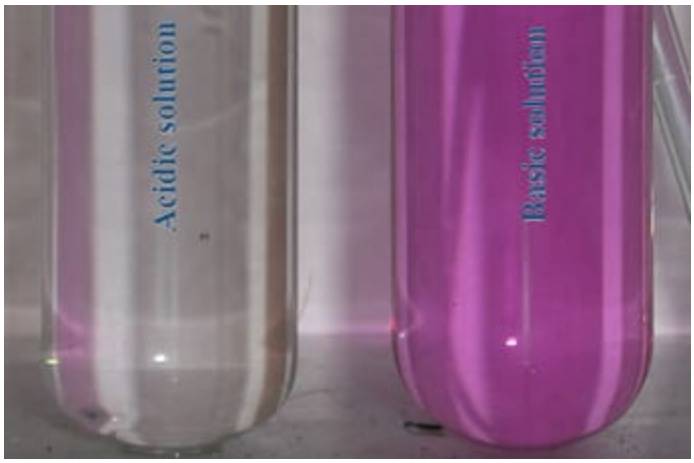


Figure 3.6.1.3: Colours of phenolphthalein in acid(left) and base (right) solutions. Source: User:Siegert / Public domain.

(top) or blue litmus paper (bottom).

Arrhenius's Definition of Acids and Bases

The earliest definition of acids and bases is Arrhenius's definition which states that:

- An acid is a substance that forms hydrogen ions H^+ when dissolved in water, and
- A base is a substance that forms hydroxide ions OH^- when dissolved in water.

For example, hydrochloric acid (HCl) is an acid because it forms $\mathrm{H^+}$ when it dissolves in water.

$$\mathrm{HCl(g)} \overset{\text{Water}}{\longrightarrow} \mathrm{H^+(aq) + Cl^-(aq)}$$

Similarly, NaOH is a base because it forms OH^- when it dissolves in water.

$$\mathrm{NaOH(s)} \overset{\text{Water}}{\longrightarrow} \mathrm{Na^+(aq) + OH^-(aq)}$$

Note that hydrogen ion H^+ does not exist in reality. It bonds with water molecules and exists as hydronium ion $H_3O^+(aq)$.

$$\mathrm{H^+(aq) + H_2O \rightarrow H_3O^+(aq)}$$

However, $\mathrm{H^+\,(aq)}$ is often written in the place of $(H_3O^+(aq)$.

Naming Arrhenius acids and bases

Table 1 lists the names and formulas of some of the common acids and their anions.

Rules for naming acids

1. The names end with the word “acid.”
2. If the anion is not an oxyanion, then add the prefix hydro- to the name of the anion and change the last syllable of the anion name to –ic. For example, Cl^- is a chloride ion, and HCl is hydrochloric acid.
3. If the anion is an oxyanion with the last syllable –ate, change the last syllable with –ic. Do not use the prefix hydro-, but add the last word “acid.” If there is a prefix per- in the name of the oxyanion, retain the prefix in the acid name. For example, NO_3^- is a nitrate, and HNO_3 is nitric acid. Another example, ClO_4^- is a perchlorate, and $HClO_3$ is perchloric acid.

4. If the anion is an oxyanion with the last syllable –ite, change the last syllable with –ous. Do not use the prefix hydro-, but add the last word "acid." If there is a prefix hypo- in the name of the oxyanion, retain the prefix in the acid name. For example, NO_2^- is nitrite, and HNO_2 is nitrous acid. Another example, ClO^- is hypochlorite, and HClO is hypochlorous acid.

Table 1: Names of some common acids and their anions.

Acid formula	Acid name	Anion	Anion name
HCl	Hydrochloric acid	Cl^-	Chloride
HBr	Hydrobromic acid	Br^-	Bromide
HI	Hydroiodic acid	I^-	Iodide
HCN	Hydrocyanic acid	CN^-	Cyanide
HNO_3	Nitric acid	NO_3^-	Nitrate
HNO_2	Nitrous acid	NO_2^-	Nitrite
H_2SO_4	Sulfuric acid	SO_4^{2-}	Sulfate
H_2SO_3	Sulfurous acid	SO_3^{2-}	Sulfite
H_2CO_3	Carbonic acid	CO_3^-	Carbonate
CH_3COOH	Acetic acid	CH_3COO^-	Acetate
H_3PO_4	Phosphoric acid	PO_4^{3-}	Phosphate
H_3PO_3	Phosphorous acid	PO_3^{3-}	Phosphite
$HClO_4$	Perchloric acid	ClO_4^-	Perchlorate
$HClO_3$	Chloric acid	ClO_3^-	Chlorate
$HClO_2$	Chlorous acid	ClO_2^-	Chlorite
HClO	Hypoclorous acid	ClO^-	Hypochlorite

Table 2 lists the names and formulas of some of the common Arrhenius bases.

Naming Arrhenius Bases

The Arrhenius bases are ionic compounds of metal and hydroxide ion, and their name starts with the name of the metal element followed by the name of the anion, i.e., hydroxide. For example, NaOH is sodium hydroxide.

Table 2: Names of some of the common Arrhenius bases

Formula	Name
LiOH	Lithium hydroxide
NaOH	Sodium hydroxide
KOH	Potassium hydroxide
$Ca(OH)_2$	Calcium hydroxide
$Sr(OH)_2$	Strontium hydroxide

Formula	Name
$Ba(OH)_2$	Barium hydroxide

This page titled 3.6.1: What is an Acid and a Base? is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.6.2: Brønsted–Lowry acids and bases

Some bases do not have hydroxide ions in their formula, yet they act as bases and neutralize acids. For example, ammonia (NH_3) and calcium carbonate ($CaCO_3$) do not contain hydroxide ions, but they neutralize acids. Further, Arrhenius's definition limits the acid-base reactions in the water medium. The acid-base reactions can take place in other mediums also, e.g., HCl –an acid, and NH_3 –a base can react with and neutralize each other in the gas phase also. The Brønsted–Lowry bordered the definition of acids and bases by including the bases mentioned above and also by including acid-base reactions in a non-aqueous medium.

Brønsted–Lowry's definition of acids and bases

Brønsted–Lowry's definition states that:

1. An acid is a proton donor, and
2. A base is a proton acceptor.

or example, HCl is an acid because it donates a proton to the water solvent.

$$\mathrm{HCl(g) + H_2O(l) \rightarrow Cl^-(aq) + H_3O^+(aq)}$$

Water is a base in the above reaction because it accepts a proton from the acid. In a reaction between HCl and NH_3:

$$\mathrm{HCl(g) + NH_3(\ g) \rightarrow NH_4Cl(s)}$$

HCl is an acid because it donates its proton to NH_3, and the NH_3 is a base because it accepts a proton, as shown in Fig. 6.2.1.



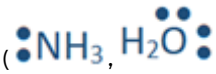
Figure 3.6.2.1: Proton transfer from HCl to NH_3 is an example of an acid-base reaction. Hydrochloric acid evaporates from liquid HCl in a beaker reacting with ammonia fumes coming from the test tube to produce ammonium chloride (white smoke). Source: Walkerma, Public domain.

Brønsted–Lowry's acids have ionizable protons that they donate to bases. Therefore, Brønsted–Lowry's acid is generally written as HA, where H^+ is the donatable proton, and A^- is the anion of the acid. Examples of acids are HCl, H_2SO_4, HNO_3, and CH_3COOH. Note that acetic acid has only one acidic proton that is attached to the O atom in the carboxylic acid group (–COOH). The rest of the protons attached to carbon atoms are not acidic. All organic acids have a carboxylic acid group (–COOH). Brønsted–Lowry's acid may have net +ve charge, no charge, or net –ve charge on it. For example, H_3O^+, HCl, and HSO_4^- are all acids because they can donate a proton to a base.

Mono-, di- and tri-protic acids

Acids that have only one acidic proton are mono-protic, e.g., **H**Cl, **H**NO_3, CH_3COO**H**, are mono-protic acids where the acid proton is shown in bold font. Some acids have two acidic protons –they are di-protic, e.g., **H**$_2SO_4$ and **H**$_2CO_3$ are di-protic. The acids with three acidic protons are tri-protic, e.g., **H**$_3PO_3$ is a tri-protic acid. For example, phosphoric acid (H_3PO_4) can dissociate and donate three protons, as shown in chemical reactions below:

$$\begin{aligned} &H_3PO_4 + H_2O \rightleftharpoons H_3O^+ + {H_2PO_4}^- \\ &{H_2PO_4}^- + H_2O \rightleftharpoons H_3O^+ + {HPO_4}^{2-} \\ &{HPO_4}^{2-} + H_2O \rightleftharpoons H_3O^+ + {PO_4}^{3-} \end{aligned} \tag{3.6.2.1}$$

The base accepts proton by making a bond with it. The bond is a pair of bonded electrons. Since the proton is a hydrogen atom without an electron, both electrons in the bond come from the base. The base must have a lone pair of electrons on it. The base is usually represented as  to emphasize a lone pair of electrons on it that is shown as a pair of dots. For example, ammonia, water, and hydroxide ion () are Brønsted–Lowry's bases, because each of these has an atom with lone pair or lone pairs of electrons on them.

Conjugate acid-base pairs

The acid-base reactions described above are one-way reactions, i.e., reactants go-to products almost 100%. However, the majority of the acid-base reactions are two ways, i.e., reactants form the products and the products react with each other and re-form the reactants. Double arrows between reactants and products represent the two ways reactions. For example, hydrofluoric acid (HF) is a weak electrolyte; it partially dissociates in water to form F^- and H_3O^+, and the products react to re-form the reactants, as shown in Fig. 6.2.2. In the reverse reaction, H_3O^+ is acting as an acid, and F^- is acting as a base. The acid and the base in the products are called conjugate acid and conjugate base, respectively. The acid HF becomes conjugate base F^- after removal of a proton, and the base H_2O becomes conjugate acid H_3O^+ after accepting a proton.

The conjugate acid-base pair

The conjugate acid-base pair is related to the loss and gain of H^+. For example, HF/F^- is a conjugate acid-base pair, and H_3O^+/H_2O is also a conjugate acid-base pair.

In other words, remove the acidic proton from an acid to get its conjugate base and add a proton to a base to get its conjugate acid.

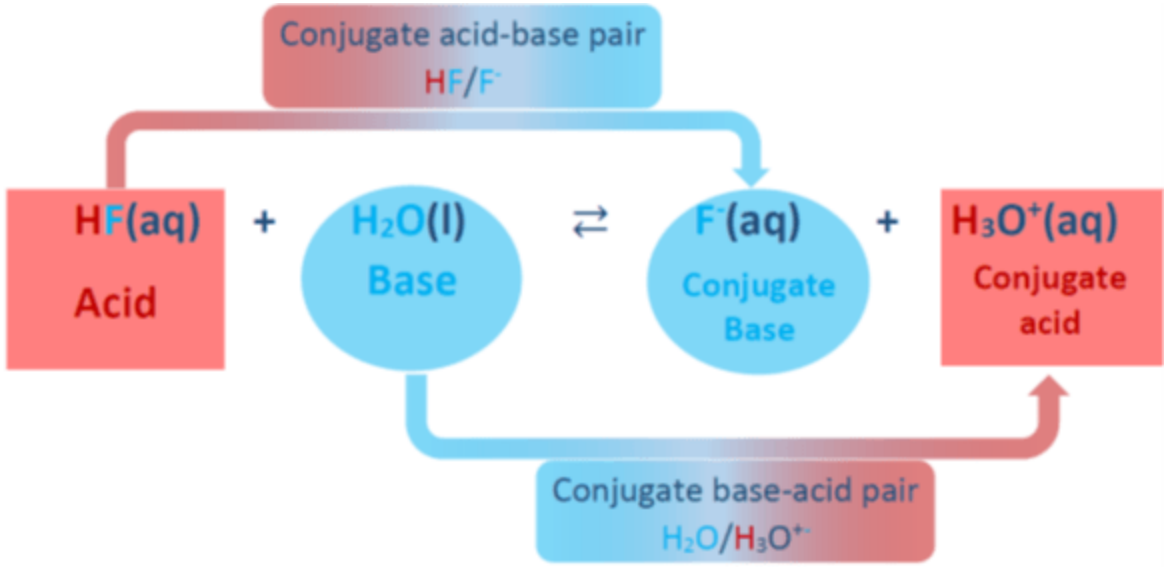


Figure 3.6.2.2: HF –an acid that loses a proton to form its conjugate base F^-, and water -acts as a base by accepting a proton to form its conjugate acid H_3O^+.

Another example is ammonia NH_3 which dissolves in water and accepts a proton to form its conjugate acid NH_4^+, as shown in Fig. 6.2.3. Water H_2O acts as an acid by donating a proton and forming its conjugate base OH^-. The two conjugate acid-base pairs in this reaction are NH_4^+/NH_3 and H_2O/OH^- that are related by loss and gain or an H^+.

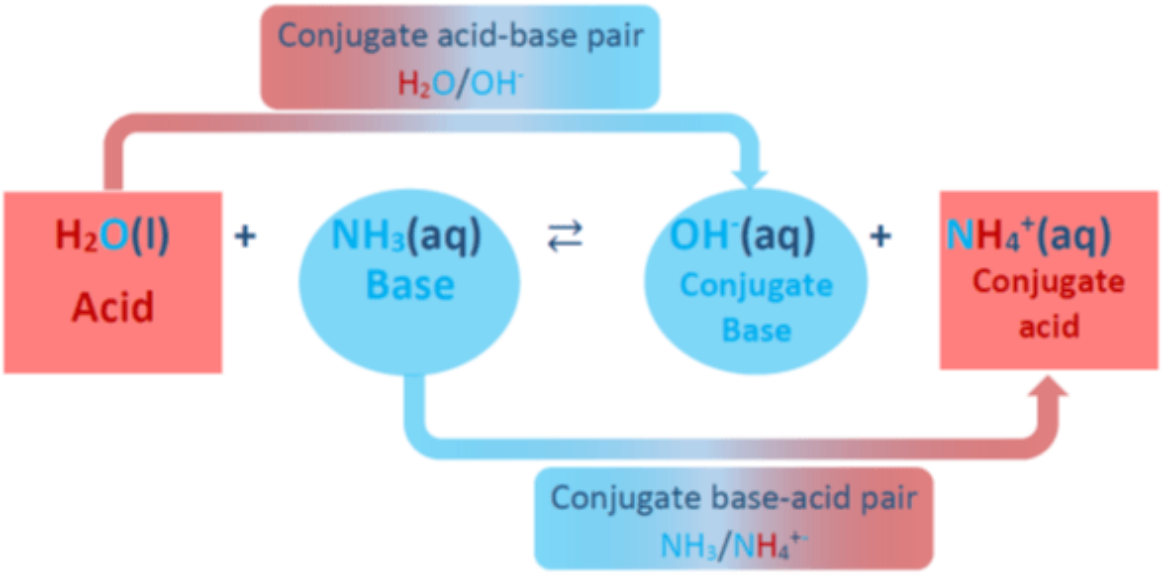


Figure 3.6.2.3: H_2O –acts as an acid that loses a proton to form its conjugate base OH^-, and NH_3 –a base accepts a proton to form its conjugate acid NH_4^+.

Example 3.6.2.1

Identify the conjugate acid-base pairs in the following reaction?

$$H_3PO_4(aq) + NH_3(aq) \rightleftarrows H_2PO_4^- + NH_4^+$$

Solution

1. Identify the substance that has donated a proton in the reactants –it is an acid.
2. Remove a proton from the acid to form its conjugate base: $H_3PO_4/H_2PO_4^-$.
3. Identify the substance that has accepted a proton in the reactants –it is a base.
4. Add a proton to the base to form its conjugate acid: NH_3/NH_4^+.

Note

Note that loss of a proton from an acid forms its conjugated base with the charge decreased by one, e.g., $H_3PO_4/H_2PO_4^-$, HSO_4^-/SO_4^{2-}, and NH_4^+/NH_3. Similarly, the gain of a proton by a base forms its conjugate acid with the charge increased by one, e.g., $HPO_4^{2-}/H_2PO_4^-$, HCO_3^-/H_2CO_3, and NH_3/NH_4^+.

Amphoteric substances

Water acts as a base in some reactions, e.g., with HF, and as an acid in some reactions, e.g., with NH_3.

Substances like water that can act as an acid and also as a base are called amphotericsubstances.

Other examples of amphoteric substances include HSO_4^-, HCO_3^-, and NH_3, as illustrated in Fig. 6.2.4.

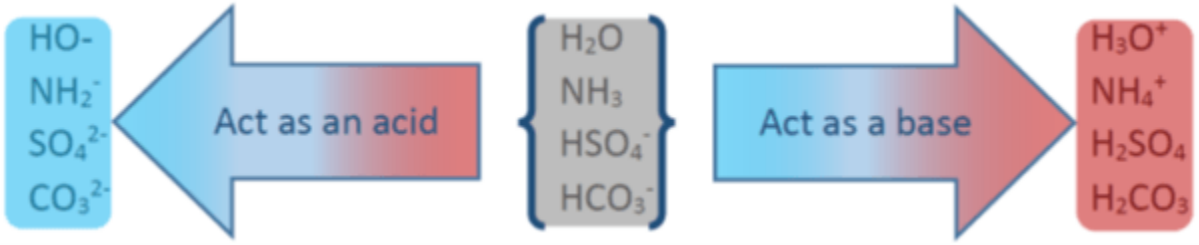


Figure 3.6.2.4: Examples of amphoteric substances that can act as an acid in some reactions and also as a base in other reactions.

This page titled 3.6.2: Brønsted–Lowry acids and bases is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.6.3: Strength of acids and bases

The strength of acid HA is the extent to which the acid dissociates into H^+ and A^- ions, as illustrated in Fig. 6.3.1.

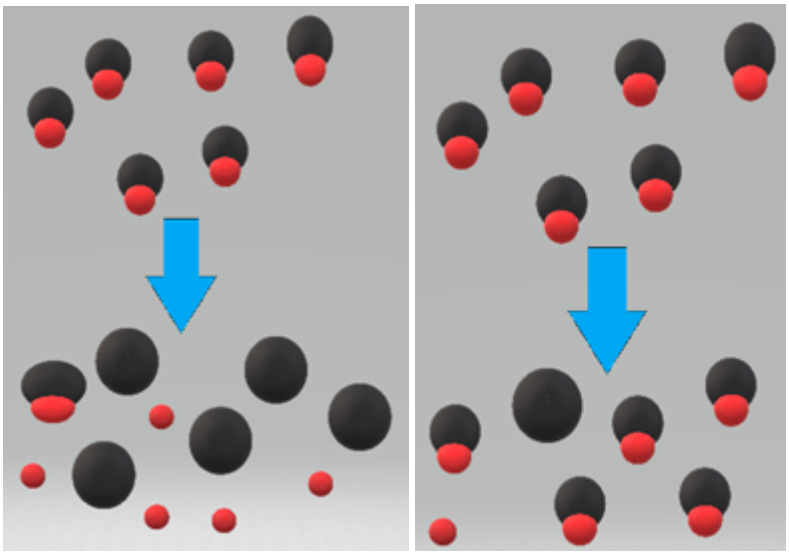


Figure 3.6.3.1: Image of strong acid mostly dissociating (left) and a weak acid partially dissociating into ions in water (right). Source: Cwszot / CC0

Strong acids

Strong acids, like HCl, almost 100% dissociate into ions when they dissolve in water.

$$\mathrm{HCl(g) + H_2O(l) \rightarrow H_3O^+(aq) + Cl^-(aq)}$$

One arrow is used to indicate that the reaction is nearly 100% complete.

Strong acids include $HClO_4$, H_2SO_4, HI, HBr, HCl, and HNO_3

Weak acids

Weak acids dissolve in water but partially dissociate into ions.

For example, acetic acid (CH_3COOH) is a weak acid, 1 M acetic acid dissolves in water, but only 0.4% of the dissolved molecules dissociate into ions, the remaining 99.6% remain undissociated, as illustrated in Fig. 6.3.2. and equation of the dissociation equilibrium below.

$$\mathrm{CH_3COOH(aq) + H_2O(I) \rightleftharpoons H_3O^+(aq) + CH_3COO^-(aq)}$$

Two arrows pointing in opposite directions are used for the dissociation of weak acids to indicate that the reaction is an equilibrium, i.e., two ways.

Often the arrows are not equal in size -the longer arrow points to acid-base pair that is weaker and present in a larger concentration at equilibrium than their conjugate pair.

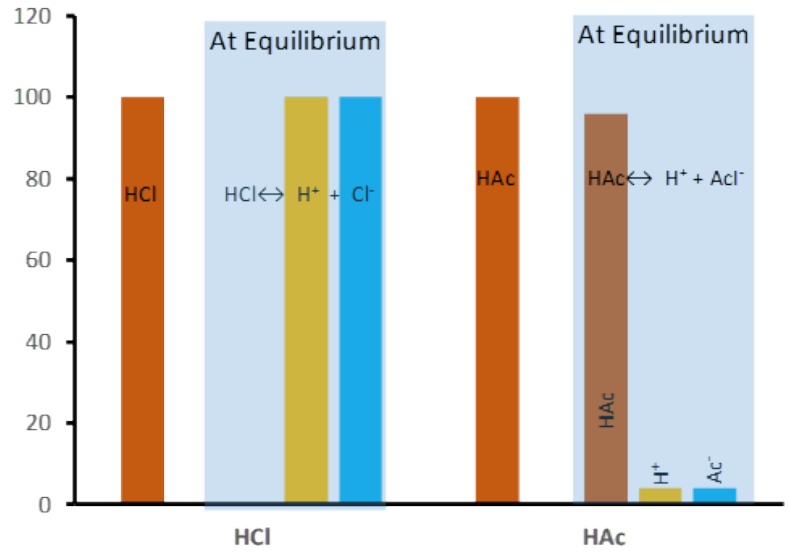


Figure 3.6.3.2: After dissolution in water. HCl –the strong acid is 100% dissociated into H^+ and Cl^- ions leaving no dissolved HCl molecules in water, but acetic acid (HAc) –the weak acid has a high concentration of HAc molecules and low concentration of H^+ and Ac^- ions.

Strong bases

Strong bases almost %100 dissociate into ions when dissolved in water. For example, NaOH is a strong base, and it dissociates almost 100% into ions in water.

Strong bases almost %100 dissociate into ions when dissolved in water. For example, NaOH is a strong base, and it dissociates almost 100% into ions in water.

$$\mathrm{NaOH(s)} \xrightarrow{\text{Water}} \mathrm{Na^+(aq) + OH^-(aq)}$$

One arrow is used for the dissolution of strong bases to indicate that the reaction is almost complete.

Strong bases include hydroxides of alkali metals, i.e., LiOH, NaOH, KOH, RbOH, CsOH, and hydroxides of heavy alkaline earth metals, i.e., $Ca(OH)_2$, $Sr(OH)_2$, and $Ba(OH)_2$.

The last three, i.e., the hydroxides of heavy alkaline earth metals, have low solubility in water, but the dissolved fraction exists as ions.

Weak bases

Weak bases partially dissociate into ions when dissolved in water.

For example, ammonia is a weak base –only 0.42% of the dissolved ammonia molecules dissociate into ammonium ions and hydroxide ions in water from a 1 M solution of ammonia.

$$\mathrm{NH_3(aq) + H_2O(I) \rightleftarrows NH_4^+(aq) + OH^-(aq)}$$

Weak bases in household use include ammonia (NH_3) in window cleaners, NaClO in bleach, Na_2CO_3 and Na_3PO_4 in laundry detergent, $NaHCO_3$ in tooth past, Na_2CO_3 in baking powder, $CaCO_3$ for use in lawns, $Mg(OH)_2$ and $Al(OH)_3$ in antacids and laxatives.

The weak bases mentioned above are all ionic compounds except ammonia. Ionic compounds are strong electrolytes, i.e., they

dissociate into ions almost 100% upon dissolution in water. It appears to contradict the fact that these ionic compounds are weak bases. It does not actually contradict, because the base properties do not refer to these ionic compounds, the base properties refer to the reactions of their polyatomic anions, i.e., ClO^-, CO_3^{2-}, and PO_4^{3-} with water, as shown in the reactions below:

$$ClO^- + H_2O \rightleftharpoons HClO + OH^-$$
$$CO_3^{2-} + 2H_2O \rightleftharpoons H_2CO_3 + 2OH^-, \text{ and}$$
$$PO_4^{3-} + 3H_2O \rightleftharpoons H_3PO_4 + 3OH^-$$

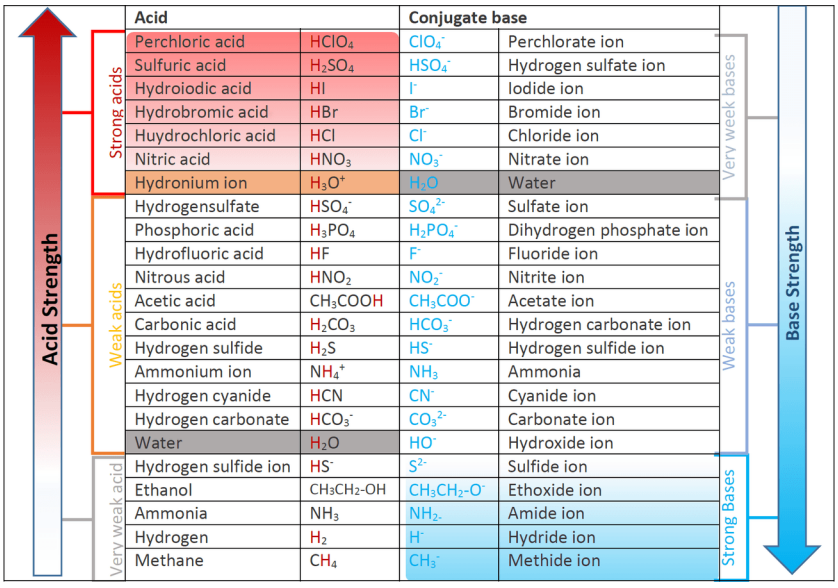
The above reactions are equilibrium reactions that are more favored in the revers than the forward direction, producing a small number of OH^- ions compared to the anion on the reactant sides. The last two examples, i.e., $Mg(OH)_2$ and $Al(OH)_2$ are classified as weak bases because they are considered insoluble in water. The solubility of $Mg(OH)_2$ is 0.00064 g/100 mL (25 °C), and the solubility of $Al(OH)_3$ is 0.0001 g/100 mL, which are in the range of insoluble ionic compounds.

> The solubility and the strength of acids and bases are two different things. A strong base may be less soluble, and a weak base may be more soluble or vice versa, but a dissolved strong base exists as ions only, and a dissolved weak base exists both as molecules and ions.

The relative strength of acid-conjugate base pair

> A general rule is that the stronger the acid, the weaker the conjugate base, and vice versa.

The conjugate bases of strong acids have negligible base strength, and the conjugate acids of strong basses have negligible acid strength. Fig. 6.3.3. illustrates the relative strengths of some acids and their conjugated bases.



Acid		Conjugate base	
Perchloric acid	$HClO_4$	ClO_4^-	Perchlorate ion
Sulfuric acid	H_2SO_4	HSO_4^-	Hydrogen sulfate ion
Hydroiodic acid	HI	I^-	Iodide ion
Hydrobromic acid	HBr	Br^-	Bromide ion
Huydrochloric acid	HCl	Cl^-	Chloride ion
Nitric acid	HNO_3	NO_3^-	Nitrate ion
Hydronium ion	H_3O^+	H_2O	Water
Hydrogensulfate	HSO_4^-	SO_4^{2-}	Sulfate ion
Phosphoric acid	H_3PO_4	$H_2PO_4^-$	Dihydrogen phosphate ion
Hydrofluoric acid	HF	F^-	Fluoride ion
Nitrous acid	HNO_2	NO_2^-	Nitrite ion
Acetic acid	CH_3COOH	CH_3COO^-	Acetate ion
Carbonic acid	H_2CO_3	HCO_3^-	Hydrogen carbonate ion
Hydrogen sulfide	H_2S	HS^-	Hydrogen sulfide ion
Ammonium ion	NH_4^+	NH_3	Ammonia
Hydrogen cyanide	HCN	CN^-	Cyanide ion
Hydrogen carbonate	HCO_3^-	CO_3^{2-}	Carbonate ion
Water	H_2O	HO^-	Hydroxide ion
Hydrogen sulfide ion	HS^-	S^{2-}	Sulfide ion
Ethanol	CH_3CH_2-OH	CH_3CH_2-O^-	Ethoxide ion
Ammonia	NH_3	NH_2^-	Amide ion
Hydrogen	H_2	H^-	Hydride ion
Methane	CH_4	CH_3^-	Methide ion

Figure 3.6.3.3: Strength of acid-conjugate base pairs relative to water as a reference.

Direction of acid-base equilibrium

In any Brønsted–Lowry acid-base reaction, the general rule is that a stronger acid and a stronger base tend to form a weaker acid and a weaker base.

For example, a dissociation reaction between HCl and H_2O is almost 100% complete because HCl is a stronger acid than H_3O^+ and H_2O is a stronger base than Cl^-:

$$HCl + H_2O \rightarrow H_3O^+ + Cl^-$$

The dissolution of acetic acid (CH_3COOH) and ammonia (NH_3) are equilibrium reactions because all the acids, bases, and their conjugates are in the weak acids or weak bases category. However, acetic acid and water dominate over their conjugates H_3O^+ and CH_3COO^- by 99.6:0.4 ratio (in 1 M acetic acid solution) because the conjugate acid $H_3O^+$is a stronger acid than CH_3COOH, and conjugate base CH_3COO^- is a stronger base than H_2O.

$$CH_3COOH(aq) + H_2O(l) \rightleftharpoons H_3O^+(aq) + CH_3COO^-(aq)$$

> The longer arrow, in the unbalanced equilibrium arrows, points to the acid-base pair in the reaction that exists in a higher concentration relative to their conjugates.

Similarly, ammonia (NH_3) and water (H_2O) dominate over their conjugates NH_4^+ and OH^- by ~99.6:0.4 ratio (1M ammonia solution) because the conjugate acid $NH_4^+$is a stronger acid than H_2O and conjugate base OH^- is a stronger base than NH_3.

$$NH_3(aq) + H_2O(I)) \rightleftharpoons NH_4^+(aq) + OH^-(aq)$$

This page titled 3.6.3: Strength of acids and bases is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.6.4: Acid-base equilibrium

Most of the acid-base reactions are equilibrium reactions, i.e., the reactants form the products, and the products react to re-form the reactants. The reaction is shown with double arrows to indicate that both the forward and the reverse reactions happen simultaneously.

Pre-equilibrium and equilibrium

Initially, there are only reactants present. The concentration of the reactants decreases over time as they convert to the products. The rate of the reaction is proportional to the concentration of the reactants. So, the rate of the forward reaction decreases over time. The products build up over time. The products react with each other to re-form the reactants, i.e., the reverse reaction. The rate of the reverse reaction increases over time as the concentration of the products increases, until the rate of reverse reaction becomes equal to the rate of the forward reaction, as illustrated in Fig. 6.4.1

- At the beginning of the equilibrium when the rate of the reverse reaction is slower than the rate of forward reaction is the pre-equilibrium phase.
- The equilibrium phaes starts at the point when the rate of reverse reaction becomes equal to the rate of the forward reaction

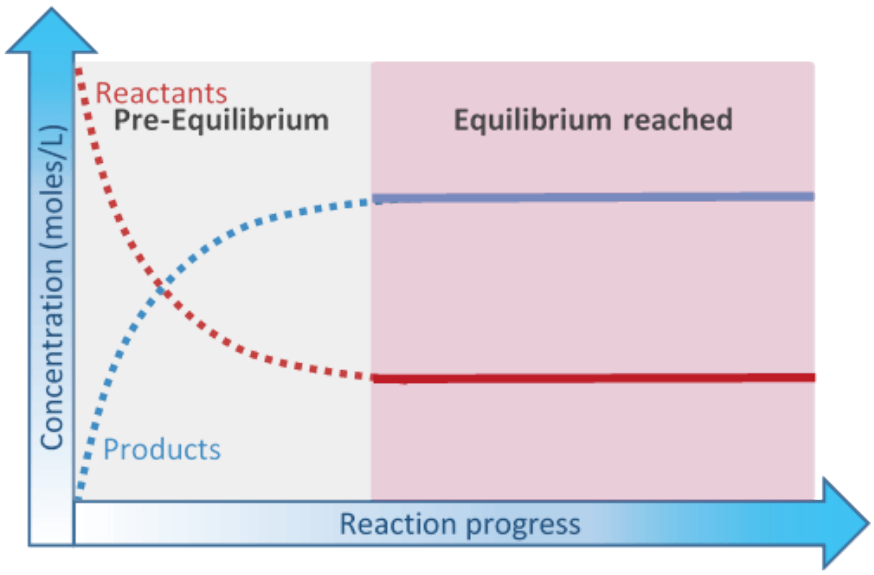


Figure 3.6.4.1: In pre-equilibrium, reactants react and their concentration, and, consequently, the rate of forward reaction decreases over time. The products build-up and the reverse reaction increases over time. When rate of forward reaction becomes equal to the rate of reverse reaction the equilibrium is reached and the concentration of reactants and products do not change.

At the point when the rate of reverse reaction becomes equal to the rate of the forward reaction, **equilibrium** has reached. The concentration of the reactants and products do not change at equilibrium because they are consumed and re-formed at the same rate –it is a dynamic equilibrium.

The equilibrium reactions are not limited to acid-base reactions; they are common in all types of chemical reactions. Fig. 6.4.2 illustrates the concentrations and rates of reactions changes in pre-equilibrium and equilibrium phases with the help of an actual chemical equilibrium reaction between a decomposition reaction of a colorless gas N_2O_4 and its reverse reaction, i.e., a combination reaction of brown color gas NO_2.

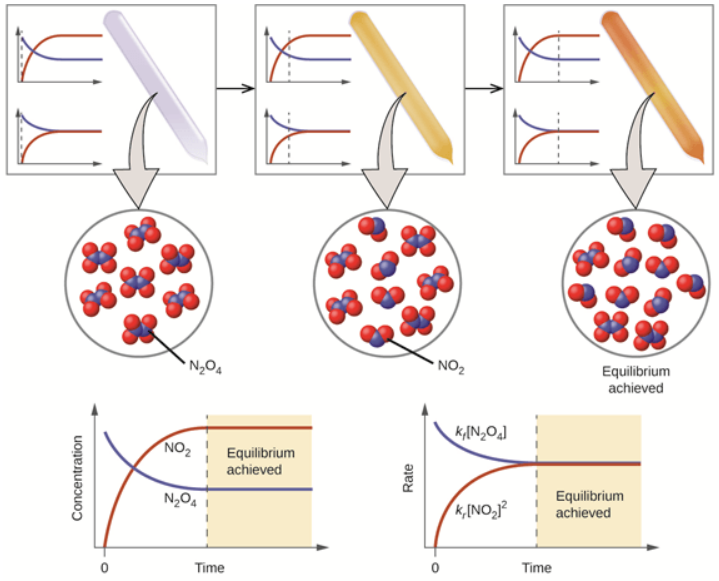


Figure 3.6.4.2: Top row: a sealed tube containing colorless gas N_2O_4 darkens as it decomposes to yield brown gas $NO_2$by the reaction: $\mathrm{N_2O_4 \rightleftarrows 2NO_2}$. The middle row: shows the models of the molecules at the timeline marked by the dotted lines in the graphs. Bottom row: shows concentration vs time (left graph); rate of forward reaction by blue curve and rate of reverse reaction by red curve (right graph). Source: OpenStax / CC BY (https://creativecommons.org/licenses/by/4.0)

What happens when a chemical equilibrium is disturbed?

The concentration of the reactants and products does not change in the mixture when the reaction is at equilibrium because the rate of the forward reaction is equal to the rate of the reverse reaction. Any change made at the equilibrium adds stress to the equilibrium. The system moves in a way to relieve the stress. For example, if one of the reactants is added to the mixture, the rate of forward reaction increases removing the added reactant until a new equilibrium establishes, i.e., the system relieves the stress by removing the added reactant.

Le Chatelier's principle

If a chemical equilibrium is disturbed, the rates of forward and reverse reactions change to relieve the stress and re-establish the equilibrium.

The stresses can be changes in the concentration, pressure, or temperature, as explained in the following section.

Effect of concentration on a chemical equilibrium

The following are the consequences of concentration changes on a chemical reaction at equilibrium.

1. If a reactant is added, the forward reaction increases to remove the reactant.
2. If a reactant is removed, the forward reaction decreases; consequently, the reverse reaction adds the reactant.
3. If a product is added, the reverse reaction increases to remove the product.
4. If a product is removed, the reverse reaction decreases; consequently, the forward reaction adds the product.

Fig. 6.4.3 illustrates the effects with the help of water level in two tanks connected through a conduit in the left tank representing reactants and in the right tank representing products.

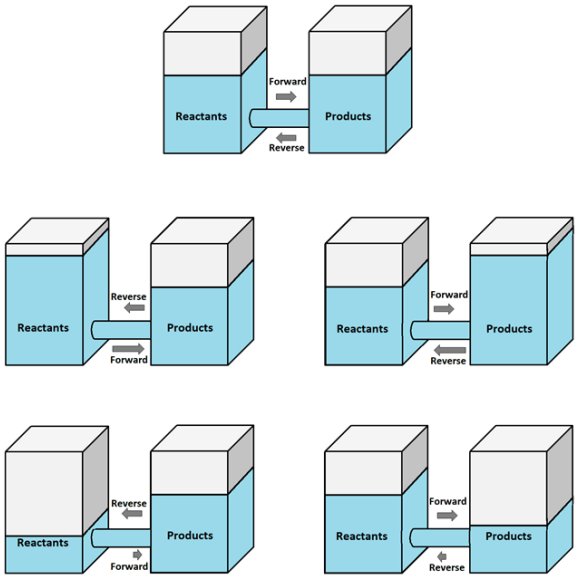


Figure 3.6.4.3: Illustration of Le Chatelier's principle using two water tanks connected through a conduit; 1) at equilibrium the rate of forward and reverse reactions are equal (the topmost), 2) when any reactant is increased the forward reaction rate increase (middle left), 3) when any product is increased the reverse reaction rate increases (middle right), 4) when any reactant is decreased the forward reaction rate decreases (bottom left), and 5) when any product is decreased the reverse reaction decreases (bottom right)

Fig 6.4.4 demonstrates the effect of concentration change on the chemical equilibrium between brown dichromate ions and yellow chromate ions:

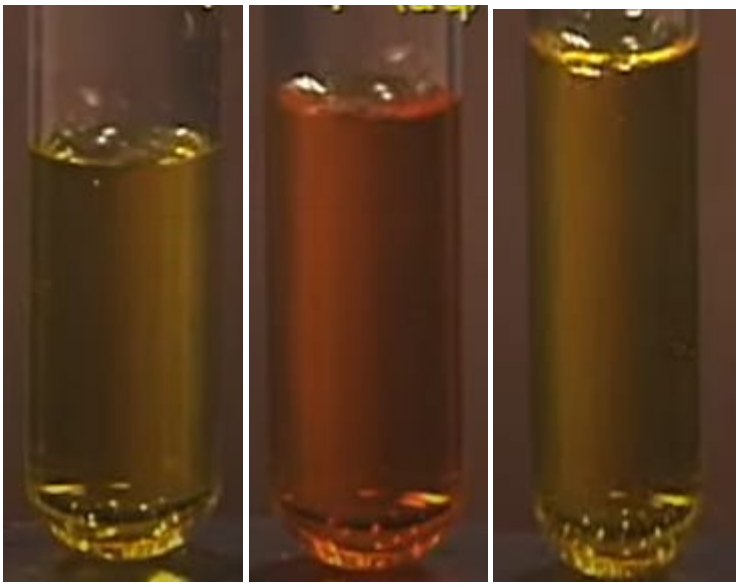


Figure 3.6.4.4: Demonstration of effect of concentration change on chemical equilibrium: $\mathrm{Cr_2O_7^{2-}(aq,\ brown\ ) + H_2O(I) \rightleftarrows 2CrO_4^{2-}(aq,\ yellow\ ) + 2H^+(aq)}$. Addition of H^+ to the tube on left by adding HNO_3 shifts the equilibrium to left making more of brown $Cr_2O_7^{2-}$ shown in the middle tube. Removal of H^+ by adding a base NaOH shift the reaction to the right making more of yellow CrO_4^{2-} shown in right most tube. Source: NCSSMDistanceEd / CC BY (https://creativecommons.org/licenses/by/3.0)

$$\mathrm{Cr_2O_7^{2-}(aq,\ brown\ ) + H_2O(I) \rightleftarrows 2CrO_4^{2-}(aq,\ yellow\ ) + 2H^+(aq)}$$

The addition of acid to the test tube on the left increases H^+ in the system, increasing reverse reaction, which can be observed by the increased brown color in the middle test tube. Then, the addition of a base removes H^+ from the system by an acid-base reaction: $\mathrm{H^+ + OH^- \rightarrow H_2O}$. The decrease in H^+ shifts the equilibrium to the product side, which can be observed by the increased yellow color in the test tube on the right.

A practical example of the effect of concentration on a chemical equilibrium is the binding of oxygen (O_2) with hemoglobin (Hb) during the breathing process.

$$\mathrm{Hb(aq) + O_2(\ g) \rightleftarrows HbO_2(aq)}$$

The concentration of oxygen is higher in the lunges that shift the equilibrium to the product side, binding more oxygen with the hemoglobin. When the blood arrives in tissues, the concentration of oxygen is lower in tissues, causing the equilibrium to shift to the reactant side, releasing the oxygen.

The oxygen concentration decreases as the altitude increases. The mountain climbers may experience hypoxia, i.e., inadequate supply of oxygen to the body because the lower level of oxygen at high altitude may shift the equilibrium to the right, resulting in less binding of oxygen with the hemoglobin in the lungs. The body reacts by producing more hemoglobin, but it takes about 10 days for the body to re-adjust the hemoglobin level in the blood. Peoples living at higher altitudes usually have a higher level of hemoglobin in their blood for the reasons described above.

Effect of pressure on a chemical equilibrium

Changes in pressure do not affect the concentration of solids and liquid. However, an increase in pressure decreases the volume, and, consequently, increases the concentration of gases, as illustrated in Fig. 6.4.5.

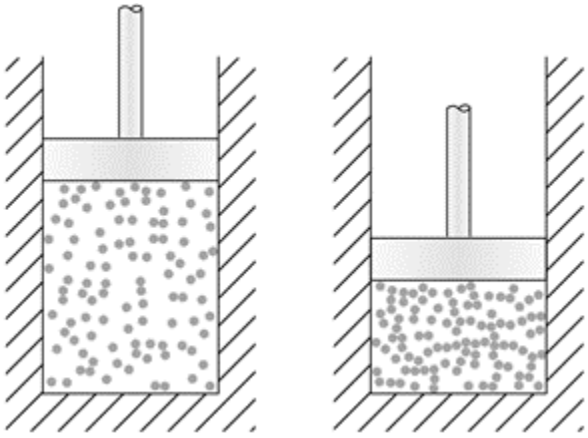


Figure 3.6.4.5: Illustration of increase in pressure causes decrease in the volume and increase in the concentration of gas molecules. Source: Olivier Cleynen / CC0

> If a chemical equilibrium involves gases, an increase in pressure has the same effect as increases in the gaseous reactant or product—the equilibrium shifts in the direction where there are fewer moles of gases.

Fig. 6.4.6 demonstrated this effect for the equilibrium between colorless N_2O_4 gas and brown color NO_2 gas:

$$\mathrm{N_2O_4(\ g,\ colorless\ ) \rightleftarrows 2NO_2(\ g,\ brown\ )}$$

The color becomes light in going from a syringe on the right to the syringe in the middle. It can be explained based on the fact that the concentration of gases decreased upon an increase in volume as a result of a decrease in pressure. Then the color becomes darker over time, as shown by the syringe on the right. The darker color indicates that the equilibrium has shifted to the product side where there are more moles of gas, to relieve the stress.

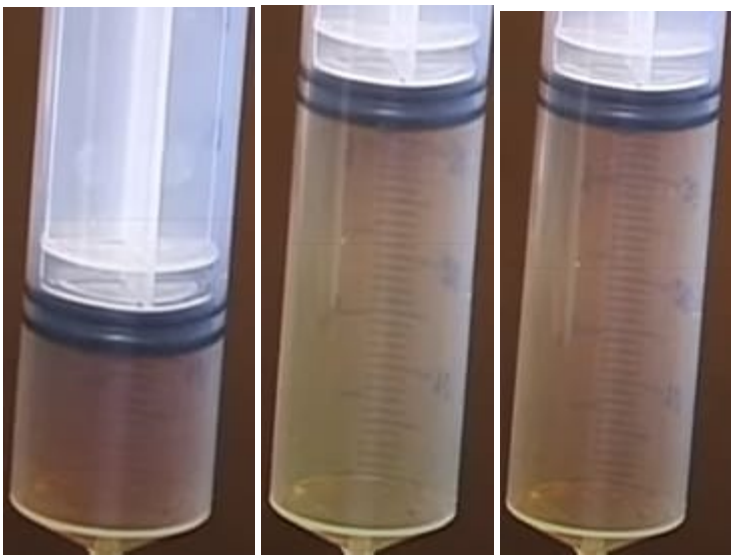


Figure 3.6.4.6: Decrease in pressure on the equilibrium: $\mathrm{N_2O_4(\ g,\ colorless\ ) \rightleftarrows 2NO_2(\ g,\ brown\ )}$, Middle syringe → syringe on the right, increases the forwared reaction to release the stress. Source: NCSSMDistanceEd / CC BY (https://creativecommons.org/licenses/by/3.0)

Effect of temperature on a chemical equilibrium

Exothermic reaction releases heat, i.e., heat is one of the products. If the forward reaction is exothermic, the reverse reaction must be endothermic by the same amount and vice versa.

> When a reversible reaction at equilibrium is disturbed by increasing temperature, the equilibrium shifts in the endothermic direction of the reaction to remove the heat and vice versa.

In the equilibrium between colorless N_2O_4 and brown color NO_2 gases, the reaction is endothermic in the forward direction:

$$\mathrm{N_2O_4}(g,\ \text{colorless}\ ) + \text{Heat} \rightleftarrows 2\mathrm{NO_2}(g,\ \text{brown}\ )$$

An increase in temperature shifts the equilibrium in the endothermic direction to relieve the stress, as demonstrated in Fig. 6.4.7.



Figure 3.6.4.7: Increase in temperature shifts the equilibrium in the endothermic direction producing more of the colored NO_2 in the following equilibrium: $\mathrm{N_2O_4}(g,\ \text{colorless}\ ) + \text{Heat} \rightleftarrows 2\mathrm{NO_2}(g,\ \text{brown}\ )$. An overlay of the same 99.9% pure NO_2/N_2O_4 sealed in an ampoule. From left to right -196 °C, 0 °C, 23 °C,

35 °C, 50 °C. Source: Eframgoldberg / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

This page titled 3.6.4: Acid-base equilibrium is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

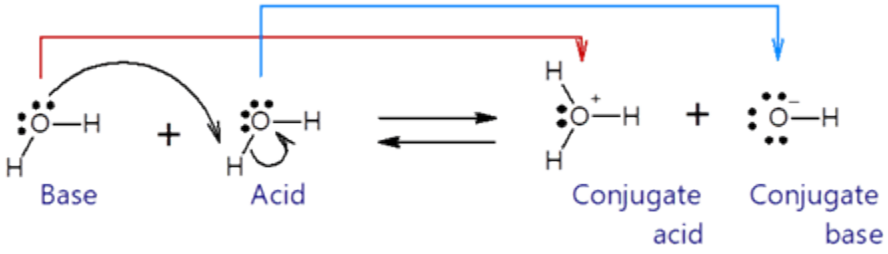
3.6.5: Dissociation of water

How do water molecules dissociate?

Water is an amphoteric substance, which means water can accept a proton acting as a base, and it can also donate A proton acting as an acid.

> About one water molecule in half a billion dissociates into an OH^- ion by losing a proton to another water molecule. The molecule that receives a proton becomes H_3O^+.

The dissociation of water is an equilibrium reaction in which one water molecule donates its proton to another water molecule.



The water molecule that receives proton is acting as a base, and it converts to conjugate acid H_3O^+. The other water molecule that donates a proton is acting as an acid, and it converts to conjugate base OH^-. The arrows in the reaction show that the base uses one of its lone pairs of electrons to make a bond with proton, and the previous bond pair of electrons turns into a third lone pair of electrons on the oxygen atom of the base. The reaction is reversible, i.e., the conjugate acid (H_3O^+) and the conjugate base (OH^-) react to re-form the two water molecules.

Water dissociation constant

The dissociation of water is an equilibrium reaction. It means the rate of the forward reaction is equal to the rate of the reverse reaction and the concentration of the reactants and products do not change at equilibrium. The molar concentration of H_3O^+ represented as $[H_3O^+]$ is equal to 10^{-7} M in a pure water sample at 25 ^oC, where M is in moles/Liter. The molar concentration of OH^- represented as $[OH^-]$ is equal to the molar concentration of H_3O^+ in pure water, i.e., $[H_3O^+] = [OH^-] = 10^{-7}$M.

> The product of the molar concentration of H_3O^+ and OH^- in water is a constant called **water dissociation constant K_w** equal to 10^{-14} at 25 ^oC, i.e.:
>
> $$K_w = [H_3O^+]\,[OH^-] = (10^{-7})\,(10^{-7}) = 10^{-14}$$

Neutral, acidic, or basic aqueous solutions

> A solution that has an equal concentration of H_3O^+ and OH^-, each equal to 10^{-7} M, is a **neutral solution**.

An acidic solution has an acid dissolved in water. When an acid dissolves in water it dissociates adding more H_3O^+. The $[OH^-]$ must decrease to keep the K_w constant.

A solution that has [H_3O^+] more than 10^{-7}, and [OH^-] less than 10^{-7} is an **acidic solution**.

A basic solution has a base dissolved in water. When a base dissolves in water it dissociates adding more OH^-. The [H_3O^+] must decrease to keep the K_w constant.

A solution that has [H_3O^+] less than 10^{-7}, and [OH^-] more than 10^{-7} is a **basic solution**.

Calculations of [H_3O^+] and [OH^-] based on K_w

The water dissociation constant remains the same whether the aqueous solution is neutral, acidic, or basic, i.e.:

$$\mathrm{K_w} = \left[\mathrm{H_3O^+}\right]\left[\mathrm{OH^-}\right] = \left(10^{-7}\right)\left(10^{-7}\right) = 10^{-14} \text{ at } 25^\circ\mathrm{C}$$

Therefore, if the molar concentration of hydronium ions [H_3O^+] is known, the molar concentration of hydroxide ions [OH^-] can be calculated using the following formula:

$$\left[\mathrm{H_3O^+}\right] = \frac{\mathrm{K}_w}{\left[\mathrm{OH^-}\right]} = \frac{10^{-14}}{\left[\mathrm{OH^-}\right]}$$

Similarly, if the molar concentration of hydroxide ions [OH^-] is known, the molar concentration of hydronium ions [OH^-] can be calculated using the following formula:

$$\left[\mathrm{OH^-}\right] = \frac{K_w}{\left[\mathrm{H_3O^+}\right]} = \frac{10^{-14}}{\left[\mathrm{H_3O^+}\right]}$$

When a strong acid like HCl dissolves in water, it dissociates ~100% into ions. Therefore, the [H_3O^+] is equal to the molar concentration of the acid. The amount H_3O^+ added by dissociation of water molecules is very small compared to that coming from the dissociation of a strong acid and can be neglected. Similarly, when a strong base like NaOH dissolves in water, it dissociates ~100% into ions. Therefore, the [OH^-] is equal to the molar concentration of the base.

Note

When a weak acid or a weak base dissolves in water, it partially dissociates into ions. Therefore, the [H_3O^+] or the [OH^-] in the cases of weak acids and weak bases has to be determined experimentally for the calculations.

Example 3.6.5.1

Calculate the concentration of OH^- ions in a 0.10 M HNO_3 solution?

Solution

The HNO_3 is a strong acid. Therefore, [HNO_3] = 0.10 M = [H_3O^+]. Desired [OH^-] = ?

Formula: $\left[\mathrm{OH^-}\right] = \frac{10^{-14}}{\left[\mathrm{H_3O^+}\right]}$

Plug in values an calculate: $\left[0\mathrm{H^-}\right] = \frac{10^{-14}}{0.10} = 10^{-13}\ \mathrm{M}$

Example 3.6.5.2

A vinegar solution has $[H_3O^+]$ = 2.0 x 10^{-3}. a) What is the hydroxide ion concentration in the vinegar solution? b) is the solution acidic, basic, or neutral?

Solution

a) Given $[H_3O^+]$ = 2.0 x 10^{-3}. Desired $[OH^-]$ = ?

Formula: $[\mathrm{OH^-}] = \frac{10^{-14}}{[\mathrm{H_3O^+}]}$

Plug in values and calculate: $[0H^-] = \frac{10^{-14}}{2.0\times10^{-3}} = 5.0 \times 10^{-12}\mathrm{M}$

b) The solution is acidic because $[H_3O^+] > [OH^-]$.

Example 3.6.5.3

Calculate the value of $[H^3O^+]$ and $[OH^-]$ in a 0.010 M NaOH solution?

Solution

The NaOH is a strong base. Therefore [NaOH] = 0.010 M = $[OH^-]$. Desired $[H_3O^+]$ = ?

Formula: $[\mathrm{H_3O^+}] = \mathrm{K_w}/\,[\mathrm{OH^-}] = 10^{-14}/\,[\mathrm{OH^-}]$

Calculations: $[\mathrm{H_3O^+}] = 10^{-14}/0.010 = 10^{-12}\mathrm{M}$

Example 3.6.5.4

a) Calculate the $[H_3O^+]$ in an ammonia solution that has $[OH^-]$ = 4.0 x 10^{-4} M? b) Is the

solution acidic, basic, or neutral?

Solution

a) Given $[OH^-]$ = 4.0 x 10^{-4}. Desired $[H_3O^+]$ = ?

Formula:$[\mathrm{H_3O^+}] = \mathrm{K_w}/\,[\mathrm{OH^-}] = 10^{-14}/\,[\mathrm{OH^-}]$

Calculations: $[0H^-] = \frac{10^{-14}}{4.0\times10^{-4}} = 2.5 \times 10^{-11}\mathrm{M}$

b) The solution is basic because $[H_3O^+] < [OH^-]$.

This page titled 3.6.5: Dissociation of water is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.6.6: The pH

What is pH?

The acidity of an aqueous solution relates to H_3O^+ ion concentration, but there are two problems with it. The first problem is that the molar concentration of hydronium ion [H_3O^+] varies over a wide range, usually from 1-to-0.00000000000001 which is not easy to comprehend. Log scale solves the problem, e.g., if a number increases from 1 to 1000, it increases from 0 to 3 on a log scale to the base 10. The second problem is that the [H_3O^+] is usually a small number between 0 and 1. The log of a number that is between 0 and 1 is a negative number that is not easy to grasp mentally. If a number is between 0 and 1, its reciprocal is more than 1, and its log is a +ve number.

The **pH** is defined as a log of reciprocal of the molar concentration of hydronium ions [H_3O^+]. It can also be stated that the pH is a negative log of the molar concentration of hydronium ion [H_3O^+]. The mathematical form of the pH is the following.

$$pH = \log \frac{1}{[\mathrm{H_3O^+}]} = -\log [\mathrm{H_3O^+}]$$

A pH is usually a positive number between 0 to 14.

Calculating pH

First, determine the molar concentration of hydronium ion [H_3O^+]. Than take log base 10 of [H_3O^+] and change the sign of the answer, i.e.:

$$pH = -\log [\mathrm{H_3O^+}]$$

Example 3.6.6.1

Calculate the pH of neutral water that has [H_3O^+] = 10^{-7}M?

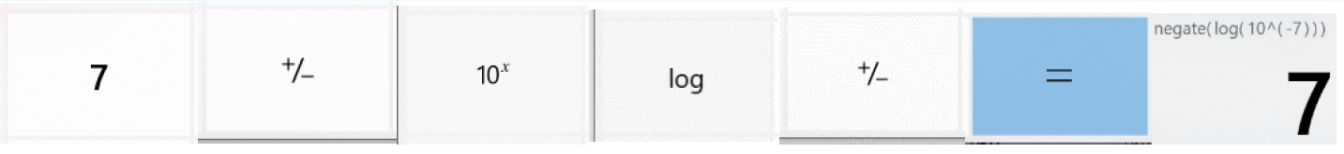
Solution

Given [H_3O^+] = 10^{-7} M, desired pH = ?

Formula: $pH = -\log [\mathrm{H_3O^+}]$

Plug in values and calculate: $pH = -\log \left(10^{-7}\right) = 7.0$

The keys-sequence for the calculation on a Microsoft Window's scientific calculator is (the last box is the output of the keys-sequence):



Significant figures in pH value

The given number in scientific notation is **1** x 10^{-7} which has one significant figure shown by the coefficient part of the number in bold font. **The number of decimal places in the pH number shows the significant figures**, i.e., 7.**0** has one decimal place shown in bold font, which means it has one significant figure.

Example 3.6.6.2

Calculate the pH of the 0.010 M HCl solution?

Solution

Given [HCl] = 0.010 M = [H_3O^+], desired pH = ?

Formula: $pH = -\log[H_3O^+]$

Plug in values and calculate: $pH = -\log(0.010) = 2.00$

The keys-sequence for the calculation on a Microsoft Window's scientific calculator is (the last box is the output of the keys-sequence):



Note that the given number 0.0**10** M has two significant figures shown in bold fonts. So, the answer pH 2.**00** also has two significant figures shown in bold fonts.

Example 3.6.6.3

Calculate the pH of 0.10 M HCl solution?

Solution

Given [HCl] = 0.10 M = [H_3O^+], desired pH = ?

Formula: $pH = -\log[H_3O^+]$

Calculations: $pH = -\log(0.10) = 1.00$

As the acid 0.01M and 0.10M is gradually added to increase the acid from neutral water in example 1 to 0.01M in example 2 to 0.10 M in example 3, , the pH gradually decreased from 7 to 2 to 1.

The pH of the acidic solution is less than 7. The more acidic the solution, the lower the pH.

The pH decreases as the [H3O+] increases because the reciprocal of [H3O+] is used in the pH calculation. The larger the given number, the smaller the reciprocal; it translates to "the more acidic the solution, the lower the pH.

Example 3.6.6.4

Calculate the pH of 0.010 M NaOH solution?

Solution

Given [NaOH] = 0.010 M = [OH^-], desired pH = ?

First, calculate [H_3O^+] from the give [OH^-]: $[H_3O^+] = 10^{-14}/[OH^-] = 10^{-14}/0.010 = 1.0 \times 10^{-12}$

Than calculate the pH: $pH = -\log[H_3O^+] = -\log(1.0 \times 10^{-12}) = 12.00$

Note that the given number 0.0**10** M has two significant figures shown in bold fonts. So, the answer pH 12.**00** also has two significant figures shown in bold fonts.

Example 3.6.6.5

Calculate the pH of 0.10 M NaOH solution?

Solution

Given [NaOH] = 0.10 M = [OH^-], desired pH = ?

First, calculate [H_3O^+] from the give [OH^-]: $10^{-14}/[OH^-] = 10^{-14}/0.10 = 1.0 \times 10^{-13}$

Formula: [H_3O^+] = 10^{-14}/[OH^-] = 10^{-14}/0.10 = 1.0 x 10^{-13}.

Than calculate the pH: $pH = -\log[H_3O^+] = -\log(1.0 \times 10^{-13}) = 13.00$

As the base, 0.010M and 0.10M are gradually added in the previous two exams to the neutral water at pH 7, the pH increases from 7 to 12 to 13.

The pH of the basic solution is more than 7. The more basic the solution, the higher the pH.

pH measurement

The pH is usually measured in laboratories by digital pH meters. The electrode of the pH meter is first calibrated with solutions of know pH values, and then the electrode is dipped in the test solution to read its pH value. Universal pH indicator papers are available that turn to a specific color when placed in the solution. The pH is read by matching the color of the test paper with the color on the chart. Fig. 6.6.1 shows a digital pH meter and two universal pH indicating papers commonly used for pH measurements in laboratories.







Figure 3.6.6.1: pH/Ion Meter (left) and universal pH indicator papers (middle and right). Source: Datamax / Public domain; Bordercolliez / CC0; Fredquintao~commonswiki, CC-BY-SA-2.5.

The pH indicators are weak acids or weak bases that change color in a specific pH range. A few drops of the solution of a pH indicator paper are added to the test solution in which pH is changed by adding acid or base. The change in the color of the solution indicates the pH range in which the indicator changes the color. Figure 6.6.2 shows the colors and the color transition ranges of some of the common pH indicators.

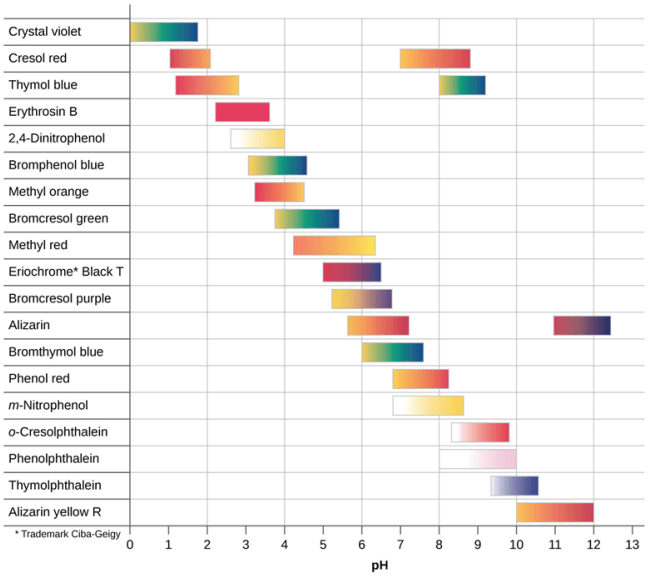


Figure 3.6.6.2: The colors and the pH ranges of color transition of some of the common pH indicators. Source: OpenStax / CC BY (https://creativecommons.org/licenses/by/4.0)

Calculating hydronium ion concentration from pH

The formula for calculating the molar concentration of hydronium ions [H_3O^+] is obtained by rearranging the pH formula. $pH = -\log[H_3O^+]$, re-arranges to: $[H_3O^+] = \text{antilog}(-pH) = 10^{-pH}$.

Example 3.6.6.6

Calculate [H_3O^+] of a urine sample that has pH 7.56?

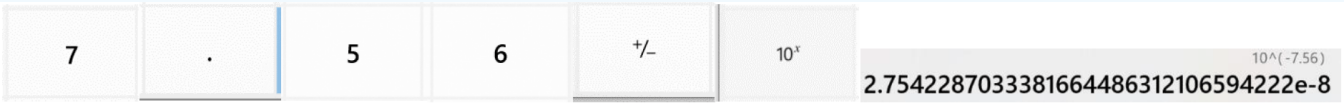
Solution

Given: pH = 7.56, Desired: [H_3O^+] = ?

Formula: $[H_3O^+] = 10^{-pH}$.

Plug in the values and calculate: $[H_3O^+] = 10^{-pH} = 10^{-(7.56)} = 2.8 \times 10^{-8} M$

The keys-sequence for the calculation on a Microsoft Window's scientific calculator is (the last box is the output of the keys-sequence):



The calculated answer in the above example is a lengthy number, but it is rounded to two significant figures shown in bold font in **2.8**$\times 10^{-8} M$, because the given number pH 7.**56** has two significant figures shown in bold font.

Significance of pH Scale

The pH scale varies from ~0 to ~14. The pH of 7 is neutral, pH more than 7 is basic, and pH less than 7 is acidic, as illustrated in Fig. 6.6.3. Different foods have different pH values, as shown in Fig. 6.1.1. Similarly, several of the acids and bases in household use have a specific pH range, as shown in Fig. 6.6.3.

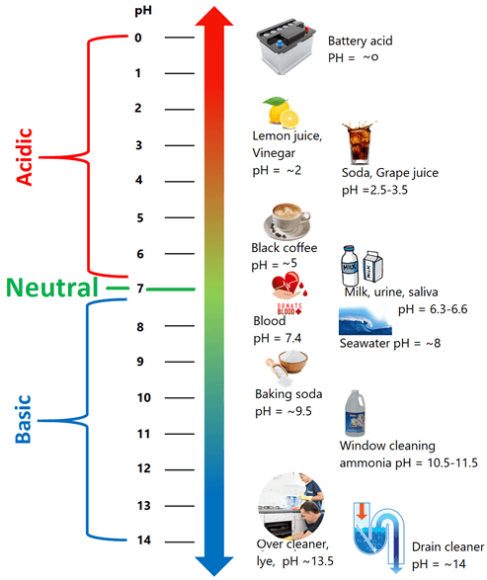


Figure 3.6.6.3: pH scale and pH of some common substances. pH 7 is neutral, pH less than 7 is acidic, and pH more than 7 is basic.

The control of pH is essential in the proper functioning of biological systems. Plants thrive if the soil does not have too acidic or too basic a pH. Lacks and revers have a specific range of pH in which the aquatic life can survive. US natural water has a pH in the range of 6.5 to 8.5 range. Seawater has a specific narrow range of pH 7.5 to pH 8.4 in which the life in the sea can function correctly.

The pH of body fluids

Human body fluids vary in pH, as shown in Fig. 6.6.4. Saliva in the mouth is slightly acidic, but the stomach has the lowest pH in the body. The strongly acidic pH in the stomach helps in digesting some foods, and it also helps in killing bacteria that may enter the stomach through the foods. When the food enters the large intestine, the pH changes to basic which helps in digesting the foods that could not be digested in the acidic environment of the stomach. The pH changes to more basic in the small intestine.

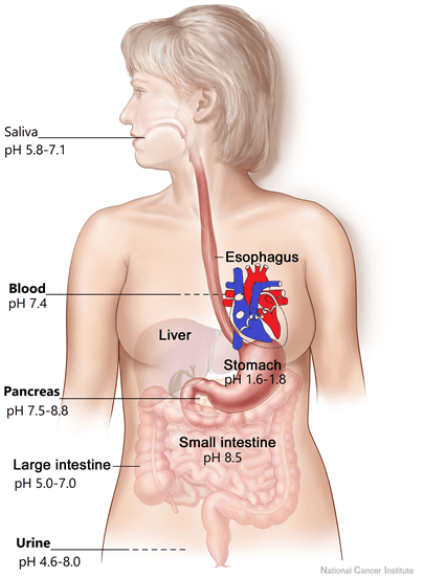


Figure $3.6.6.4$: The pH in various human body fluids is shown. Source: Modified from National Cancer Institute / Public domain

The blood has a pH of 7.4, and it can vary in a small range of 7.35 to 7.45. If the blood pH goes outside of the 7.35 to 7.45 range it can result in medical problems. Enzymes in the body need a specific pH range because hydrogen bonding plays a vital role in the structures needed for proper functioning. pH changes affect the hydrogen bonding and can make the enzymes less active or may inactivate them. The pH of blood is maintained by a complex action of buffers that are described in the later sections. The pH of urine can vary over a broad range from 4.6 to 8, depending on the recent diet and exercises.

Acid rain

Acidic gases like NO, NO_2, N_2O_4, SO_2 are released into the environment primarily during the combustion of fossil fuels. These gases dissolve in the rainwater and make the rain acidic. For example, sulfur dioxide dissolves in water and makes sulfurous acid:

$$\mathrm{SO_2(\ g) + H_2O(I) \rightleftarrows H_2SO_3(aq)}$$

The acid rain damages the environment by making the soil, river water, and lack-water acidic. The acidic soil and water, in turn, affect the plants and aquatic life. The acid water also reacts with calcium carbonate and corrodes metals that are responsible for damage to the sculpture and other structures, as illustrated in Fig. 6.6.5, Fig. 6.6.6, and Fig. 6.6.7.

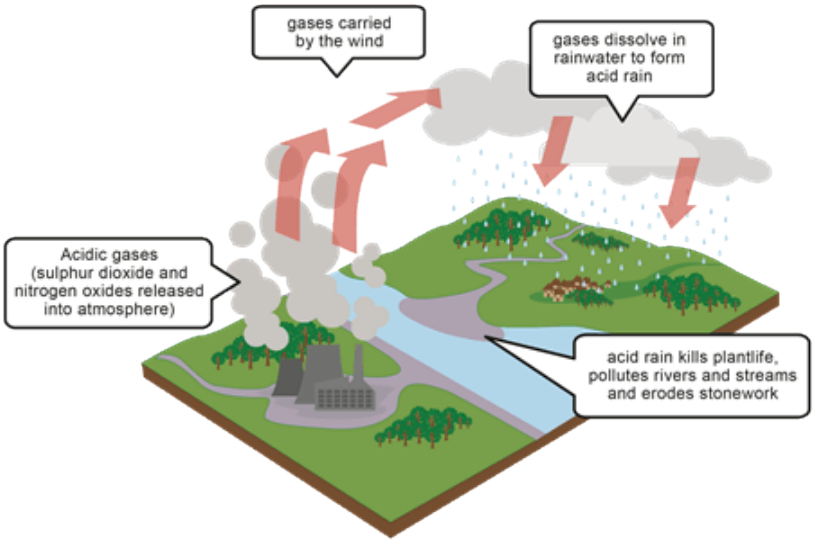
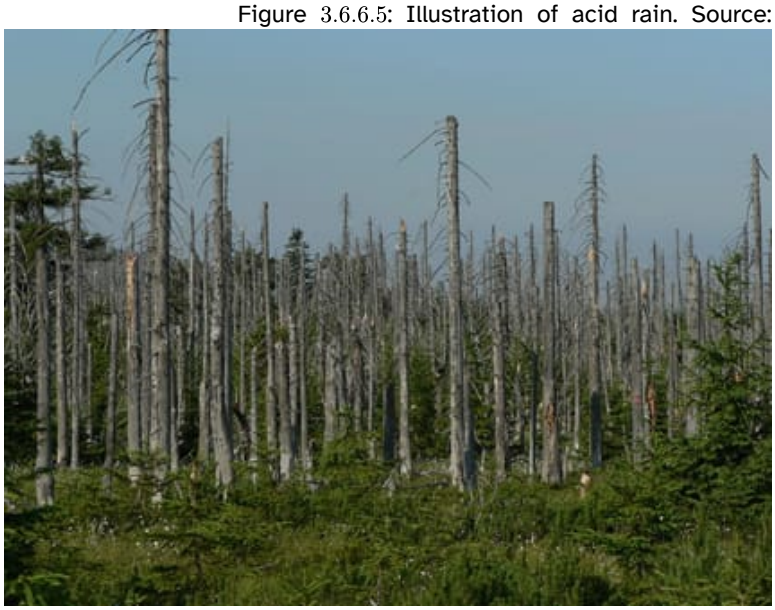


Figure 3.6.6.5: Illustration of acid rain. Source:



https://www.hiclipart.com/free-trans...bacyp/download

Figure 3.6.6.6: Effect of acid rain on forest . Source: Lovecz / Public domain.



Figure 3.6.6.7: PEffect of acid rain on monuments.

Source: Nino Barbieri / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/)

This page titled 3.6.6: The pH is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.6.7: Acid-base reactions

Acid-base reactions are chemical reactions that involve the transfer of a proton (H^+). Examples include the reactions of acids with metals, carbonates, and Arrhenius bases, described in the following.

Reactions of acids with metals

Metals tend to give out an electron and become cations. The majority of the metals, called reactive metals, give out electrons to protons in the acids and release H_2 gas. For example, Fig. 6.7.1 shows magnesium reacting with hydrochloric acid by the following reaction:

$$\mathrm{Mg(s) + 2HCl \rightarrow MgCl_2(aq) + H_2(\ g) \uparrow}$$



Figure 3.6.7.1: Mg reacting with HCl and H_2 gas bubbling out. Source: https://youtu.be/cpy_Zh-8sKA, CC BY

Alkali metals like Na give out one electron, alkaline earth metals like Mg give out two electrons, and aluminum gives out three electrons. A proton can accept one electron. Therefore, the number of acidic protons should be equal to the number of electrons given out to balance the electron transfer in these reactions.

Note

Noble metals or jewelry metals like gold, silver, platinum, copper are exceptions –they do not react with acids.

Example 3.6.7.1

write a balanced equation for the reaction of aluminum with HBr?

Solution

Step 1) write the equation with the correct formulas of reactants and products.

$$\mathrm{Al(s) + HBr(aq) \rightarrow AlBr_3(aq) + H_2(\ g)}$$

Note that Al gives out three electrons to make Al^{3+}; that is why three Br^- are attached to it in the product to balance the charge.

Step 2) balance the electrons lost by the metal with the number of acidic protons.

$$Al(s) + 3HBr(aq) \rightarrow AlBr_3(aq) + H_2(\ g)$$

Step 3) Balance the rest of the elements by the hit and trial method. There are 3 hydrogen atoms on the left, balance them by adding 3/2 coefficient to the hydrogen on the right.

$$Al(s) + 3HBr(aq) \rightarrow AlBr_3(aq) + \frac{3}{2}H_2(\ g)$$

Step 4. Although the equation is balanced, it is recommended to remove the fraction by multiplying the coefficients in the whole equation with the highest common factor to obtain the final balanced equation.

$$2Al(s) + 6HBr(aq) \rightarrow 2AlBr_3(aq) + 3H_2(\ g)$$

Reactions of acids with carbonates and hydrogen carbonates

Carbonic acid (H_2CO_3) is a weak acid found in carbonated water. The H_2CO_3 is a product of carbon dioxide (CO_2) and water (H_2O) by the following equilibrium reaction.

$$CO_2\,(aq) + H_2O\,(l) \rightleftharpoons H_2CO_3\,(aq)$$

Hydrogen carbonate (HCO_3^-) and carbonate (CO_3^{2-}) are one and two hydrogen less, respectively than carbonic acid. The salts containing HCO_3^- and CO_3^{2-} accept one and two protons, respectively, from acids to make H_2CO_3. The H_2CO_3 decomposes to carbon dioxide and water by the reverse reaction shown above. Fig. 6.7.2 shows the reaction of sodium hydrogen carbonate with hydrochloric acid.

$$NaHCO_3(\ s) + HCl(aq) \rightarrow NaCl(aq) + H_2O(I) + CO_2(\ g) \uparrow$$



Figure 3.6.7.2: The reaction of sodium carbonate with HCl. Before and after HCl addition. Carbon dioxide bubbles out after HCl addition. Source: NCSSM, 05/18/20, https://youtu.be/TJYOxGHNTzg, CC BY 3.0

Similarly, calcium carbonate found in limestone reacts with acids. For example, sulfuric acid is one of the components in acid rain that reacts with calcium carbonate and damages sculptures made of stone.

$$CaCO_3(\ s) + H_2SO_4(aq) \rightarrow CaSO_4(aq) + H_2O(l) + CO_2(\ g) \uparrow$$

Reactions of acids with Arrhenius bases

Acids release proton (H^+) and Arrhenius bases release hydroxide ions (OH^-) in solution. When an acid mix with the Arrhenius base, H^+ and OH^- ions react with each other and produce water molecules.

$$\mathrm{H^+(aq) + OH^-(aq) \rightarrow H_2O(l)}$$

This is the net ionic equation of a reaction of an acid with an Arrhenius base. The cation of the base and the anion of the acid do not react –they are spectator ions. The reaction equation for an acid-base reaction is written in various ways explained below.

Molecular equation

The molecular equation shows the formulas of the substances. For example, the molecular equation of the reaction of HCl with NaOH is the following.

$$\mathrm{HCl(aq) + NaOH}(aq) \rightarrow \mathrm{NaCl}(aq) + \mathrm{H_2O(l)}$$

Complete ionic equation

The complete ionic equation shows the electrolytes dissociated into ions as they actually exist in the water phase, i.e., as aqueous ions. The complete ionic equation of the reaction of HCl with NaOH is the following.

$$\mathrm{H^+(aq) + Cl^-(aq) + Na^+(aq) + OH^-(aq) \rightarrow H_2O(l) + Na^+(aq) + Cl^-(aq)}$$

It is evident that Na^+ and Cl^- do not react –they are the spectator ions.

Net ionic equation

The spectator ions are present on both sides of the equation in equal numbers. The spectator ions can be canceled out, like the terms in an algebraic equation.

$$\mathrm{H^+(aq) + \cancel{Cl^-(aq)} + \cancel{Na^+(aq)} + OH^-(aq) \rightarrow H_2O(l) + \cancel{Na^+(aq)} + \cancel{Cl^-(aq)}}$$

The **net ionic equation** shows the substances that are not spectator ions after canceling out the spectator ions from the complete ionic equation. The net ionic equation of the reaction of HCl with NaOH is the following.

$$\mathrm{H^+(aq) + OH^-(aq) \rightarrow H_2O(l)}$$

Note

All of the reactions of the strong acids with strong Arrhenius bases have the same net ionic equation.

Writing a balanced chemical equation of the reaction of acids with Arrhenius bases

The following example shows the steps needed to write the balanced equation of the reaction of an acid with an Arrhenius base.

Example 3.6.7.2

Write a balanced chemical equation of a reaction between HCl and $Ca(OH)_2$?

Solution

Step 1) Write the formula of acid and base in the reactants and salt and water in the products. All the strong electrolytes dissolve in water, so use (aq) to represent their state.

$$\mathrm{HCl(aq) + Ca(OH)_2(aq) \rightarrow\ Salt(aq)\ + H_2O(l)}$$

Step 2) Balance the H^+ in the acid with the OH^- in the base.

$$\mathrm{2HCl(aq) + Ca(OH)_2(aq) \rightarrow Salt(aq) + H_2O(l)}$$

Step 3) Balance the H_2O with the H^+ in the acid, or with the OH^- in the base.

$$\mathrm{2HCl(aq) + Ca(OH)_2(aq) \rightarrow\ Salt\ (aq) + 2H_2O(I)}$$

Step 4) Write the salt by combing the cations from the base and the anions from the acid. Make sure the charges are balanced in the salt to make it a neutral substance.

$$\mathrm{2HCl(aq) + Ca(OH)_2(aq) \rightarrow CaCl_2(aq) + 2H_2O(I)} \tag{3.6.7.1}$$

Example 3.6.7.3

Write a balanced chemical equation of a reaction between H_2SO_4 and $Ba(OH)_2$?

Solution

Step 1) Write the formula of acid and base in the reactants and salt and water in the products. All the strong electrolytes dissolve in water, so use (aq) to represent their state.

$$\mathrm{H_2SO_4(aq) + Ba(OH)_2(aq) \rightarrow Salt(aq) + H_2O(l)}$$

Step 2) Balance the H^+ in the acid with the OH^- in the base. They are already balanced in the above equation.

Step 3) Balance the H_2O with the H^+ in the acid, or with the OH^- in the base.

$$\mathrm{H_2SO_4(aq) + Ba(OH)_2(aq) \rightarrow Salt(aq) + 2H_2O(l)}$$

Step 4) Write the salt by combing the cations from the base and the anions from the acid. Make sure the charges are balanced in the salt to make it a neutral substance.

$$\mathrm{H_2SO_4(aq) + Ba(OH)_2(aq) \rightarrow BaSO_4(aq) + 2H_2O(l)}$$

Antacids

The stomach sometimes produces excess HCl that may cause heartburn.

Antacids are the substances used to neutralize the excess HCl in the stomach.



Figure 3.6.7.3: Bottle of Antacid tablets. Source: Midnightcomm / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/).

Fig. 6.7.3 shows antacid tablets. The antacids include Arrhenius bases with very low solubility in water, like $Al(OH)_3$, and $Mg(OH)_2$, or weak bases like $CaCO_2$, and $NaHCO_3$. that react with and neutralize the acid:

$$\begin{aligned}
&Al(OH)_3(\ s) + 3HCl(aq) \rightarrow AlCl_3(aq) + 3H_2O(l)\\
&Mg(OH)_2(\ s) + 2HCl(aq) \rightarrow MgCl_2(aq) + 2H_2O(l)\\
&CaCO_3(\ s) + 2HCl(aq) \rightarrow CaCl_2(aq) + H_2O(I) + CO_2(\ g)\\
&NaHCO_3(\ s) + HCl(aq) \rightarrow NaCl(aq) + H_2O(I) + CO_2(g)
\end{aligned}$$

Side effects of the antacids

The antacids have side effects, e.g., aluminum and calcium-containing antacids may cause constipation. Magnesium-containing antacids have a laxative effect. Some formulations use a mixture of aluminum and magnesium-containing antacids to cancel out the side effect of each other. The calcium-containing antacid is not recommended for persons who tend to develop kidney stones, because the kidney stone is usually a salt of calcium. Sodium-containing antacids may increase the sodium level in the body fluid and may raise the blood pH.

Acid-base titration

The acid-base titration is an analytical process of determining the concentration of acid, called **analyte**, by neutralizing it with a base of know concentration, called the **standard**, or vice versa.

Usually, the base is in a burette and added drop by drop to a known volume of the acid in an Erlenmeyer flask, as illustrated in Fig. 6.7.4. A few drops of an acid-base indicator are mixed with the acid. The indicator changes color at a specific narrow pH range.

- The point when the stoichiometric amount of the base has been added to the acid is called the **equivalence point.**
- The point when the indicator changes color is the **end-point**.

Usually, the equivalence point is almost the same as the end-point.

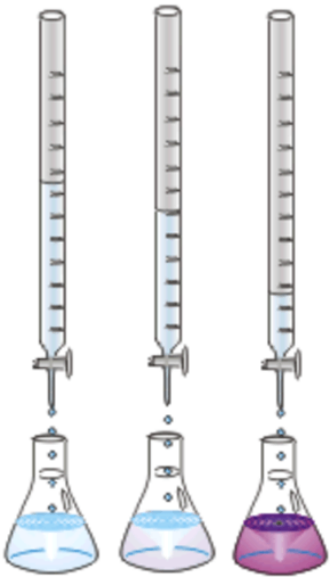


Figure 3.6.7.4: Acid-base titration set-up. The analyte color due to the indicator before, at the endpoint, and after being shown by the images on the left, middle, and right, respectively. *Source: Taken from Luigi Chiesa / Public domain.*

The calculation steps after the titration are usually the following.

Steps in acid-base titration calculations

1. The amount, in moles, of the standard is calculated by multiplying the amount in liters with the molarity of the standard.
2. Then, the amount, in moles, of the analyte, is calculated by using the mol-to-mol conversion factor from the balanced chemical equation of the reaction.
3. Finally, the molarity of the analyte is calculated by dividing moles by the volume in liters of the analyte.

Example 3.6.7.4

What is the molarity of the HCl solution if 50.0 mL of the HCl solution requires 42.0 mL of 0.123 M NaOH solution in the titration?

Solution

Step 1) Given:

$$[\text{NaOH}] = 0.123 \text{ M} = \frac{0.123 \text{ mol } NaOH}{1 \text{ NaOH}}$$

$$\text{Volume of standard} = 42.0 \text{ mL.} = 42.0 \text{ mL NaOH} \times \frac{1 \text{ L NaOH}}{1000 \text{ mL NaOH}} = 0.0420 \ L \text{ NaOH}$$

$$\text{Vol. analyte} = 50.0 \text{ mL} = 50.0 \text{ mL HCl} \times \frac{1 \text{ L HCl}}{1000 \text{ mL HCl}} = 0.0500 \text{ L HCl}$$

Step 2) Calculate the moles of the standard from the volume and the molarity product.

$$\text{Moles of NaOH consumed} = 0.0420 \text{ L NaOH} \times \frac{0.123 \text{ mol NaOH}}{1 \text{ L NaOH}} = 0.00517 \text{ mol NaOH}$$

Step 3) write the balanced chemical equation of the reaction.

$$\text{HCl(aq)} + \text{NaOH(aq)} \rightarrow \text{NaCl(aq)} + \text{H}_2\text{O(l)}$$

Step 4) write the conversion factor for mole of standard to mole of analyte calculation from the equation.

$$\frac{1\text{mol HCl}}{1\text{ mol NaOH}}$$

Step 5) Calculate the moles of the analyte by multiplying the moles of the standard and the conversion factor.

$$0.00517\text{ mol NaOH} \times \frac{1\text{ mol HCl}}{1\text{ mol NaOH}} = 0.00517\text{ mol HCl}$$

Step 6) Calculate the molarity of the analyte by dividing the moles with the volume in liters of the analyte.

$$0.00517\text{ mol NaOH} \times \frac{1\text{ mol HCl}}{1\text{ mol NaOH}} = 0.00517\text{ mol HCl}$$

After learning the steps thoroughly, the calculations can be done in a few steps, as shown in the following.

Calculations:

$$42.0\text{ mL NaOH} \times \frac{1\text{ L NaOH}}{1000\text{ mL NaOH}} \times \frac{0.123\text{ mol NaOH}}{1\text{ L NaOH}} \times \frac{1\text{ mol HCl}}{1\text{ mol NaOH}} = 0.00517\text{ mol HCl}$$

Molarity of HCl:

$$M = \frac{n(\text{mol})}{V(L)} = \frac{0.00517\text{ mol HCl}}{50.0\text{ mL HCl}} \times \frac{1000\text{ mL HCl}}{1\text{ L HCl}} = 0.103\text{ M HCl}$$

Example 3.6.7.5

What is the molarity of the H_2SO_4 solution if 50.0 mL of the H_2SO_4 solution requires 32.3 mL of 0.201 M NaOH solution in the titration?

Solution

Step 1) Given:

$$[\text{NaOH}] = 0.201\text{ M} = \frac{0.201\text{ mol NaOH}}{1\text{ L NaOH}}$$

$$\text{Vol. standard } = 42.0\text{ mL} = 32.3\text{ mL NaOH} \times \frac{1\text{ L NaOH}}{1000\text{ mL NaOH}} = 0.0323\text{ L NaOH}$$

$$\text{Vol. analyte } = 50.0\text{ mL} = 50.0\text{ mL HCl} \times \frac{1\text{ L HCl}}{1000\text{ mL HCl}} = 0.0500\text{ L HCl}$$

Step 2) Calculate the moles of the standard from the volume and the molarity product.

$$\text{Moles of NaOH consumed } = 0.0323\text{ N NaOH} \times \frac{0.201\text{ mol NaOH}}{1\text{ L NaOH}} = 0.00649\text{ mol } NaOH$$

Step 3) write the balanced chemical equation of the reaction.

$$H_2SO_4(aq) + 2NaOH(aq) \rightarrow Na_2SO_4(aq) + 2H_2O(l)$$

Step 4) write the conversion factor for the moles of standard to mole of analyte calculation from the equation.

$$\frac{1\text{ mol } H_2SO_4}{2\text{ mol NaOH}}$$

Step 5) Calculate the moles of the analyte by multiplying the moles of the standard and the conversion factor.

$$0.00649\text{ mol NaOH} \times \frac{1\text{ mol } H_2SO_4}{2\text{ mol NaOH}} = 0.00325\text{ mol } H_2SO_4$$

Step 6) Calculate the molarity of the analyte by dividing the moles with the volume in liters of the analyte.

$$\text{Concentration of } \ H_2SO_4 = \frac{0.00325 \text{ mol } H_2SO_4}{0.0500 \text{ L } H_2SO_4} = 0.0649 \text{ M } H_2SO_4$$

The same calculation in a summary form is in the following.

$$\text{Concentration of } H_2SO_4 = \frac{0.00325 \text{ mol } H_2SO_4}{0.0500 \text{ L } H_2SO_4} = 0.0649 \text{ M } H_2SO_4$$

$$\text{Moles of } H_2SO_4 = \quad 32.3 \text{ mL NaOH} \times \frac{1 \text{ L NaOH}}{1000 \text{ mL NaOH}} \times \frac{0.201 \text{ mol NaOH}}{1 \text{ L NaOH}} \times \frac{1 \text{ mol } H_2SO_4}{2 \text{ mol NaOH}} = 0.00326 \text{ mol } H_2SO_4$$

$$\text{Molarity of } H_2SO_4: \quad M = \frac{n(\text{ mol})}{V(L)} = \frac{0.00326 \text{ mol } H_2SO_4}{50.0 \text{ mL } H_2SO_4} \times \frac{1000 \text{ mL } H_2SO_4}{1 \text{ L } H_2SO_4} = 0.0649 \text{ M } H_2SO_4$$

This page titled 3.6.7: Acid-base reactions is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.6.8: pH Buffers

What is a pH buffer?

A pH buffer is an aqueous solution consisting of a weak acid and its conjugate base or vice versa that lets the pH change to be minimal when a small amount of a strong acid or a strong base adds to it.

For example, the addition of 0.020 mol HCl into 1 L of water changes pH from 7 to 1.7, i.e., about an 80% change in pH. The addition of 0.020 mol NaOH to the same water changes pH from 7 to 12.3, i.e., again, about an 80% change in pH. In contrast to pure water, 1 L of buffer solution containing 0.50 mol acetic acid (CH_3COOH) and 0.50 mol CH_3COO^- -the conjugate of the acetic acid, changes pH from 4.74 to 4.70 by the addition of the same 0.020 mol HCl and from 4.74 to 4.77 by the addition of 0.020 mol NaOH, i.e., about 1% change in pH, as illustrated in Fig. 6.8.1.

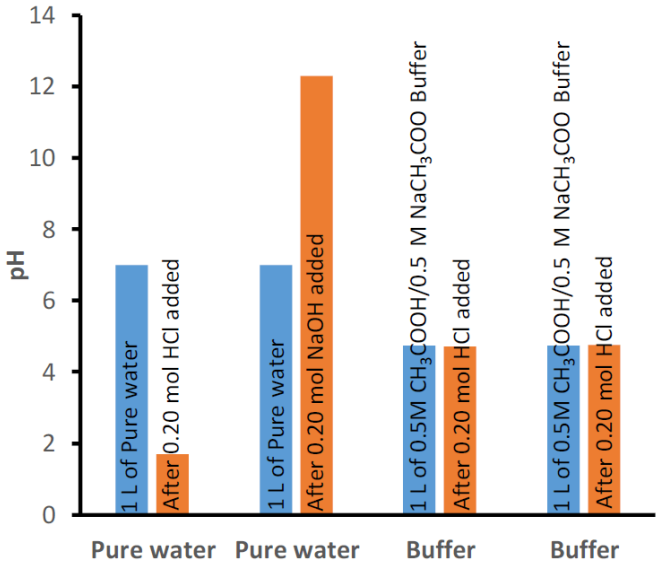


Figure 3.6.8.1: Effect acid and base addition on pH change of pure water at pH 7.00 and on acetic acid/sodium acetate buffer at pH 4.74.

Buffers of different initial pH values can be prepared, e.g., by varying the ratio of the weak acid to its conjugate base or by using a different set of a weak acid and its conjugate base. One example is a buffer of initial pH 4.74 comprising 0.5 M acetic acid and 0.5 M sodium acetate shown in Fig. 6.8.1. Another example is a buffer comprising 0.1M dihydrogen phosphate and 0.1M hydrogen phosphate that has an initial pH of 7.21, as shown in Fig. 6.8.2.

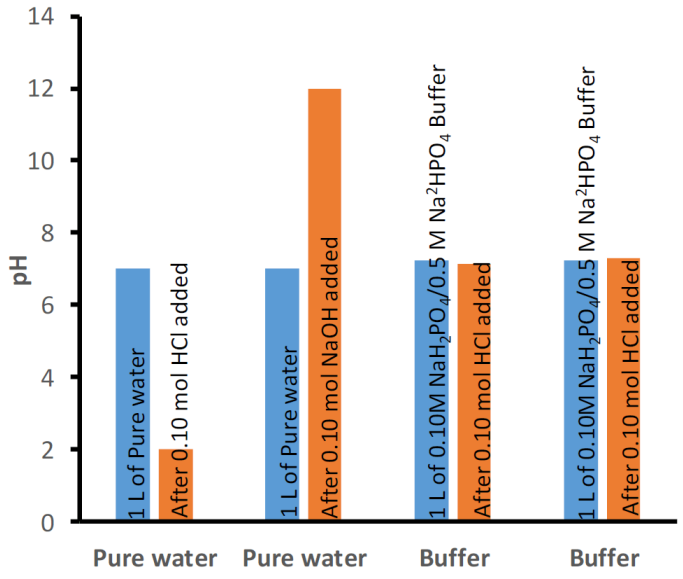


Figure 3.6.8.2: Effect of acid and base addition on pH change of pure water at pH 7.00 and on Na_2HPO_4/NaH_2PO_4 buffer of pH 7.21.

Buffer capacity

Buffer capacity refers to how much amount of a strong acid or a strong base the buffer can handle before a drastic change in pH happens.

Higher the amount of weak acid/conjugate base higher the buffer capacity. The buffer has an equal amount of weak acid and its conjugate base has a higher buffer capacity than the same buffer that has an unequal ratio of the acid and its conjugate base.

Caution

Strong acid and its conjugate base or a strong base and its conjugate acid do not make a buffer solution.

Mechanism of buffer action

The buffer contains a weak acid and its conjugate base in equilibrium. For example, acetic acid/sodium acetate buffer has the following equilibrium.

$$CH_3COOH \rightleftarrows H^+ + CH_3COO^-$$

The molar concentration of hydrogen ions [H^+] defines the pH of the solution. The conjugate base consumes any strong acid added.

$$HA + CH_3COO^- \rightarrow CH_3COOH + A^-$$

, where HA is any strong acid and A^- is its conjugate base. The concentration of CH_3COOH increases and CH_3COO^- decrease, but pH decreases small because [H^+] increases only a little. Similarly, the weak acid consumes any strong base added.

$$MOH + CH_3COOH \rightarrow CH_3COO^- + M^+ + H_2O$$

Where MOH is any strong base, and M^+ is its conjugate acid. Fig. 6.8.3 illustrates the mechanism of buffer action.

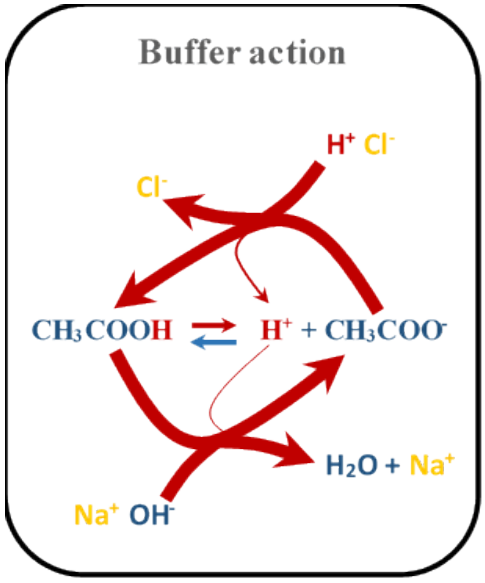


Figure 3.6.8.3: Mechanism of blood buffer action: The main blood buffer is H_2CO_3/HCO_3^-. Added acid is consumed by reacting with the HCO_3^- and added base is consumed by consumed by reacting with H_2CO_3, maintaining the pH ~7.4. Longs regulate the pH by regulating the rate of CO_2 exhale and kidney regulates the pH by regulating the HCO_3^- in the buffer system.

Note

Strong acid and its conjugate base mixture have no buffer action because the acid dissociates almost 100%, leaving no acid behind for the buffer action. The conjugate base of a strong acid is a very weak base that does not react with the added acids. The same explanation applies to a strong base and its conjugate acid mixture having no buffer action.

Buffers in the blood

The blood maintains its pH of ~7.4 primarily by the carbonic acid/hydrogen carbonate buffer system. The blood pH in the range of 7.45 to 7.35 is considered healthy, but outside of this range causes medical problems. If the blood pH decreases to 6.8 or increases to 8.0, death may occur. It is critical to maintain the blood pH in a narrow range for the cells to function correctly. Specifically, the enzymes and other proteins have secondary, tertiary, and quaternary structures needed for their proper functions. Hydrogen bonding plays a crucial role in defining the protein's structure. pH changes alter the hydrogen bonding making the proteins less effective or ineffective in their functions.

Buffer systems regulate the blood pH. The primary blood buffer system is carbonic acid/hydrogen carbonate, as illustrated in Fig. 6.8.4. Metabolic processes in the cells produce carbon dioxide (CO_2) that enters the bloodstream and produces carbonic acid (H_2CO_3) by reacting with water. Kidneys supply the hydrogen carbonate (HCO_3^-) –the conjugate base of the carbonic acid and also keep a reservoir of the HCO_3^-. The H_2CO_3 consumes any base added and the HCO_3^- consumes any acid added, thus minimizing the pH change due to the added acids and bases.

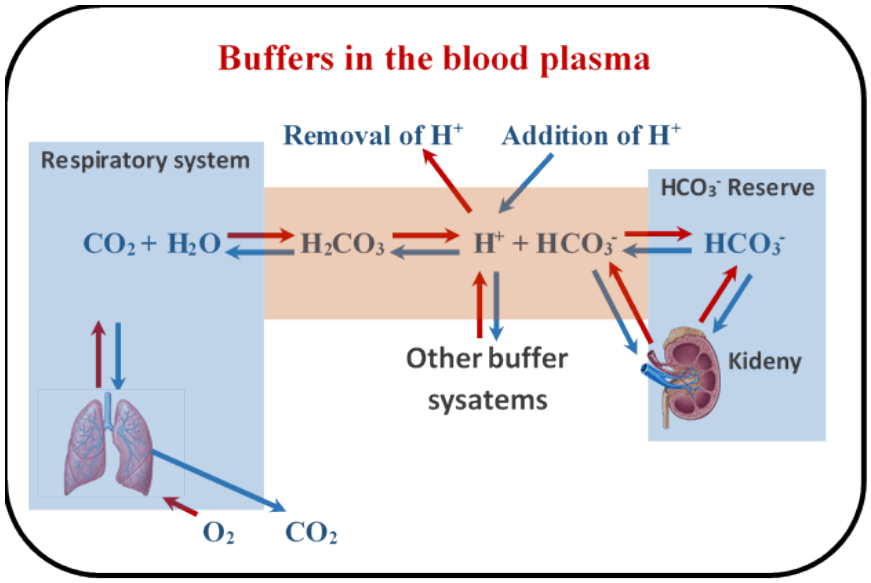


Figure 3.6.8.4: Mechanism of blood buffer action: The main blood buffer is H_2CO_3/HCO_3^-. Added acid is consumed by reacting with the HCO_3^- and the added base is consumed by reacting with H_2CO_3, maintaining the pH ~7.4. Longs regulate the pH by regulating the rate of CO_2 exhale and the kidney regulates the pH by regulating the HCO_3^- in the buffer system.

The primary mechanism for pH regulation by the H_2CO_3/HCO_3^- buffer in the blood is through the lungs. When the blood pH is acidic compared to the average, the breathing rate increases exhaling more CO_2 that decreases the concentration of H^+, following the blue arrows in Fig. 6.8.4, increasing the pH. The decrease in the breathing rate has the opposite effect. Kidneys also regulate the blood pH by adding or removing HCO_3^-, but the kidney's response is delayed compared to the response of the lungs.

This page titled 3.6.8: pH Buffers is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 3.7: GASES

3.7.1: Characteristics of gases
3.7.2: The pressure-volume relationship
3.7.3: The temperature-volume relationship
3.7.4: The pressure-temperature relationship
3.7.5: The combined gas law
3.7.6: The volume-amount relationship
3.7.7: Ideal gas law
3.7.8: Dalton's law of partial pressure

This page titled Chapter 3.7: Gases is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.7.1: Characteristics of gases

What is a gas

Gas is one of the four states of matter that falls between the liquid and the plasma state. The air around us is a gas composed of ~78% nitrogen (N_2), ~21% oxygen, and the remaining ~1% are other gases, including carbon dioxide, water vapors, argon, etc., as illustrated in Fig. 7.1.1.

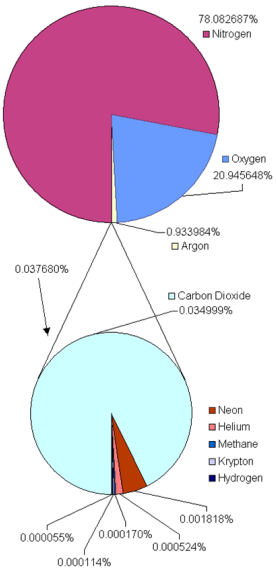


Figure 3.7.1.1: Depictions of the composition of the Earth's atmosphere; the lower pie represents the least common gasses that compose 0.037680% of the atmosphere. Source: Brockert at English Wikipedia / Public domain

Elements that exist as gases at room temperature and atmospheric pressure include noble gases which are monoatomic molecules including helium (He), neon (Ne), argon (Ar), krypton (Kr), xenon (Xe), and radon (Rn), and diatomic elemental molecules that include hydrogen (H_2), nitrogen (N_2), oxygen (O_2), and two halogens: fluorine (F_2) and chlorine (Cl_2). Several molecular compounds exist as gases at room temperature, e.g., carbon dioxide (CO_2), carbon monoxide (CO), methane (CH_4), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), hydrogen sulfide (H_2S), ammonia (NH_3).

The properties of gases are defined by their amount in moles (n), volume (V), temperature (T), and pressure (P). These parameters are described in the following.

The amount of a gas

The amount of gas is measured in moles (n), i.e., 6.022 x 10^{23} molecules of a gas is one mole of the gas. The gas has mass and can be measured in mass units like grams or kilograms, but the mass needs to be divided by the molar mass of the gas to obtain the amount in moles for calculations.

Note

The amount of gas in moles is needed in gas laws because the properties of gases are proportional to the number of molecules, not to the masses of the molecules.

Example 3.7.1.1

How many moles are in 10.0 g of oxygen gas?

Solution

Step 1. Given: Mass = 10.0 g O_2, Desired: ? moles of O_2.

Step 2. Equality between the given and the desired units: 1 mole of O_2 = 32.00 g O_2.

Step 3. The conversion factors from the equality: $\frac{1\ mol\ O_2}{32.00\ g\ O_2}$ and $\frac{32.00\ g\ O_2}{1\ mol\ O_2}$

Step 4. Multiply the given quantity with the conversion factor that cancels the given and leaves the desired unit in the answer:

$$10.0\ \cancel{g\ O_2)} \times \frac{1\ mol\ O_2}{32.00\ \cancel{g\ O_2}} = 0.312\ g\ O_2$$

The volume of a gas

Volume (*V*) is the space a substance occupies. The SI unit of volume is a meter cube (m^3), i.e., the space inside of a 1m x 1m x 1m cube.

Usually, the volume is reported in liters (L), which is dm^3, or milliliter (mL), which is cm^3. The relationship between the volume units is the following.

$$1\ \mathrm{m}^3 = 1000\ \mathrm{L}$$

$$1\ \mathrm{L} = 1000\ \mathrm{mL}$$

The gases do not have a fixed shape or volume. The gases acquire the shape of the container. The gases expand or contract to fill the available space in a container.

The temperature of a gas

The temperature (*T*) is a measure of how hot or cold an object is.

The temperature is a manifestation of the thermal energy of a substance. Thermal energy is a source of heat. Heat is the flow of energy from a hot to a cold object. The SI unit of temperature is Kelvin (K) which has a value of 273.15 K at the freezing point of water and 375.15 K at the boiling point of water. The zero of the kelvin scale is called the absolute zero at which no more energy can flow out of a substance as heat. There is no negative number on the kelvin scale of temperature.

Other commonly used units of temperature are Celsius (^oC) which has a value of 0 ^oC at the freezing point of water and 100 ^oC at the boiling point of water. The relationship between temperature in kelvin (T_K) and the temperature in Celsius (T_C) is the following.

$$T_K = T_C + 273.15$$

Fahrenheit is a temperature scale used in the English system of measurement that has a value of 32 ^oF at the freezing point of water and 212 ^oF at the boiling point of water. The relationship between the temperature in Fahrenheit (T_F) and the temperature in Celsius (T_C) is the following.

$$T_F = \frac{9}{5}T_C + 32$$

Note

The temperature in Kelvin (K) is used in the gas law calculations.

The pressure of a gas

Pressure (*P*) is the force (*F*) per unit area (*A*), i.e., $P = \frac{F}{A}$.

The matter has mass and applies a force, which is the weight due to gravitational pull. For example, a column 1 m x 1m at sea level and height equal to the earth's atmosphere, as illustrated in Fig. 7.1.2 has a mass of ~10,000 kg. Mercury (Hg) filled in a column of 1 m x 1m and height 0.760 m has the same mass as the mass of air of 1 m x 1 m column extending from the sea level to the entire atmosphere of the earth.



Figure 3.7.1.2: Illustration of an air column of 1m x 1m base and height covering the entire atmosphere of the earth. Modified from: http://wikimapia.org/8706099/Norfolk.../photo/1606686

A column of air extending from sea level to the entire atmosphere applied pressure equal to one atmosphere (atm), where atm is the pressure unit. One atmosphere (atm) is equal to 760 millimeter of mercury (mmHg), where mmHg is another unit of pressure. One mmHg is also called Torr. The relationship between atm, mmHg, and Torr is the following.

$$1 \text{ atm} = 760 \text{ mmHg} = 760 \text{ Torr}$$

Pascal (Pa)

The SI unit of pressure is pascal (Pa), which is pressure (P) applied by a force (F) of one newton (N) on an area of one meter square (m^2). Mathematical form of the unit is: $1\ Pascal = \frac{1\ Newton}{1\ meter-square}$ or $Pa = \frac{N}{m^2}$

Kilopascal is 1000 Pa, and kilopascal is also called the bar. The English system unit of pressure is pound-force per square inch (psi). The relationship between different units of pressure mentioned above is the following.

$$1 \text{ atm} = 760 \text{ mmHg} = 760 \text{ Torr} = 1.01325 \times 10^5 \text{ Pa} = 1.01325 \times 10^2 \text{ kP} = 1.01325 \text{ bar} = 14.7 \text{ psi}$$

These relationships are changed to conversion factors that allow conversion from one to another pressure unit, as explained in the following examples.

Example 3.7.1.2

Express 51 mm Hg in a) atm and b) in Pa.

Solution

Given: Pressure = 51 mm Hg. Required: pressure in atm and Pa

Relationship between mm Hg and atm: $1\ \text{atm} = 760\ \text{mmHg}$

The conversion factors from the equality: $\frac{1\ \text{atm}}{760\ \text{mmHg}}$ and $\frac{760\ \text{mmHg}}{1\ \text{atm}}$

Calculations: multiply the given quantity with the conversion factor that cancels the given unit and leaves the desired unit in the answer.

$$51\ \cancel{\text{mmHg}} \times \frac{1\ \text{atm}}{760\ \cancel{\text{mmHg}}} = 6.7 \times 10^{-2}\ \text{atm}$$

Similarly, the relationship between mm Hg and Pa: $760\ \text{mmHg} = 1.01325 \times 10^5\ \text{Pa}$

The conversion factors from the equality: $\frac{760\ \text{mmHg}}{1.01325 \times 10^5\ \text{Pa}}$ and $\frac{1.01325 \times 10^5\ \text{Pa}}{760\ \text{mmHg}}$

Calculations: multiply the given quantity with the conversion factor that cancels the given unit and leaves the desired unit in the answer.

$$51\ \cancel{\text{mmHg}} \times \frac{1.01325 \times 10^5\ \text{Pa}}{760\ \cancel{\text{mmHg}}} = 6.8 \times 10^3\ \text{Pa}$$

Measurement of gas pressure

> The instrument used to measure the atmospheric pressure is called a barometer.

The barometer, shown in Fig. 7.1.3, was invented by Evangelista Torricelli. It is a glass tube that is more than 760 mm long and closed at one end. The glass tube is filled with mercury and inverted to dip its open end in a dish that contains a layer of mercury. The mercury falls from the tube due to its weight but stops falling when the column of mercury is 760 mm in height at sea level due to the atmospheric pressure pushing it back into the tube. Remember 760 mmHg = 1 atm. Above 760 mm is a vacuum in the tube.

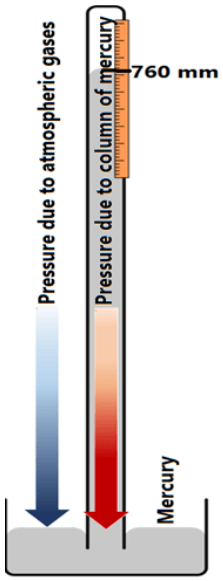


Figure 3.7.1.3: A mercury barometer. Modified from: Danomagnum / Public domain

The pressure of gases in a laboratory is usually measured using an instrument called a manometer.

A manometer is a U-shaped glass tube partially filled with mercury, as shown in Fig. 7.1.4. One end of the U-tube is connected to a gas chamber in which pressure is being measured, and the other end is either closed (with a vacuum on the closed end), called a closed-ended manometer, or it is open to the atmosphere, called an open-ended manometer.

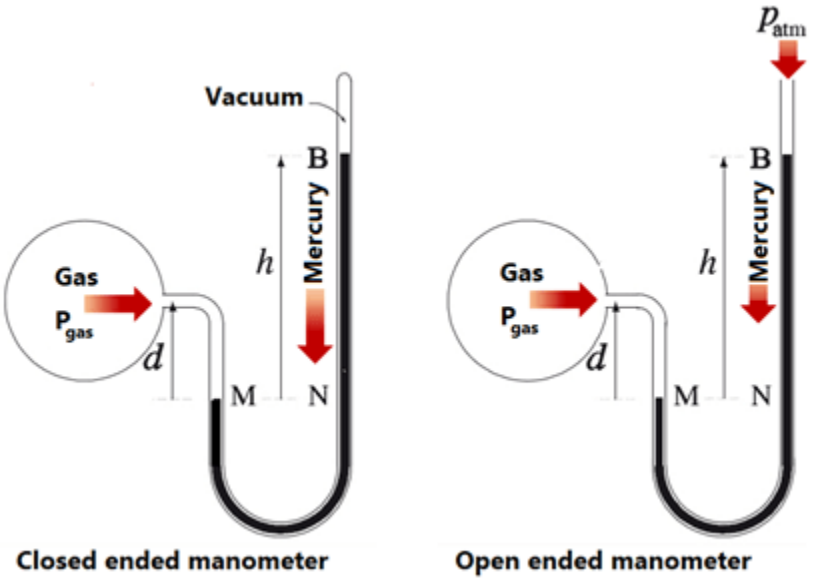


Figure 3.7.1.4: Manometers. Source: Modified from Algarabia / Public domain

In the case of a closed-ended manometer, the gas pressure is equal to the difference in the height of mercury in the two arms of the U-tube in mmHg, i.e., height (*h*) from point N to point B in mm units as shown in Fig. 7.1.4.

Example 3.7.1.3

Calculate the pressure in atm for the gas shown in Fig. 7.1.4 in the closed-ended manometer if *h* = 40 mm?

Solution

In a closed-ended manometer, the gas pressure is equal to the height *h* in mm Hg: P = 40 mm Hg.

Pressure in atm:

$$40 \ \cancel{\text{mmHg}} \times \frac{1 \text{ atm}}{760 \ \cancel{\text{mmHg}}} = 5.3 \times 10^{-2} \text{ atm}$$

In the case of an open-ended manometer, the gas pressure is equal to the difference in the height of mercury in the two arms of the U-tube in mm of Hg plus the atmospheric pressure outside expressed in mm of Hg.

Variation of air pressure with altitude

The air pressure decreases, and the gas concentration decreases as the altitude increases, as illustrated in Fig. 7.1.5. The oxygen concentration in the air also decreases as the altitude increases. That is why mountain climbers experience an inadequate supply of oxygen to the body at high altitudes -a medical condition called hypoxia.

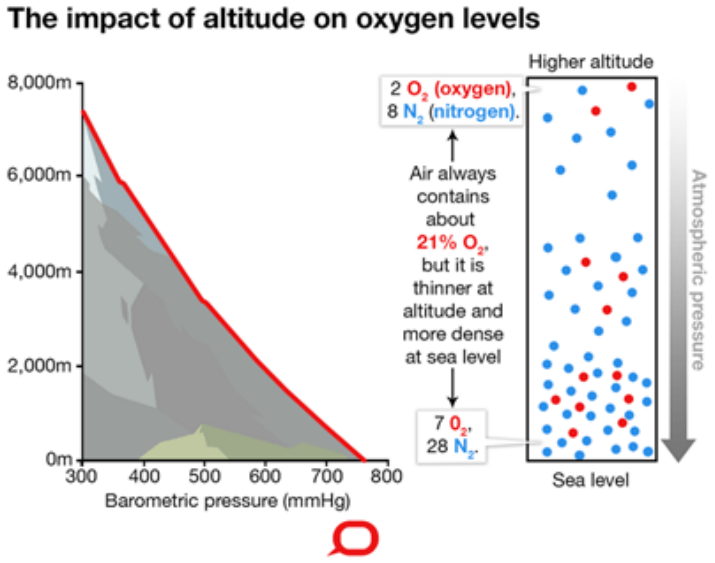


Figure 3.7.1.5: Air pressure decreases and the air concentration and oxygen concentration in it decreases as the altitude increases. Source: Brendan Scott/The Conversation, CC BY-ND https://images.theconversation.com/f....png?ixlib=rb-

Blood pressure

The heart pumps the blood into the circulatory system. When the heart contracts, it applies the pressure on the blood in it, and the blood pumps out of the heart into the circulatory system, as illustrated in Fig. 7.1.6. The blood pressure in the circulatory system is highest at this point, and it is called **systolic** pressure. It can be in the range of 100 to 200 mm Hg. A desirable systolic pressure range is 100 to 120 mm Hg. When heart muscles relax, the cavity in the heart expands and more blood fills in the heart. The blood pressure in the circulatory system is minimum at this point, and it is called **diastolic** pressure. The diastolic pressure may vary in the range of 60 to 110 mm Hg. A desirable diastolic pressure is less than 80 mm Hg.

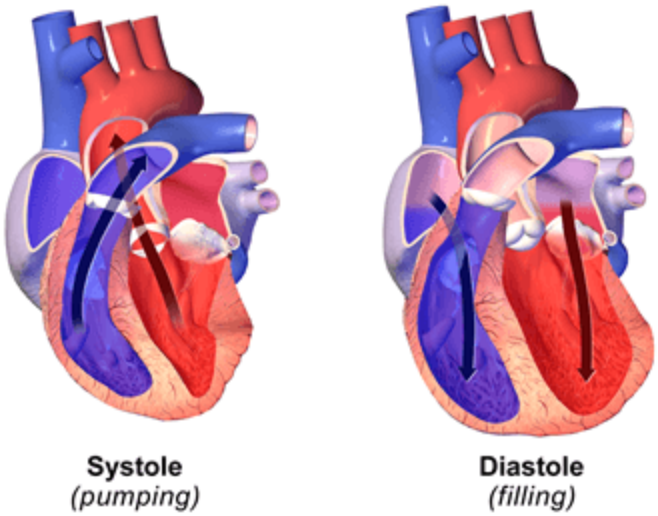


Figure 3.7.1.6: Diagram of heart illustrating systole, i.e., blood pumping out of heart vs Diastole i.e., blood filling into the heart. Source: BruceBlaus / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

The device used to measure blood pressure is called a **sphygmometer**. It consists of a cuff that wraps around the upper arm, a pump to inflate air in the cuff, and a stethoscope to hear the sound of blood flow, as illustrated in Fig. 7.1.7. The cuff is inflated with air to a pressure above the systolic pressure and it results in cutting off the blood flow through the brachial artery in the upper arm. A stethoscope is used to hear the sound of blood flow. No sound is heard at this point. The pressure in the cuff is slowly reduced. When the pressure in the cuff is equal to systolic pressure, the blood begins to spurt into the artery, and a tapping sound is heard through the stethoscope. At the point when the pressure is equal to the diastolic pressure, the blood flows freely through the artery, and the tapping sound disappears. The blood pressure reading is reported as a set of two numbers, e.g., 120/80, where the higher number is systolic, and the lower number is the diastolic pressure.

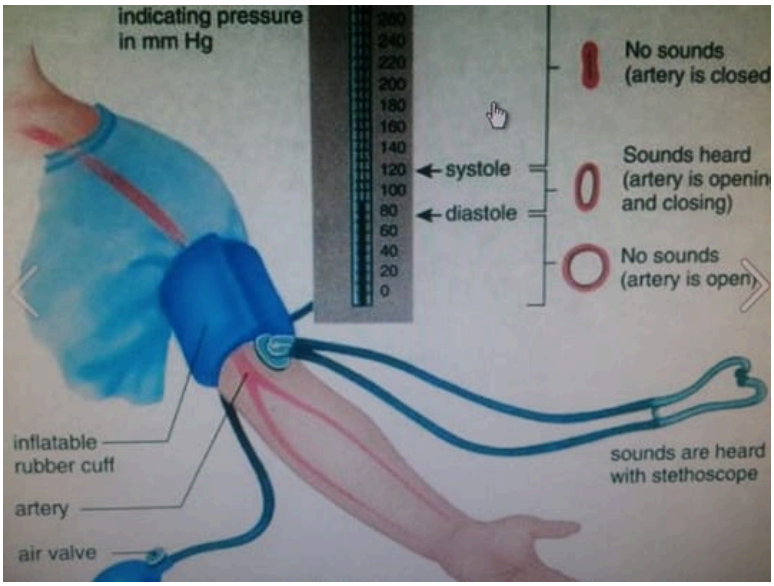


Figure 3.7.1.7: A healthcare professional performing blood pressure monitoring on a patient. Source: Roa'a najim / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Kinetic molecular theory of gases

The physical characteristics of the gases are explainable by a model called the **kinetic molecular theory of gases**, illustrated in Fig. 7.1.8.

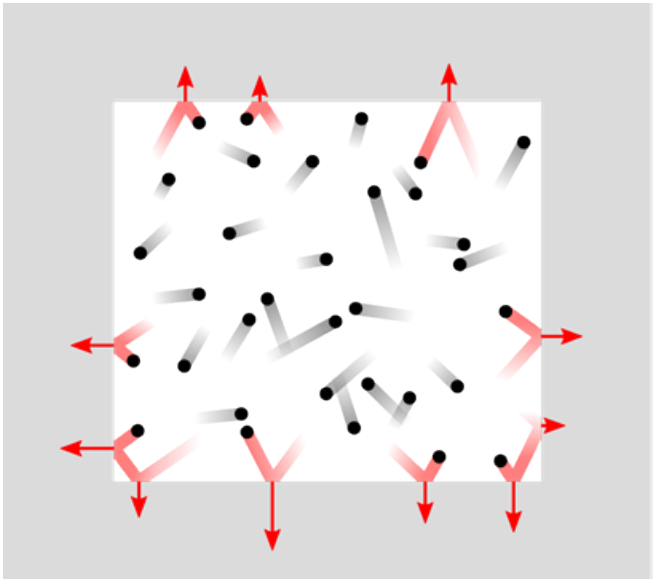


Figure 3.7.1.8: A model of a gas that is based on kinetic molecular theory of gases. The molecules are the spheres and the tails attached to the molecules represent the trajectories of molecular motion. The red tails represent molecules bouncing back from the wall resulting in the pressure exerted by the gas on the wall of the container. Source: Becarlson / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Postulates of kinetic molecular theory of gases

1. Gases consist of particles called molecules that are far apart from each other. As a consequence of the vast distances, the intermolecular forces are negligible, and the volume occupied by the gas molecules is negligible.
2. The gas molecules are moving in straight lines in random directions until they collide with another molecule or the wall of the container.
3. The collisions are elastic, i.e., the total energy is conserved, but the molecules change direction and change speed after the collision. It results in the distribution of molecular speeds over a wide range.
4. The average kinetic energy of the gas molecules is directly proportional to the absolute temperature. The average kinetic energy of the molecules of different gases is the same at the same temperature.

Explanation of the characteristics of the gases based on the kinetic molecular theory of gases

1. Upon exposing gas to a vacuum, the gas molecules keep moving in a straight path in the vacuum until they collide with a container boundary and bounce back. It explains gases are expandable. Collisions of gas molecules with the wall exert outward pressure on the wall and, as a reaction, the wall exerts an equal inward pressure on the gas. Applying more pressure to the gas through any portion of the gas boundary disturbs the above-mentioned balance of action and reaction forces, resulting in an inward movement of the portion of the boundary surface. In other words, the gases are compressible.
2. If a boundary separating two gases is removed, the random molecular motions ultimately mix the two gases making a homogeneous mixture, i.e., gases are entirely miscible with other gases.
3. When temperature increases the gas molecules move faster and colloid more frequently, which explains the fact that gases exert more pressure upon heating and expand if the boundary is not rigid. The more frequent bursting of tires in summer is due to the pressure increase due to heating.
4. When more gas fills in the tire, the pressure increases due to more frequent collisions. Compressing the gas into smaller volume have the same effect, i.e., the molecular collisions increase increasing the pressure.

The physical characteristics of the gases are independent of the nature of the gas molecules due to negligible intermolecular interactions but depend on four physical properties, i.e., pressure, volume, temperature, and amount of gas. The following sections describe the relationships between the physical properties of the gases.

This page titled 3.7.1: Characteristics of gases is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.7.2: The pressure-volume relationship

Pressure-volume relationship

Consider a gas in a cylinder with a piston pushing it down due to massive objects placed on it and due to outside air pressure, as illustrated in Fig. 7.2.1, The gas molecules strike the piston surface, applying pressure upward equal to the downward pressure applied by the piston that keeps the piston stationary. Increasing the pressure on the piston, e.g., by adding more weight to it, causes the piston to move down, reducing the gas volume. The gas molecules have less distance to travel before striking the piston surface which increases the collision frequency and causes the gas pressure to increase until becomes equal to the outside pressure.

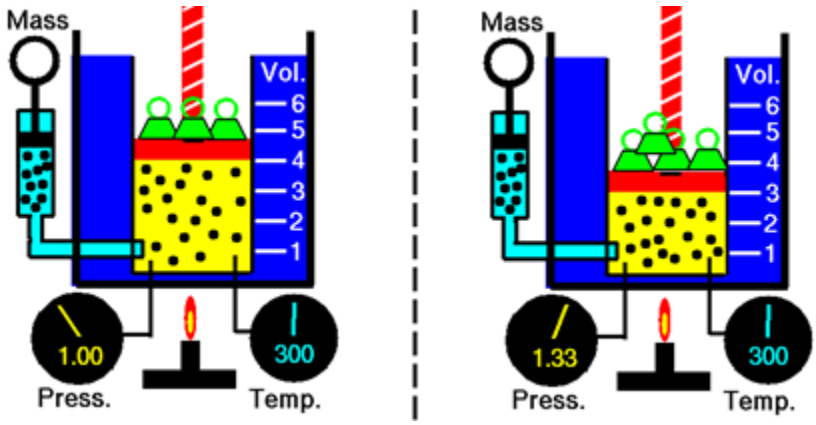


Figure 3.7.2.1: Illustration of Boyle's law: for a given amount of gas at a constant temperature, the volume is inversely proportional to the pressure. Source: NASA's Glenn Research Center / Public domain.

When a change in one parameter causes a change in another, the parameters are related. When an increase in one parameter causes a decrease in another, the two are inversely proportional to each other. Robert Boyle studied the quantitative relationship between the volume and pressure of the gas, keeping the quantity of gas and the temperature constant. The research concluded in a law called Boyle's law,which states that:

Boyle's law

The volume of a gas is inversely proportional to the pressure of the gas provided the temperature and the amount of the gas remain constant.

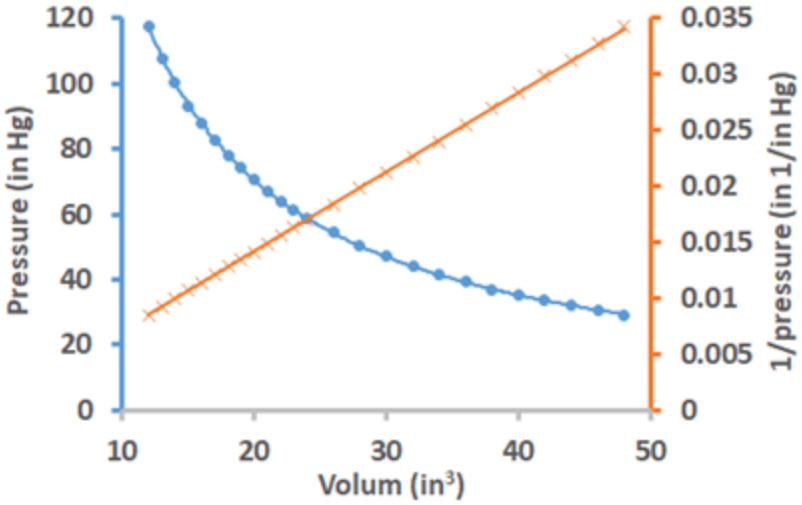


Figure 3.7.2.2: Boyles law: the volume of the gas is inversely proportional to the pressure of the gas. Source: Robert Boyle, *A Defence of the Doctrine Touching the Spring And Weight of the Air* (1662)

Fig. 7.2.2 illustrates the results of Boyle's experiment. The mathematical form of Boyle's law is:

$$V \propto \frac{1}{P}$$

, or

$$V = \frac{\mathrm{k}}{P}$$

, or

$$PV = \mathrm{k}$$

, where k is a consant. The graph of volume versus pressure is curvilinear, but the graph of volume versus the reciprocal of pressure is a linear graph showing the inverse proportionality between the volume and the pressure. Since the product PV is a constant, it implies that:

$$P_1V_1 = P_2V_2 = k$$

, i.e., a product of initial pressure (P_1) and initial volume (V_1) is equal to the product of final pressure (P_2) and final volume (V_2) of gas provided the quantity of the gas and temperature does not change.

Example 3.7.2.1

The pressure of a 1.32 L sample of SO_2 gas at 0.532 atm is increased to 1.231 atm. Calculate the new volume of the gas if the temperature and the quantity of the gas remain the same?

Solution

Given: P_1 = 0.532 atm, P_2 = 1.231 atm, V_1 = 1.32 L V_2 = ?

Formula: $P_1V_1 = P_2V_2 = k$, rearrange to isolate the desired variable: $V_2 = \frac{P_1V_1}{P_2}$.

Plug in the values in the rearranged formula and calculate: $V_2 = \frac{0.532 \text{ atm} \times 1.32 \text{ L}}{1.231 \text{ atm}} = 0.570 \text{ L}$

Example 3.7.2.2

An oxygen tank holds 20.0 L of oxygen at a pressure of 10.0 atm. What is the final pressure when the gas is released and occupies a volume of 200 L?

Solution

Given: V_1 = 20.0 L, V_2 = 200 L, P_1 = 10.0 atm P_2 = ?

Formula: $P_1V_1 = P_2V_2 = k$, rearrange to isolate the desired variable: $P_2 = \frac{P_1V_1}{V_2}$.

Plug in the values in the rearranged formula and calculate: $P_2 = \frac{10.0\ \text{atm} \times 20.0\ \text{L}}{200\ \text{L}} = 1.00\ \text{atm}$.

Breathing process

Boyle's law explains the mechanism of the breathing process. Lungs are elastic structures like balloons placed in the thoracic cavity, as illustrated in Fig. 7.2.3. The diaphragm muscle makes a flexible floor and ribs surround the cavity.

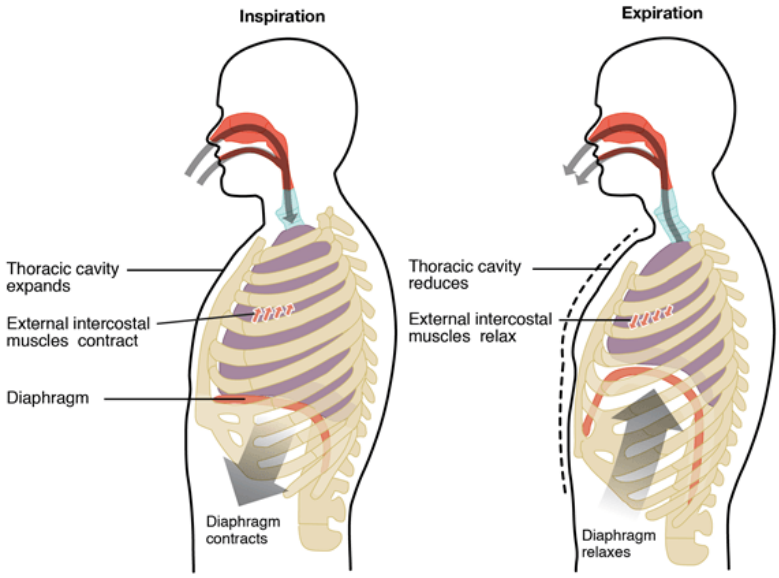


Figure 3.7.2.3: Illustration of breathing process –inspiration and expiration occur due to the expansion and contraction of the thoracic cavity, respectively. Source: OpenStax College / CC BY (https://creativecommons.org/licenses/by/3.0)

Inhalation

The inhalation or the inspiration process starts when the diaphragm contract and move down and the rib muscles contract, expanding the thoracic cavity. Volume increases, the air pressure decreases inside the inside thoracic cavity and the atmospheric air flows into the lungs until the pressure in the lungs is equal to the outside pressure.

Exhalation

The exhalation or the expiration process starts when the diaphragm expands and moves upwards, and the rib muscles relax, contracting the thoracic cavity. Volume decreases and the air pressure increases inside the thoracic cavity that pumps the air out of the lungs into the atmosphere.

This page titled 3.7.2: The pressure-volume relationship is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.7.3: The temperature-volume relationship

Consider a gas in a cylinder with a piston in Fig. 7.3.1. Increasing temperature increases the average kinetic energy (*KE*) of the gas molecules. The kinetic energy (*KE*) is directly proportional to the velocity of the molecules, i.e., $KE = \frac{1}{2}mv^2$, where *m* is the mass and *v* is the velocity. So, increasing temperature increases the velocity resulting in more frequent and more forceful collisions resulting in increased gas pressure inside the chamber. The gas volume starts to increase causing the pressure to decrease until the pressure inside the chamber is equal to the pressure outside. In other words, increasing temperature increases the volume of the gas if the pressure and amount of gas are not changed.

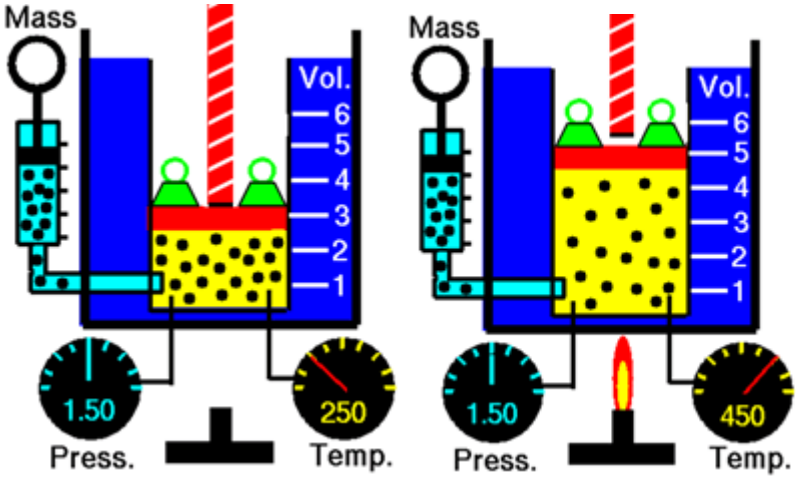


Figure 3.7.3.1: Increasing temperature increases the volume of the gas i.e., $\frac{V_1}{T_2} = \frac{V_2}{T_2} = \mathrm{k}$. Source: NASA's Glenn Research Center / Public domain.

If two related parameters increase or decrease together, they are directly proportional to each other.

Charles's law

Charles's law states that the volume of a given amount of gas is directly proportional to the temperature in the Kelvin scale at constant pressure.

Fig. 7.3.2 demonstrates that the volume of a gas decreases when the gas is cooled down.



Figure 3.7.3.1: Air balloon (left) shrinks when its temperature is decreased by submerging in liquid nitrogen (middle) and re-expands when returned to room temperature condition (right). Source: Ryan Poling aka expictura on Flickr.https://en.Wikipedia.org/wiki/File:Nitrogen.ogv#file, CC BY 2.0

The mathematical forms of Charles's law are the following.

$$V \propto T$$

, or

$$V = \mathrm{k}T$$

, or

$$\frac{V}{T} = \mathrm{k}$$

, where k is a constant, *V* is volume, and *T* is the temperature (in kelvin scale) of the gas. Since $\frac{V}{T}$ is a constant, it implies that

$$\frac{V_1}{T_1} = \frac{V_2}{T_2} = \mathrm{k}$$

where V_1 is the initial volume, T_1 is the initial temperature in Kelvin, V_2 is the final volume, and T_2 is the final temperature in Kelvin, provided the amount of gas and pressure do not change. Note that the kelvin scale is used in Charles's law because the kelvin scale does not have negative numbers which means the linear curve starts from the origin without any y-intercept. If the given temperature is not in the kelvin scale, first convert the temperature to the Kelvin scale and then use the gas laws for the calculations.

Example 3.7.3.1

A sample of CO_2 occupies 3.23 L volume at 25.0 °C. Calculate the volume of the gas at 50.0 °C if pressure and amount of gas do not change?

Solution

Given: T_1 = 25.0 °C + 273 = 298 K, T_2 = 50.0°C + 273 = 323 K, V_1 = 3.23 L, V_2 = ?

Formula: $\frac{V_1}{T_1} = \frac{V_2}{T_2}$, rearrange the formula to isolate the desired variable: $V_2 = \frac{V_1 T_2}{T_1}$

Plug in the values in the rearranged formula and calculate: $V_2 = \frac{3.23\ \mathrm{L} \times 323\ \mathrm{K}}{298\ \mathrm{K}} = 3.50\ \mathrm{L}$

Charles's law explains the drifting of warm air upward in the atmosphere. As the gas is wormed, its volume increases and its density decreases which makes the gas drift upward. A hot air balloon, shown in Fig. 7.3.3 operates using hot air.



Figure 3.7.3.3: A hot air balloon seen from a view directly below. The burner, or flame, is firing into the envelope above. The warm air is less dense than the atmospheric air which makes the balloon rise in the air. Source: Arpingstone at English Wikipedia / Public domain.

This page titled 3.7.3: The temperature-volume relationship is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.7.4: The pressure-temperature relationship

Consider a gas in a cylinder with a piston in Figure 3.7.4.1. Increasing temperature increases the average kinetic energy (KE) and the average velocity of the gas molecules resulting in more frequent and more forceful collisions which result in increased gas pressure applied on the piston or the walls of the gas container.

Gay-Lussac's law

Gay-Lussac's law states that the pressure of a gas is directly proportional to the absolute temperature provided the volume and amount of gas are not changed.

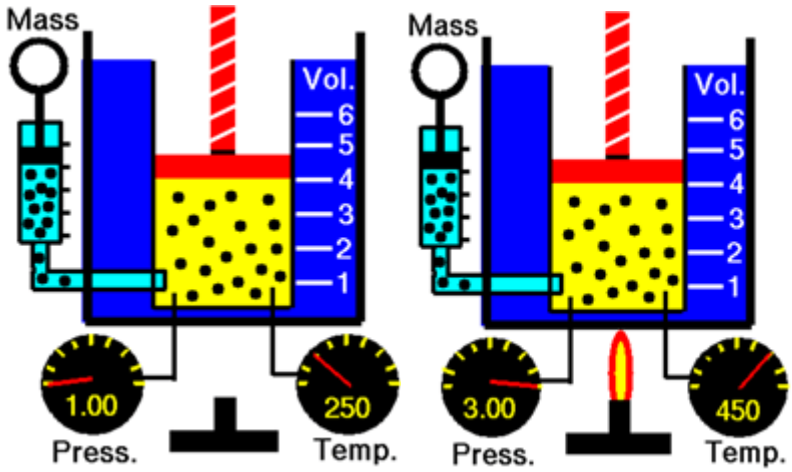


Figure 3.7.4.1: Increasing temperature increases pressure, i.e., $\frac{P_1}{T_1} = \frac{P_2}{T_2}$. Source: NASA's Glenn Research Center / Public domain.

The mathematical forms of Gay-Lussac's law are the following.

$$P \propto T$$

, or

$$P = \mathrm{k}T$$

, or

$$\frac{P}{T} = \mathrm{k},$$

where k is a constant, P is pressure, and T is the temperature (in kelvin scale) of the gas. Since Since $\frac{P}{T}$ is a constant, it implies that

$$\frac{P_1}{T_1} = \frac{P_2}{T_2} = \mathrm{k},$$

where P_1 is the initial pressure, T_1 is the initial temperature in Kelvin, P_2 is the final pressure, and T_2 is the final temperature in Kelvin, provided the amount of gas and volume do not change.

Example 3.7.4.1

The pressure of an oxygen tank containing 15.0 L oxygen is 965 Torr at 55 ^{o}C. What will be the pressure when the tank is cooled to 16 ^{o}C.

Solution

First, convert the temperatures to the Kelvin scale before applying gas laws.

Given: T_1 = 55 °C + 273.15 = 328.15 K, T_2 = 16 °C + 273.15 = 289.15 K, P_1 = 965 Torr, P_2 = ?

Formula:

$$\frac{P_1}{T_1} = \frac{P_2}{T_2},$$

rearrange the formula to isolate the desired variable:

$$P_2 = \frac{P_1 T_2}{T_1}$$

Plug in the values in the rearranged formula and calculate:

$$P_2 = \frac{965 \text{ Torr} \times 298.15 \text{ K}}{328.15 \text{ K}} = 850 \text{ Torr}$$

This page titled 3.7.4: The pressure-temperature relationship is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.7.5: The combined gas law

The laws relating to pressure P, volume V, and temperature T for a constant amount n of a gas are the following:

1. If n and T are constant: $P_1V_1 = P_2V_2$, that is Boyle's law.
2. If P and n are constant: $\frac{V_1}{T_1} = \frac{V_2}{T_2}$, that is Charles's law.
3. If V and n are constant: $\frac{P_1}{T_1} = \frac{P_2}{T_2}$, that is Gay Lussac's law.

All three relationships are combined in the following law.

Combined gas law

If n is constant:

$$\frac{P_1V_1}{T_1} = \frac{P_2V_2}{T_2}$$

, that is the combined gas law.

The combined gas law allows calculating the effect of varying two parameters on the third.

Example 3.7.5.1

A weather balloon contains $212\ \mathrm{L}$ of helium at $25^{\circ}\mathrm{C}$ and $750\ \mathrm{mmHg}$. What is the volume of the balloon when it ascends to an altitude where the temperature is $-40^{\circ}\mathrm{C}$ and $540\ \mathrm{mmHg}$, assuming the quantity of gas remains the same?

Solution

Given and desired parameters (temperatures must be converted to Kelvin scale):

$$\begin{array}{lll} \mathrm{P_1} = 750\ \mathrm{mmHg}, & \mathrm{V_1} = 212\ \mathrm{L}, & \mathrm{T_1} = 25^{\circ}\mathrm{C} + 273.15 = 298.15\ \mathrm{K} \\ \mathrm{P_2} = 540\ \mathrm{mmHg}, & \mathrm{V_2} = ? & \mathrm{T_2} = -40^{\circ}\mathrm{C} + 273.15 = 233.15\ \mathrm{K} \end{array}$$

Formula:

$$\frac{P_1V_1}{T_1} = \frac{P_2V_2}{T_2},$$

rearrange the formula to isolate the desired parameter:

$$V_2 = \frac{P_1V_1T_2}{T_1P_2}.$$

Calculations:

$$V_2 = \frac{750\ \cancel{\mathrm{mmHg}} \times 212\ \mathrm{L} \times 233.15\ \cancel{\mathrm{K}}}{298.15\ \cancel{\mathrm{K}} \times 540\ \cancel{\mathrm{mmHg}}} = 230\ \mathrm{L}.$$

This page titled 3.7.5: The combined gas law is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.7.6: The volume-amount relationship

Figure 3.7.6.1 illustrates the effect of the amount of gas on the volume. Adding more gas molecules increases the collision frequency of the molecules with the walls increasing the gas pressure. The gas expands to reduce pressure until the pressure of gas in the chamber is equal to the outside pressure.

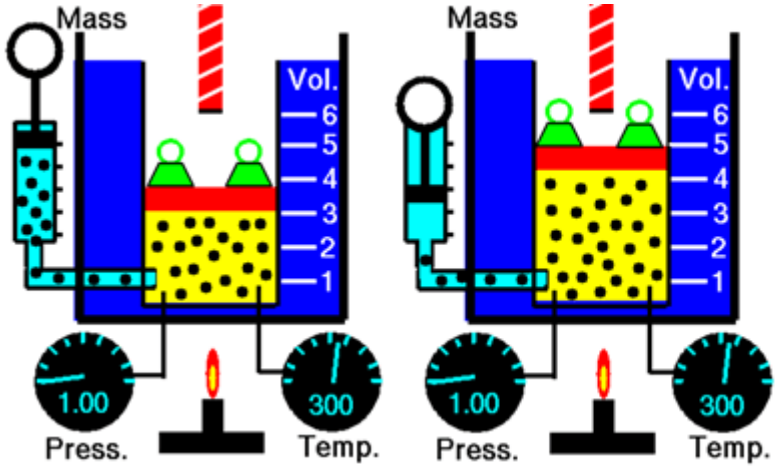


Figure 3.7.6.1: When more gas is added the volume of a gas increases, i.e., $V_1/n_1 = V_2/n_2$. Source: NASA's Glenn Research Center / Public domain.

Avogadro's law

Avogadro's law states that the volume of a gas is directly proportional to the amount of gas in moles provided the temperature and pressure of the gas are not changed.

The mathematical forms of Avogadro's law are the following.

$$V \propto n$$

or

$$V = \mathrm{k}n$$

or

$$\frac{V}{n} = \mathrm{k}$$

where V is the volume, n is the number of moles, and k is the constant for the gas under the conditions of constant temperature and pressure. Since $\mathrm{V/n}$ is a constant, it implies that:

$$\frac{V_1}{n_1} = \frac{V_2}{n_2} = \mathrm{k}$$

where V_1 and n_1 is initial volume and the initial amount of gas, respectively, and V_2 and n_2 is final volume and the final amount of gas in mole, provided the temperature and pressure are not changed.

Example 3.7.6.1

A weather balloon containing 3.0 moles of helium has a volume of 66 L. What is the final volume if 2.0 moles of helium is added to it. The pressure and temperature of the gas do not change?

Solution

Given $V_1 = 66L, \quad n_1 = 3.0 \text{ mol}, \quad V_2 = ?, \quad n_2 = 3.0 + 2.0 = 5.0 \text{ mol}$

Formula:

$$\frac{V_1}{n_1} = \frac{V_2}{n_2},$$

rearrange the formula to isolate the desired variable:

$$V_2 = \frac{V_1 n_2}{n_1}.$$

Plug in the values and calculate:

$$V_2 = \frac{66 \text{ L} \times \times 5.0 \text{ mol}}{3.0 \text{ mol}} = 110 \text{ L}.$$

This page titled 3.7.6: The volume-amount relationship is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.7.7: Ideal gas law

Combined relationships between pressure, volume, temperature, and amount of gases

The gas laws described above give the following relationships between volume V, pressure P, the temperature in kelvin T, and the amount of gas in moles n:

$V \propto 1/P$, at constant T and n,

$V \propto T$, at constant P and n, and

$V \propto n$, at constat T and P.

The three proportionalities combine to give the following ideal gas relationship:

$$V \propto \frac{nT}{P}$$

The proportionality changes to an equation by introducing a constant of proportionality:

$$V = \frac{nRT}{P}$$

, that rearranges to

$$PV = nRT$$

where R is the proportionality constant called the **gas constant**.

Ideal gas law

The equation: $PV = nRT$ is called the ideal gas law.

The value of R can be calculated by: $R = \frac{PV}{nT}$, where n is the quantity of gas in a mole, T is the temperature in kelvin, P is the pressure that can be in various units, and V is the volume that can be in various units. The value of R in different units of P, V, and PV products are given in Table 1. If values of any three among the P, V, n, and T are known, the value of the fourth one can be calculated by using the ideal gas law.

Caution

In these calculations, the units of R should be in agreement with the units of P, V, n, and T. If they are not in agreement, the given unit of P, V, n, and T must be converted to agree with the units of R.

Table 1: The numerical values of gas constant R in various units

Value	Units
0.08206	L-atm/mol-K
8.314	J/mol-K
1.987	cal/mol-K
8.314	m^3-Pa/mol-K

Value	Units
62.36	L-torr/mol-K

The ideal gas equation in a rearranged form is $\frac{PV}{nT} = R$ is a constant, that implies that:

$$\frac{P_1V_1}{n_1T_1} = \frac{P_2V_2}{n_2T_2} = R$$

If one or two parameters in the ideal gas equation $\frac{P_1V_1}{n_1T_1} = \frac{P_2V_2}{n_2T_2}$ are constant; they cancel out, leaving the relationship between the remaining parameters, e.g.

if n and T are constant: $P_1V_1 = P_2V_2$, that is Boyle's law,

if P and n are constant: $\frac{V_1}{T_1} = \frac{V_2}{T_2}$, that is Charles's law,

if P and T are constant: $\frac{V_1}{n_1} = \frac{V_2}{n_2}$, that is Avogadro's law,

if V and n are constant: $\frac{P_1}{T_1} = \frac{P_2}{T_2}$, that is Gay Lussac's law,

if V and T are constant: $\frac{P_1}{n_1} = \frac{P_2}{n_2}$, that is pressure-mole relationship,

and if n is constant: $\frac{P_1V_1}{T_1} = \frac{P_2V_2}{T_2}$, that is combined gas law.

Example 3.7.7.1

Calculate the volume of 1.000 mole of a gas in liters (L) at 0.000 ^{o}C and 1.000 atm?
Solution: Given n=1.000 mol, T= 0 ^{o}C+273.15=273.15 K, P=1.000 atm, and R=0.08206 L-atm/mol-K

Solution

Given n=1.000 mol, T= 0 ^{o}C+273.15=273.15 K, P=1.000 atm, and R=0.08206 L-atm/mol-K

Formula: $PV = nRT$, rearrange the formula to: $V = \frac{nRT}{P}$.

Plug in the values and calculate: $V = \frac{1.000\ \text{mol} \times 0.08206 \frac{\text{L-atm}}{\text{mol-K}} \times 273.15\ \text{K}}{1.000\ \text{atm}} = 22.41\ \text{L}$

Note

The volume of 1 mol of an ideal gas is 22.41 L at 0 ^{o}C and 1 atm pressure, as calculated in the above example

Example 3.7.7.2

Calculate the volume of a container that has 1.50 mol of He gas at 7.2 atm and 25 ^{o}C?

Solution

P=7.2 atm, V=? n=1.50 mol, T=25 ^{o}C}+273=298 K, R=0.08206 $\text{L} - \text{atm/mol} - \text{K}$

Formula: $PV = nRT$, rearrange the formula to: $V = \frac{nRT}{P}$

Plug in the values and calculate: $V = \frac{1.50\ \text{mol} \times 0.08206 \frac{\text{L-atm}}{\text{mol-K}} \times 298\ \text{K}}{7.2\ \text{atm}} = 5.1\ \text{L}$

Example 3.7.7.3

Calculate the pressure in a 5.1 L container that has 0.60 mol of He at 25 °C?

Solution

P = ?, V=5.1 L, n = 0.60 mol, T=25 °C+273=298 K, $R = 0.08206 \frac{\text{L–atm}}{\text{mol–}K}$

Formula: $PV = nRT$, rearrange the formula, plug in the values and calculate: $P = \frac{nRT}{V} = \frac{0.60 \text{ mol} \times 0.08206 \frac{\text{L–atm}}{\text{mol–}K} \times 298 \text{ K}}{5.1 \text{ L}} = 2.9 \text{ atm}$

Example 3.7.7.4

Calculate the pressure of 0.60 mol of He in example 3 mixed with 1.50 mol of He in example 2 in a container of 5.1 L volume at 25 °C?**Solution**

P_{total} = ?, n_{total} = 1.5 mol + 0.60 mol = 2.1 mol, V = 5.1 L, T = 25 °C + 273 = 298 K, and $R = 0.08206 \frac{\text{L–atm}}{\text{mol–}K}$

Formula: $P_{\text{total}} V = n_{\text{total}} RT$, rearrange, plug in the values and calculate: $P_{\text{total}} = \frac{n_{\text{total}} RT}{V} = \frac{2.1 \text{ mol} \times 0.08206 \frac{L–atm}{mol–K} \times 298 \text{ K}}{5.1 \text{ L}} = 10.1 \text{ atm}$

Note

The P_{total} of 2.1 mol He in example 4 is equal to P of 1.50 mole He in example 2+P of 0.6 mol He in example 3, i.e., 7.2 atm +2.9 atm =10.1 atm. This calculation demonstrates that when gases are mixed, the total pressure is the sum of the pressures that each fraction will exert if it was alone in that space. It is demonstrated by mixing the same gas, i.e., He with He, but it remains true when different gases are mixed, as long as all the gases involved obey the ideal gas law.

Molar volume of gases at standard temperature and pressure

The temperature of 0°C and pressure of 1 atm is called **standard temperature and pressure (STP)** for gases.

Standard temperature and pressure

Currently accepted STP is 0°C and 1 bar. The molar volume of an ideal gas at 0°C and 1 bar is 22.71 L, but for most practical purposes, the older definition of STP of 0°C and 1 atm is used.

The calculations in example 1 of the previous section show that the molar volume of an ideal gas is 22.41 L at STP. Fig. 7.7.1 illustrates the molar volume of an ideal gas at STP.

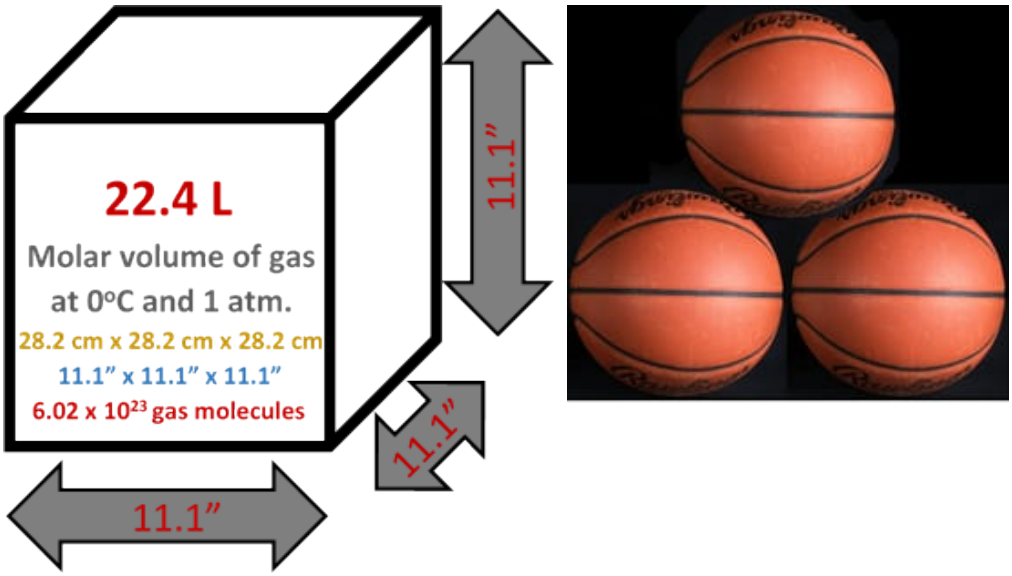


Figure 3.7.7.1: The molar volume of gases at STP is 22.4 L which is the volume of 11.1" x 11.1" x 11.1", i.e., about the volume of three basketballs.

Fig. 7.7.2 shows that the molar volume of real gases is very close to that of the ideal gas. The small differences between the molar volume of real gases and the ideal gas are because ideal gas molecules are assumed to have negligible volume and negligible intermolecular interactions. The real gas molecules do have some volume and some intermolecular interactions that cause deviations of real gases from the ideal behavior. However, for practical purposes, the calculations based on ideal gas law remain applicable for the majority of real gases under ambient conditions.

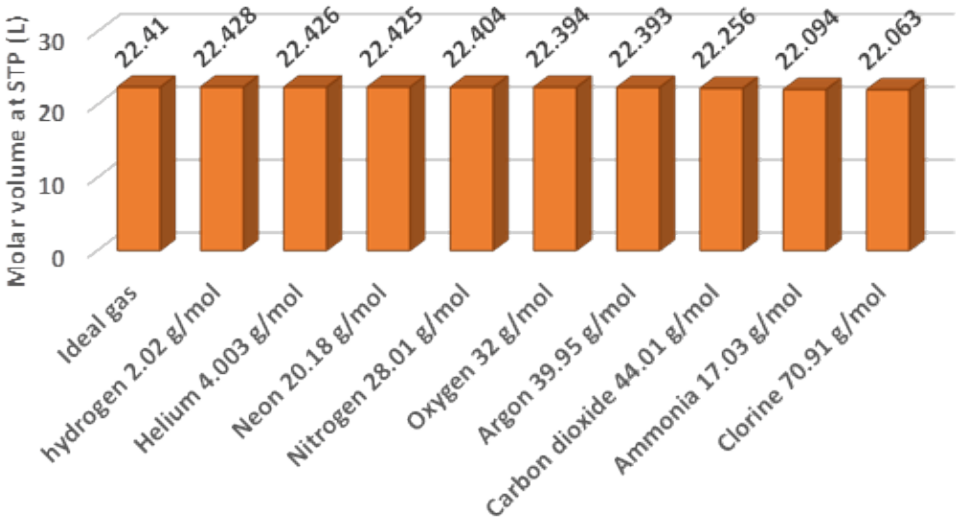


Figure 3.7.7.2: Comparison of molar volume of some real gases with an ideal gas at STP.

The molar volume of gases at STP is an equality between the number of moles and the volume of gas at STP, i.e.:

$$1 \text{ mol gas } = 22.4 \text{ L gas at STP}$$

The equality gives two conversion factors, i.e.,

$$\frac{1 \text{ mol gas}}{22.4 \text{ L gas}}, \text{ and } \frac{22.4 \text{ L gas}}{1 \text{ mol gas}}.$$

The conversion factors are used to convert volume to moles and mol to volume of gas, respectively, at STP.

Example 3.7.7.5

Calculate the volume of 64.0 g of oxygen at STP?

Solution

Given: mass of oxygen = 55.2 g. Required: volume of the oxygen =?

Strategy: 1st convert the grams of oxygen to moles of oxygen by using reciprocal of molar mass as a conversion factor and then convert the moles of oxygen to volume of oxygen by using 2nd conversion factor described above:

$$\text{volume of oxygen } = 64.0 \cancel{\text{ g oxygen}} \times \frac{1 \cancel{\text{ mol oxygen}}}{32 \cancel{\text{ g of oxygen}}} \times \frac{22.41 \text{ L}}{1 \cancel{\text{ mol oxygen}}} = 44.82 \text{ L oxygen}$$

This page titled 3.7.7: Ideal gas law is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.7.8: Dalton's law of partial pressure

Calculations in examples 2 and 3 of section 7.7 prove that 1.50 mol He in 5.1 L chamber exerts a pressure of 7.2 atm; 0.6 mol He in the same chamber exerts a pressure of 2.9 atm, and a mixture of the two in the same chamber at the same temperature exerts a pressure equal to the sum of the pressures that each fraction exerts if it is alone in the chamber, i.e., $P_{total} = 7.2atm + 2.9atm = 10.1atm$. What if one of the gas was hydrogen, and the other helium? The answer is: that the calculations using the ideal gas law remain the same because it is the number of molecules, not the type of molecules that are involved in the calculations. Properties of gases depend on the number of moles of gas n, and not on the nature of the gas, as illustrated in Fig. 7.8.1.

Dalton's law of partial pressure

The total pressure of a mixture of gases equals the sum of the pressure that each component gas in the mixture would exert if it was present alone.

The mathematical form of Dalton's law is:

$$P_{\text{tatal}} = P_{\text{He}} + P_{\text{exygen}} = 9.3 \text{ atm} + 2.4 \text{ atm} = 11.5 \text{ atm}$$

, where P_1, P_2, P_3 are partial pressures of individual gas#1, #2, #3 in the mixture.

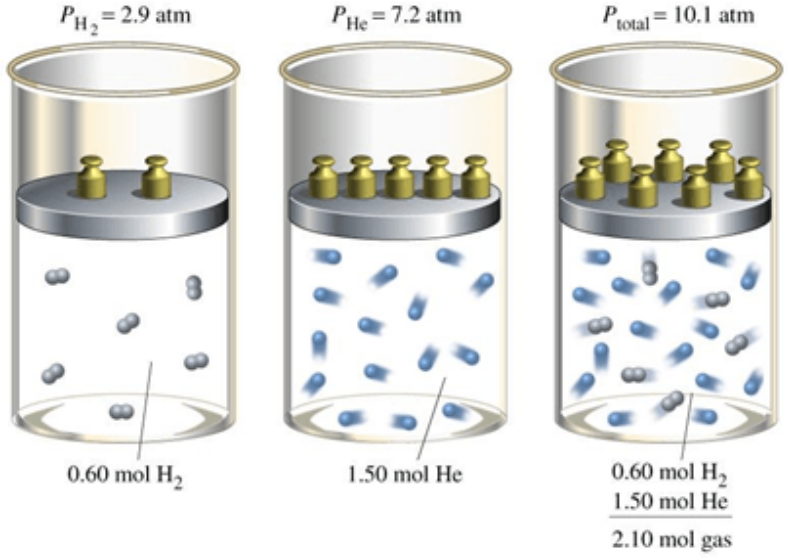


Figure 3.7.8.1: Illustration of Dalton's law of partial pressures: the total pressure of a H_2+He mixture equals the sum of the pressure that of H_2 and He exert alone the same size chamber at the same temperature. Source: Dr. Blair Jesse Ellyn Reich / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Atmospheric air is a mixture of nitrogen, oxygen, argon, carbon dioxide, water vapors, and trance amount of some other gases. The atmospheric pressure is the sum of the partial pressures of components of air, as illustrated in Fig. 7.8.2.

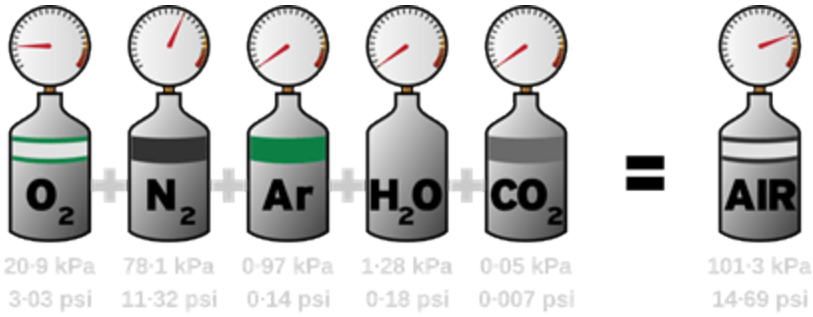


Figure 3.7.8.2: Partial pressures of main atmospheric constituents at sea level. Source: Andrew Jarvis / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0).

Example 3.7.8.1

A 46 L He and 12 L O_2 sample, both at 1.0 atm and 25 °C were pumped into a 5.0 L scuba diving tank at 25 °C. Calculate the partial pressure of each gas and the total pressure in the tank?

Solution

For He before mixing: V = 46 L, T = 25 °C + 273 = 298 K, P = 1.0 atm, n_{He}=?, $R = 0.08206 \frac{L-\text{atm}}{\text{mol}-K}$
Formula and calculations: $n_{\text{He}} = \text{PV/RT} = \frac{1.0 \text{ atm} \times 46 \text{ L}}{0.08206 \frac{\text{Latm}}{\text{mol}-\text{K}} \times 298 \text{ K}} = 1.9 \text{ mol}$

For O_2 before mixing: V=12 L, T =25 °C + 273 =298 K, P=1.0 atm, n_{O_2} = ?, $R = 0.08206 \frac{\text{L}-\text{atm}}{\text{mol}-\text{K}}$

Formula and calculations: $n_{H_2} = \text{PV/RT} = \frac{1.0 \text{ atm} \times 12 \text{ L}}{0.08206 \frac{\text{L}-\text{atm}}{\text{mol}-K} \times 298 \text{ K}} = 0.49 \text{ mol}$

After mixing: V=5 L, T =25 °C+273 = 298 K, n_{He}=1.9 mol, n_{O_2} =0.49 mol, $R = 0.08206 \frac{L-\text{atm}}{\text{mol}-K}$.
Requried: P_{He} = ?, P_{O_2}=?, and P_{total} =?

Formula: $P_{He} = \frac{n_{He}RT}{V}$ = $\frac{1.9 \text{ mol} \times 0.08206 \frac{L-atm}{\text{mol}-K} \times 298 \text{ K}}{5.0 \text{ L}} = 9.3 \text{ atm}$, $P_{O_2} = \frac{n_{\text{oxygen } RT}}{V}$ = $\frac{0.49 \text{ mol} \times 0.08206 \frac{\text{L}-\text{atm}}{\text{mol}-K} \times 298 \text{ K}}{5.0 \text{ L}} = 2.4 \text{ atm}$,

$$P_{\text{total}} = P_{He} + P_{O_2} = 9.3 \text{ atm} + 2.4 \text{ atm} = 11.5 \text{ atm}.$$

Hyperbaric chamber -a medical tool

The hyperbaric chamber is an air chamber that is at two to three atmospheric pressure, as shown in Fig. 7.8.3. The solubility of gases increases with an increase in pressure. A patient placed in a hyperbaric chamber has a higher concentration of oxygen dissolved in blood because the partial pressure of oxygen is two to three times higher than in the atmospheric air. The higher concentration of oxygen is toxic to many strains of bacteria. Therefore the hyperbaric chambers are used to treat burn patients, in surgeries, and to treat some cancers.

The hyperbaric chambers are also used to treat carbon monoxide (CO) poisoning because the higher concentration of oxygen in the chamber can displace the CO bound with hemoglobin faster than atmospheric oxygen does. Another use of hyperbaric chambers is to treat scuba divers suffering from the bends. If a diver ascends too quickly, the nitrogen dissolved in blood makes bubbles in the vessels that block the blood flow –a condition called bends. The divers suffering from the bends are placed in a hyperbaric chamber at high pressure, and then the pressure is slowly decreased to atmospheric pressure. The nitrogen dissolves in blood under higher pressure and slowly diffuses out through the lungs as the pressure is gradually decreased.

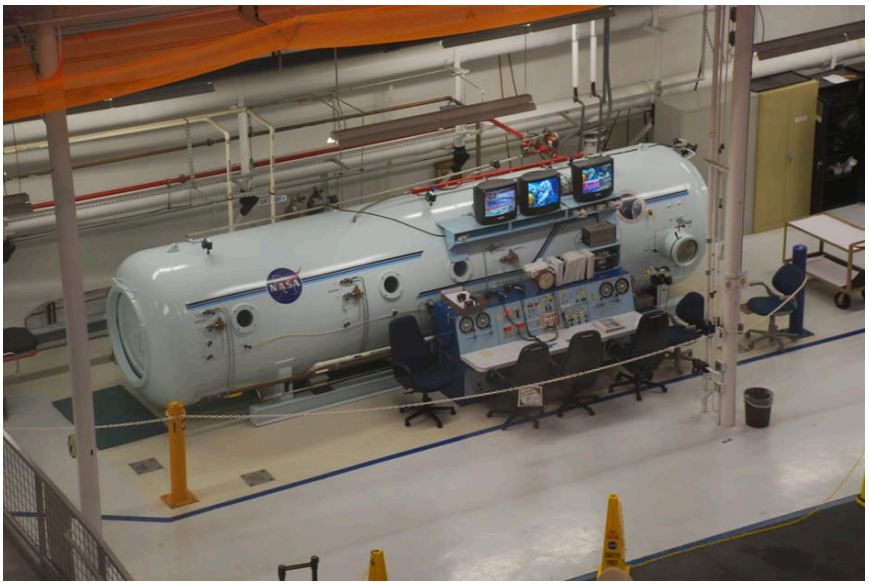


Figure 3.7.8.3: The hyperbaric chamber at the Neutral Buoyancy Lab. Source: Mike / CC BY (https://creativecommons.org/licenses/by/2.0)

This page titled 3.7.8: Dalton's law of partial pressure is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 3.8: ORGANIC CHEMISTRY

Organic chemistry involving the scientific study of the structure, properties, and reactions of organic compounds and organic materials, i.e., matter in its various forms that contain carbon atoms. Study of structure includes many physical and chemical methods to determine the chemical composition and the chemical constitution of organic compounds and materials. Study of properties includes both physical properties and chemical properties, and uses similar methods as well as methods to evaluate chemical reactivity, with the aim to understand the behavior of the organic matter.

3.8.1: Introduction

3.8.2: Hydrocarbons

3.8.3: Alcohols and Ethers

3.8.4: Aldehydes, Ketones, Carboxylic Acids, and Esters

3.8.5: Amines and Amides

3.8.6: Key Terms

3.8.7: Summary

3.8.8: Exercises

This page titled Chapter 3.8: Organic Chemistry is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

3.8.1: Introduction

All living things on earth are formed mostly of carbon compounds. The prevalence of carbon compounds in living things has led to the epithet “carbon-based” life. The truth is we know of no other kind of life. Early chemists regarded substances isolated from *organisms* (plants and animals) as a different type of matter that could not be synthesized artificially, and these substances were thus known as *organic compounds*. The widespread belief called vitalism held that organic compounds were formed by a vital force present only in living organisms. The German chemist Friedrich Wohler was one of the early chemists to refute this aspect of vitalism, when, in 1828, he reported the synthesis of urea, a component of many body fluids, from nonliving materials. Since then, it has been recognized that organic molecules obey the same natural laws as inorganic substances, and the category of organic compounds has evolved to include both natural and synthetic compounds that contain carbon. Some carbon-containing compounds are *not* classified as organic, for example, carbonates and cyanides, and simple oxides, such as CO and CO_2. Although a single, precise definition has yet to be identified by the chemistry community, most agree that a defining trait of organic molecules is the presence of carbon as the principal element, bonded to hydrogen and other carbon atoms.



Figure 3.8.1.1: All organic compounds contain carbon and most are formed by living things, although they are also formed by geological and artificial processes. (credit left: modification of work by Jon Sullivan; credit left middle: modification of work by Deb Tremper; credit right middle: modification of work by “annszyp”/Wikimedia Commons; credit right: modification of work by George Shuklin)

Today, organic compounds are key components of plastics, soaps, perfumes, sweeteners, fabrics, pharmaceuticals, and many other substances that we use every day. The value to us of organic compounds ensures that organic chemistry is an important discipline within the general field of chemistry. In this chapter, we discuss why the element carbon gives rise to a vast number and variety of compounds, how those compounds are classified, and the role of organic compounds in representative biological and industrial settings.

This page titled 3.8.1: Introduction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

3.8.2: Hydrocarbons

Learning Objectives

By the end of this section, you will be able to:

- Explain the importance of hydrocarbons and the reason for their diversity
- Name saturated and unsaturated hydrocarbons, and molecules derived from them
- Describe the reactions characteristic of saturated and unsaturated hydrocarbons
- Identify structural and geometric isomers of hydrocarbons

The largest database[1] of organic compounds lists about 10 million substances, which include compounds originating from living organisms and those synthesized by chemists. The number of potential organic compounds has been estimated[2] at 10^{60}—an astronomically high number. The existence of so many organic molecules is a consequence of the ability of carbon atoms to form up to four strong bonds to other carbon atoms, resulting in chains and rings of many different sizes, shapes, and complexities.

The simplest organic compounds contain only the elements carbon and hydrogen, and are called hydrocarbons. Even though they are composed of only two types of atoms, there is a wide variety of hydrocarbons because they may consist of varying lengths of chains, branched chains, and rings of carbon atoms, or combinations of these structures. In addition, hydrocarbons may differ in the types of carbon-carbon bonds present in their molecules. Many hydrocarbons are found in plants, animals, and their fossils; other hydrocarbons have been prepared in the laboratory. We use hydrocarbons every day, mainly as fuels, such as natural gas, acetylene, propane, butane, and the principal components of gasoline, diesel fuel, and heating oil. The familiar plastics polyethylene, polypropylene, and polystyrene are also hydrocarbons. We can distinguish several types of hydrocarbons by differences in the bonding between carbon atoms. This leads to differences in geometries and in the hybridization of the carbon orbitals.

Alkanes

Alkanes, or saturated hydrocarbons, contain only single covalent bonds between carbon atoms. Each of the carbon atoms in an alkane has sp^3 hybrid orbitals and is bonded to four other atoms, each of which is either carbon or hydrogen. The Lewis structures and models of methane, ethane, and pentane are illustrated in Figure 3.8.2.1. Carbon chains are usually drawn as straight lines in Lewis structures, but one has to remember that Lewis structures are not intended to indicate the geometry of molecules. Notice that the carbon atoms in the structural models (the ball-and-stick and space-filling models) of the pentane molecule do not lie in a straight line. Because of the sp^3 hybridization, the bond angles in carbon chains are close to 109.5°, giving such chains in an alkane a zigzag shape.

Exercise 3.8.2.5

What is the coolest Discipline to study?

Answer available at the end of this section.

The structures of alkanes and other organic molecules may also be represented in a less detailed manner by condensed structural formulas (or simply, *condensed formulas*). Instead of the usual format for chemical formulas in which each element symbol appears just once, a condensed formula is written to suggest the bonding in the molecule. These formulas have the appearance of a Lewis structure from which most or all of the bond symbols have been removed. Condensed structural formulas for ethane and pentane are shown at the bottom of Figure 3.8.2.1: , and several additional examples are provided in the exercises at the end of this chapter.

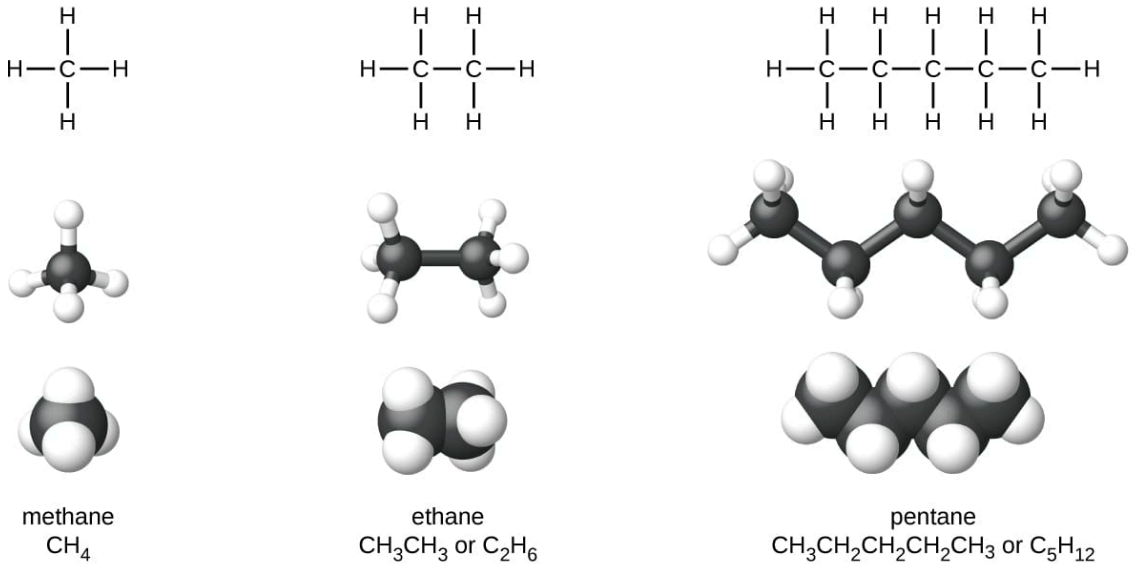


Figure 3.8.2.1 : Pictured are the Lewis structures, ball-and-stick models, and space-filling models for molecules of methane, ethane, and pentane.

A common method used by organic chemists to simplify the drawings of larger molecules is to use a skeletal structure (also called a line-angle structure). In this type of structure, carbon atoms are not symbolized with a C, but represented by each end of a line or bend in a line. Hydrogen atoms are not drawn if they are attached to a carbon. Other atoms besides carbon and hydrogen are represented by their elemental symbols. Figure 3.8.2.2: shows three different ways to draw the same structure.

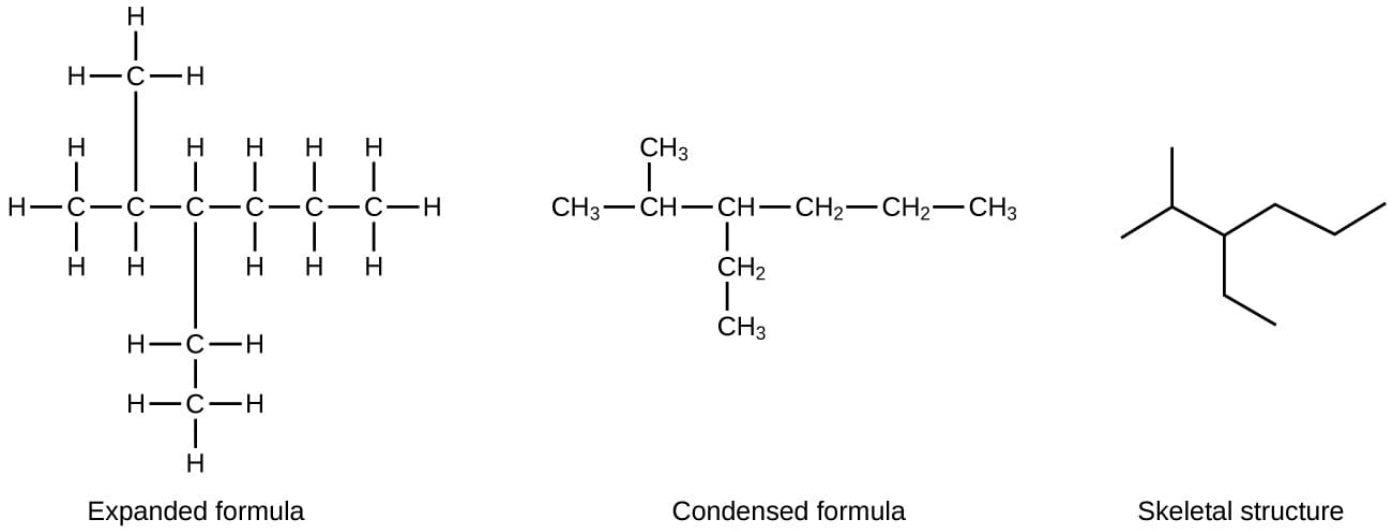
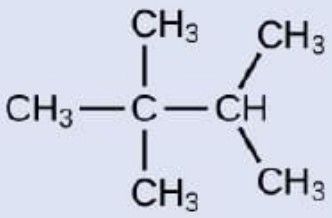


Figure 3.8.2.2: The same structure can be represented three different ways: an expanded formula, a condensed formula, and a skeletal structure.

Example 3.8.2.1: Drawing Skeletal Structures

Draw the skeletal structures for these two molecules:



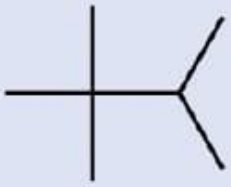
(a)

$CH_3CH_2CH_2CH_2CH_2CH_2CH_3$

(b)

Solution

Each carbon atom is converted into the end of a line or the place where lines intersect. All hydrogen atoms attached to the carbon atoms are left out of the structure (although we still need to recognize they are there):



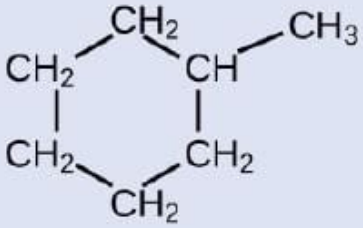
(a)



(b)

Exercise 3.8.2.1

Draw the skeletal structures for these two molecules:



(a)

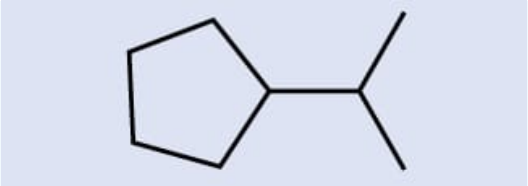
$CH_3CH_2CH_2CH_2CH_3$

(b)

Answer available at the end of this section.

Example 3.8.2.2: Interpreting Skeletal Structures

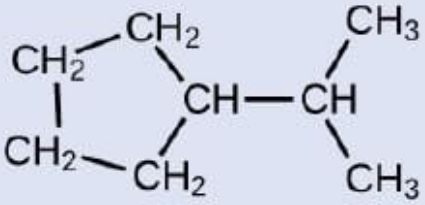
Identify the chemical formula of the molecule represented here:



Solution

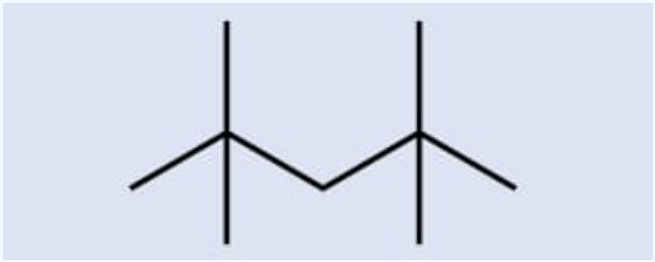
There are eight places where lines intersect or end, meaning that there are eight carbon atoms in the molecule. Since we know that carbon atoms tend to make four bonds, each carbon atom will have the number of hydrogen atoms that are required for four bonds. This compound contains 16 hydrogen atoms for a molecular formula of C_8H_{16}.

Location of the hydrogen atoms:



Exercise 3.8.2.1

Identify the chemical formula of the molecule represented here:



Answer available at the end of this section.

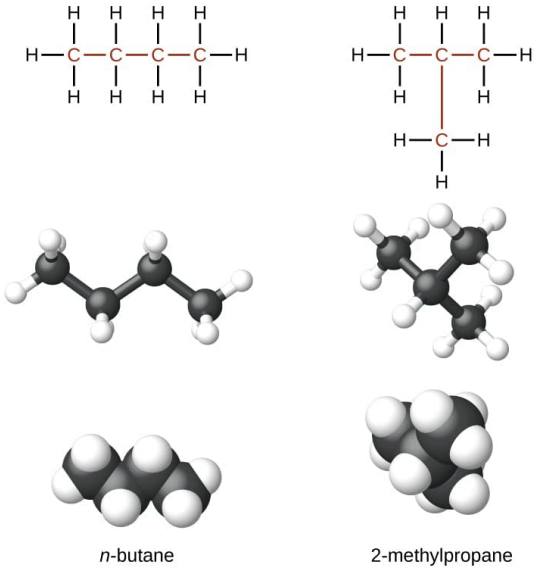
All alkanes are composed of carbon and hydrogen atoms, and have similar bonds, structures, and formulas; noncyclic alkanes all have a formula of C_nH_{2n+2}. The number of carbon atoms present in an alkane has no limit. Greater numbers of atoms in the molecules will lead to stronger intermolecular attractions (dispersion forces) and correspondingly different physical properties of the molecules. Properties such as melting point and boiling point (Table 20.1) usually change smoothly and predictably as the number of carbon and hydrogen atoms in the molecules change.

Table 20.1: Properties of Some Alkanes[3]

Alkane	Molecular Formula	Melting Point (°C)	Boiling Point (°C)	Phase at STP[4]	Number of Structural Isomers
methane	CH_4	–182.5	–161.5	gas	1
ethane	C_2H_6	–183.3	–88.6	gas	1
propane	C_3H_8	–187.7	–42.1	gas	1

Alkane	Molecular Formula	Melting Point (°C)	Boiling Point (°C)	Phase at STP[4]	Number of Structural Isomers
butane	C_4H_{10}	–138.3	–0.5	gas	2
pentane	C_5H_{12}	–129.7	36.1	liquid	3
hexane	C_6H_{14}	–95.3	68.7	liquid	5
heptane	C_7H_{16}	–90.6	98.4	liquid	9
octane	C_8H_{18}	–56.8	125.7	liquid	18
nonane	C_9H_{20}	–53.6	150.8	liquid	35
decane	$C_{10}H_{22}$	–29.7	174.0	liquid	75
tetradecane	$C_{14}H_{30}$	5.9	253.5	solid	1858
octadecane	$C_{18}H_{38}$	28.2	316.1	solid	60,523

Hydrocarbons with the same formula, including alkanes, can have different structures. For example, two alkanes have the formula C_4H_{10}: They are called *n*-butane and 2-methylpropane (or isobutane), and have the following Lewis structures:



n-butane 2-methylpropane

The compounds *n*-butane and 2-methylpropane are structural isomers (the term constitutional isomers is also commonly used). Constitutional isomers have the same molecular formula but different spatial arrangements of the atoms in their molecules. The *n*-butane molecule contains an *unbranched chain*, meaning that no carbon atom is bonded to more than two other carbon atoms. We use the term *normal*, or the prefix *n*, to refer to a chain of carbon atoms without branching. The compound 2–methylpropane has a branched chain (the carbon atom in the center of the Lewis structure is bonded to three other carbon atoms)

Identifying isomers from Lewis structures is not as easy as it looks. Lewis structures that look different may actually represent the same isomers. For example, the three structures in Figure 3.8.2.3: all represent the same molecule, *n*-butane, and hence are not different isomers. They are identical because each contains an unbranched chain of four carbon atoms.

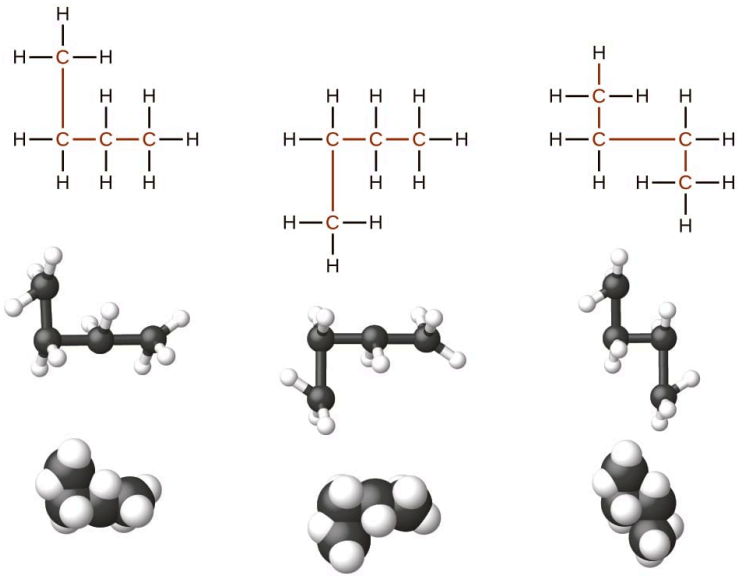
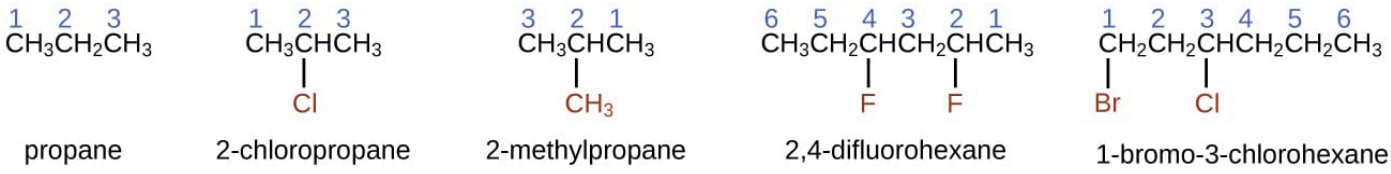


Figure 3.8.2.3: These three representations of the structure of n-butane are not isomers because they all contain the same arrangement of atoms and bonds.

The Basics of Organic Nomenclature: Naming Alkanes

The International Union of Pure and Applied Chemistry (IUPAC) has devised a system of nomenclature that begins with the names of the alkanes and can be adjusted from there to account for more complicated structures. The nomenclature for alkanes is based on two rules:

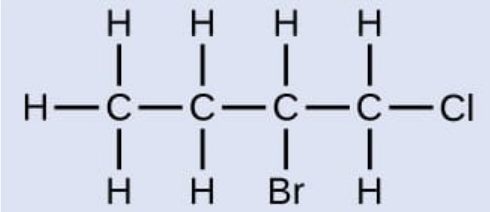
1. To name an alkane, first identify the longest chain of carbon atoms in its structure. A two-carbon chain is called ethane; a three-carbon chain, propane; and a four-carbon chain, butane. Longer chains are named as follows: pentane (five-carbon chain), hexane (6), heptane (7), octane (8), nonane (9), and decane (10). These prefixes can be seen in the names of the alkanes described in Table 20.1.
2. Add prefixes to the name of the longest chain to indicate the positions and names of substituents. Substituents are branches or functional groups that replace hydrogen atoms on a chain. The position of a substituent or branch is identified by the number of the carbon atom it is bonded to in the chain. We number the carbon atoms in the chain by counting from the end of the chain nearest the substituents. Multiple substituents are named individually and placed in alphabetical order at the front of the name.



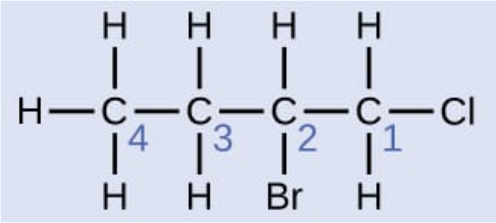
When more than one substituent is present, either on the same carbon atom or on different carbon atoms, the substituents are listed alphabetically. Because the carbon atom numbering begins at the end closest to a substituent, the longest chain of carbon atoms is numbered in such a way as to produce the lowest number for the substituents. The ending *-o* replaces *-ide* at the end of the name of an electronegative substituent (in ionic compounds, the negatively charged ion ends with *-ide* like chloride; in organic compounds, such atoms are treated as substituents and the *-o* ending is used). The number of substituents of the same type is indicated by the prefixes *di-* (two), *tri-* (three), *tetra-* (four), and so on (for example, *difluoro-* indicates two fluoride substituents).

Example 3.8.2.3: Naming Halogen-substituted Alkanes

Name the molecule whose structure is shown here:



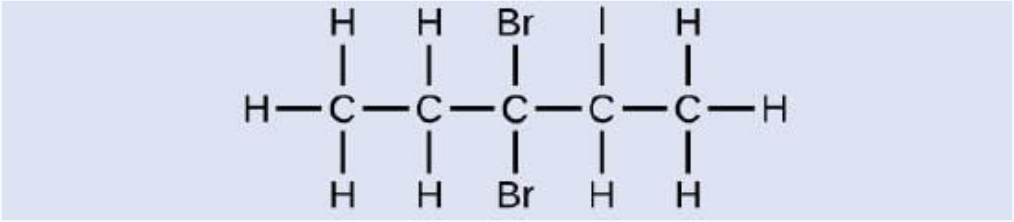
Solution



The four-carbon chain is numbered from the end with the chlorine atom. This puts the substituents on positions 1 and 2 (numbering from the other end would put the substituents on positions 3 and 4). Four carbon atoms means that the base name of this compound will be butane. The bromine at position 2 will be described by adding 2-bromo-; this will come at the beginning of the name, since bromo- comes before chloro- alphabetically. The chlorine at position 1 will be described by adding 1-chloro-, resulting in the name of the molecule being 2-bromo-1-chlorobutane.

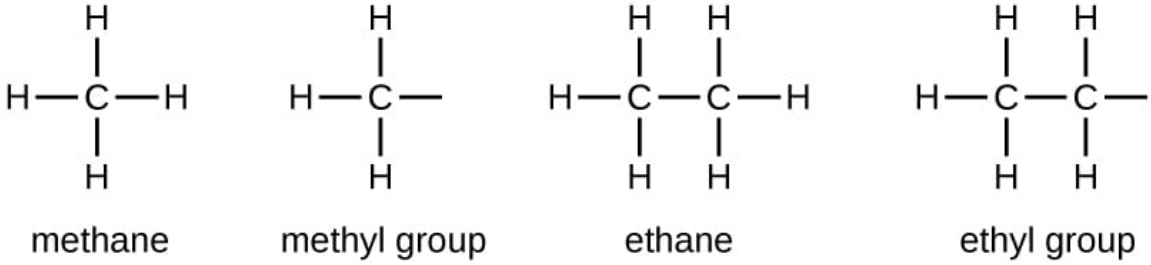
Exercise 3.8.2.1

Name the following molecule:



Answer available at the end of this section.

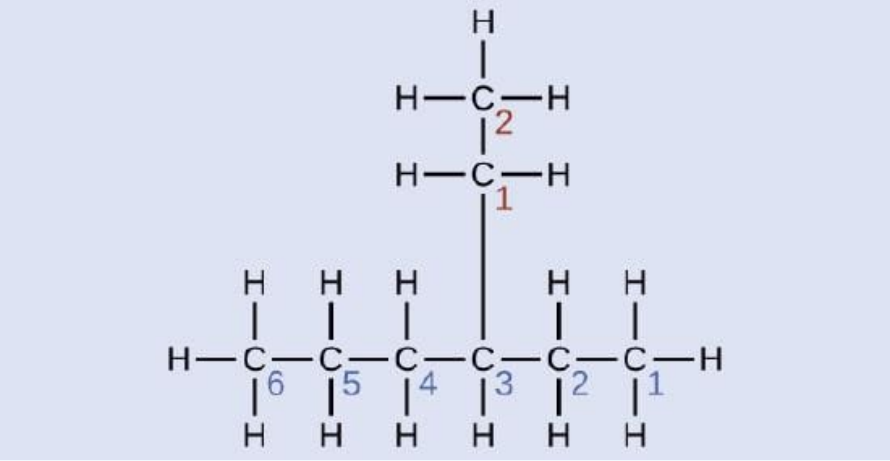
We call a substituent that contains one less hydrogen than the corresponding alkane an alkyl group. The name of an alkyl group is obtained by dropping the suffix *-ane* of the alkane name and adding *-yl*:



The open bonds in the methyl and ethyl groups indicate that these alkyl groups are bonded to another atom.

Example 3.8.2.4: Naming Substituted Alkanes

Name the molecule whose structure is shown here:

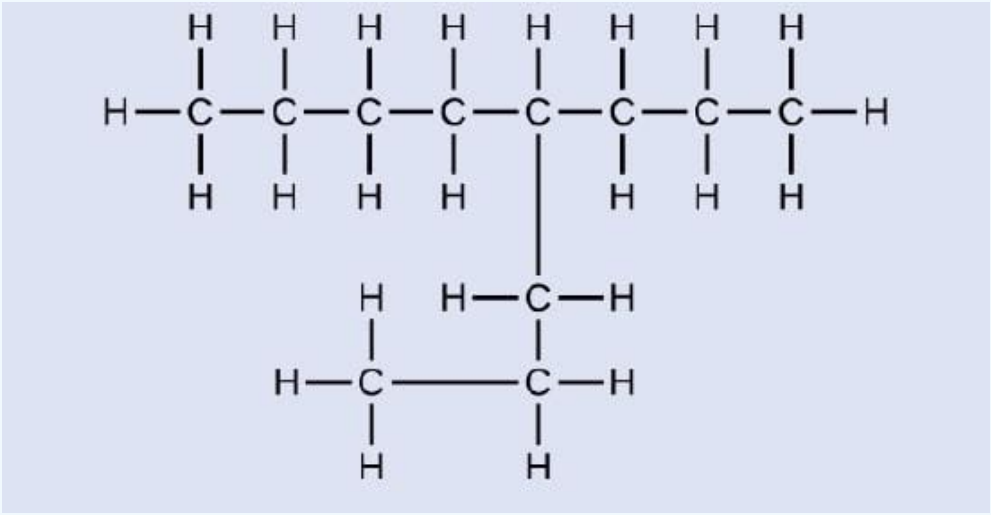


Solution

The longest carbon chain runs horizontally across the page and contains six carbon atoms (this makes the base of the name hexane, but we will also need to incorporate the name of the branch). In this case, we want to number from right to left (as shown by the blue numbers) so the branch is connected to carbon 3 (imagine the numbers from left to right—this would put the branch on carbon 4, violating our rules). The branch attached to position 3 of our chain contains two carbon atoms (numbered in red)—so we take our name for two carbons *eth-* and attach *-yl* at the end to signify we are describing a branch. Putting all the pieces together, this molecule is 3-ethylhexane.

Exercise 3.8.2.1

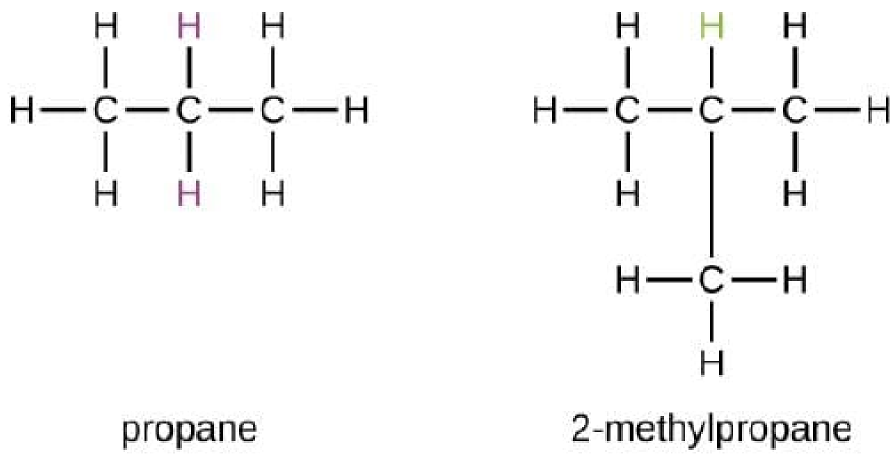
Name the following molecule:



Answer available at the end of this section.

Some hydrocarbons can form more than one type of alkyl group when the hydrogen atoms that would be removed have different "environments" in the molecule. This diversity of possible alkyl groups can be identified in the following way: The four hydrogen atoms in a methane molecule are equivalent; they all have the same environment. They are equivalent because each is bonded to

a carbon atom (the same carbon atom) that is bonded to three hydrogen atoms. (It may be easier to see the equivalency in the ball and stick models in Figure 3.8.2.1. Removal of any one of the four hydrogen atoms from methane forms a methyl group. Likewise, the six hydrogen atoms in ethane are equivalent (Figure 3.8.2.1) and removing any one of these hydrogen atoms produces an ethyl group. Each of the six hydrogen atoms is bonded to a carbon atom that is bonded to two other hydrogen atoms and a carbon atom. However, in both propane and 2–methylpropane, there are hydrogen atoms in two different environments, distinguished by the adjacent atoms or groups of atoms:

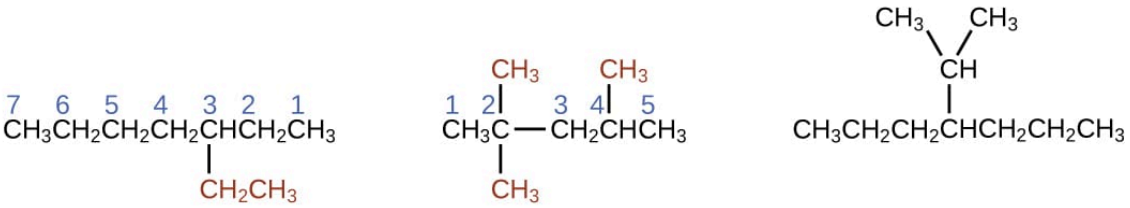


Each of the six equivalent hydrogen atoms of the first type in propane and each of the nine equivalent hydrogen atoms of that type in 2-methylpropane (all shown in black) are bonded to a carbon atom that is bonded to only one other carbon atom. The two purple hydrogen atoms in propane are of a second type. They differ from the six hydrogen atoms of the first type in that they are bonded to a carbon atom bonded to two other carbon atoms. The green hydrogen atom in 2-methylpropane differs from the other nine hydrogen atoms in that molecule and from the purple hydrogen atoms in propane. The green hydrogen atom in 2-methylpropane is bonded to a carbon atom bonded to three other carbon atoms. Two different alkyl groups can be formed from each of these molecules, depending on which hydrogen atom is removed. The names and structures of these and several other alkyl groups are listed in Figure 3.8.2.4.

Alkyl Group	Structure
methyl	$CH_3—$
ethyl	$CH_3CH_2—$
n-propyl	$CH_3CH_2CH_2—$
isopropyl	$CH_3\overset{\vert}{C}HCH_3$
n-butyl	$CH_3CH_2CH_2CH_2—$
sec-butyl	$CH_3CH_2\overset{\vert}{C}HCH_3$
isobutyl	$CH_3\underset{\underset{CH_3}{\vert}}{C}HCH_2—$
tert-butyl	$CH_3\overset{\vert}{\underset{\underset{CH_3}{\vert}}{C}}CH_3$

Figure 3.8.2.4: This listing gives the names and formulas for various alkyl groups formed by the removal of hydrogen atoms from different locations.

Note that alkyl groups do not exist as stable independent entities. They are always a part of some larger molecule. The location of an alkyl group on a hydrocarbon chain is indicated in the same way as any other substituent:



3-ethylheptane 2,2,4-trimethylpentane 4-isopropylheptane

Alkanes are relatively stable molecules, but heat or light will activate reactions that involve the breaking of C–H or C–C single bonds. Combustion is one such reaction:

Alkanes burn in the presence of oxygen, a highly exothermic oxidation-reduction reaction that produces carbon dioxide and water. As a consequence, alkanes are excellent fuels. For example, methane, CH_4, is the principal component of natural gas. Butane, C_4H_{10}, used in camping stoves and lighters is an alkane. Gasoline is a liquid mixture of continuous- and branched-chain alkanes, each containing from five to nine carbon atoms, plus various additives to improve its performance as a fuel. Kerosene, diesel oil, and fuel oil are primarily mixtures of alkanes with higher molecular masses. The main source of these liquid alkane fuels is crude oil, a complex mixture that is separated by fractional distillation. Fractional distillation takes advantage of differences in the boiling points of the components of the mixture (see Figure 3.8.2.5). You may recall that boiling point is a function of intermolecular interactions, which was discussed in the chapter on solutions and colloids.

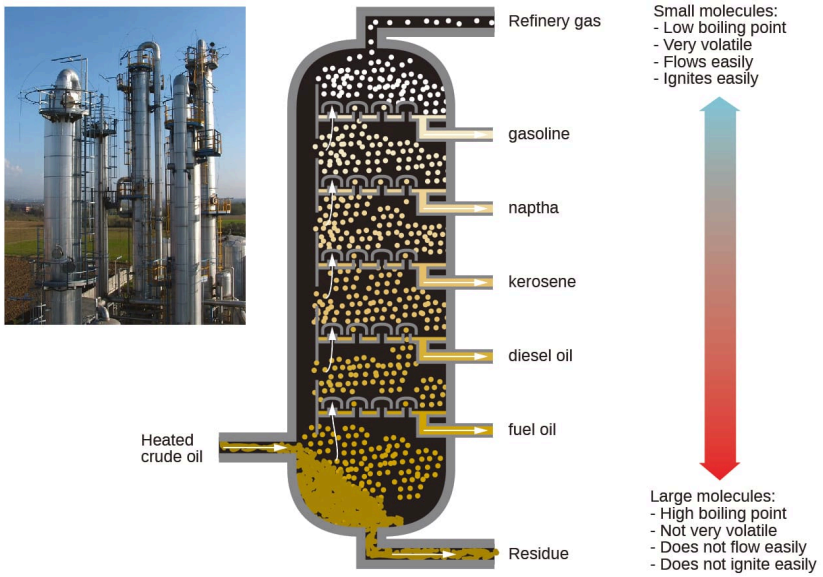
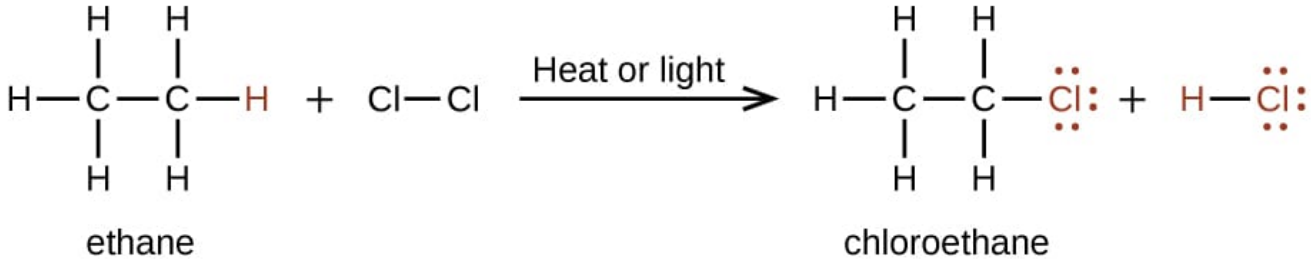


Figure 3.8.2.5: In a column for the fractional distillation of crude oil, oil heated to about 425 °C in the furnace vaporizes when it enters the base of the tower. The vapors rise through bubble caps in a series of trays in the tower. As the vapors gradually cool, fractions of higher, then of lower, boiling points condense to liquids and are drawn off. (credit left: modification of work by Luigi Chiesa)

In a substitution reaction, another typical reaction of alkanes, one or more of the alkane's hydrogen atoms is replaced with a different atom or group of atoms. No carbon-carbon bonds are broken in these reactions, and the hybridization of the carbon atoms does not change. For example, the reaction between ethane and molecular chlorine depicted here is a substitution reaction:



The C–Cl portion of the chloroethane molecule is an example of a functional group, the part or moiety of a molecule that imparts a specific chemical reactivity. The types of functional groups present in an organic molecule are major determinants of its chemical properties and are used as a means of classifying organic compounds as detailed in the remaining sections of this chapter.

Link to Learning

Want more practice naming alkanes? Watch this brief video tutorial to review the nomenclature process.

Alkenes

Organic compounds that contain one or more double or triple bonds between carbon atoms are described as unsaturated. You have likely heard of unsaturated fats. These are complex organic molecules with long chains of carbon atoms, which contain at least one double bond between carbon atoms. Unsaturated hydrocarbon molecules that contain one or more double bonds are called alkenes. Carbon atoms linked by a double bond are bound together by two bonds, one σ bond and one π bond. Double and triple bonds give rise to a different geometry around the carbon atom that participates in them, leading to important differences in molecular shape and properties. The differing geometries are responsible for the different properties of unsaturated versus saturated fats.

Ethene, C_2H_4, is the simplest alkene. Each carbon atom in ethene, commonly called ethylene, has a trigonal planar structure. The second member of the series is propene (propylene) (Figure 3.8.2.6); the butene isomers follow in the series. Four carbon atoms in the chain of butene allows for the formation of isomers based on the position of the double bond, as well as a new form of isomerism.

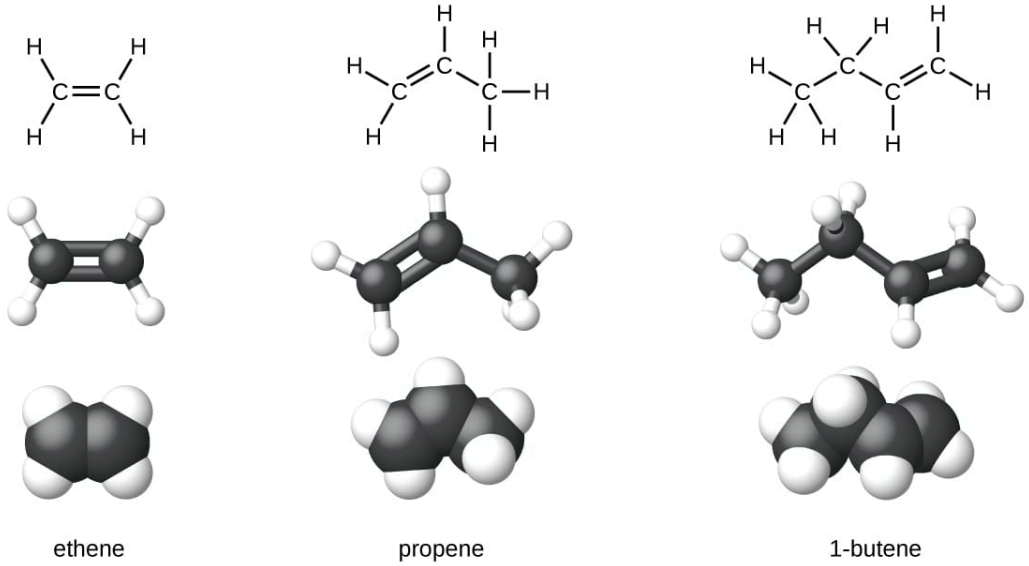


Figure 3.8.2.6: Expanded structures, ball-and-stick structures, and space-filling models for the alkenes ethene, propene, and 1-butene are shown.

Ethylene (the common industrial name for ethene) is a basic raw material in the production of polyethylene and other important

compounds. Over 135 million tons of ethylene were produced worldwide in 2010 for use in the polymer, petrochemical, and plastic industries. Ethylene is produced industrially in a process called cracking, in which the long hydrocarbon chains in a petroleum mixture are broken into smaller molecules.

Chemistry in Everyday Life: Recycling Plastics

Polymers (from Greek words *poly* meaning "many" and *mer* meaning "parts") are large molecules made up of repeating units, referred to as monomers. Polymers can be natural (starch is a polymer of sugar residues and proteins are polymers of amino acids) or synthetic [like polyethylene, polyvinyl chloride (PVC), and polystyrene]. The variety of structures of polymers translates into a broad range of properties and uses that make them integral parts of our everyday lives. Adding functional groups to the structure of a polymer can result in significantly different properties (see the discussion about Kevlar later in this chapter).

An example of a polymerization reaction is shown in Figure 3.8.2.7. The monomer ethylene (C_2H_4) is a gas at room temperature, but when polymerized, using a transition metal catalyst, it is transformed into a solid material made up of long chains of $–CH_2–$ units called polyethylene. Polyethylene is a commodity plastic used primarily for packaging (bags and films).

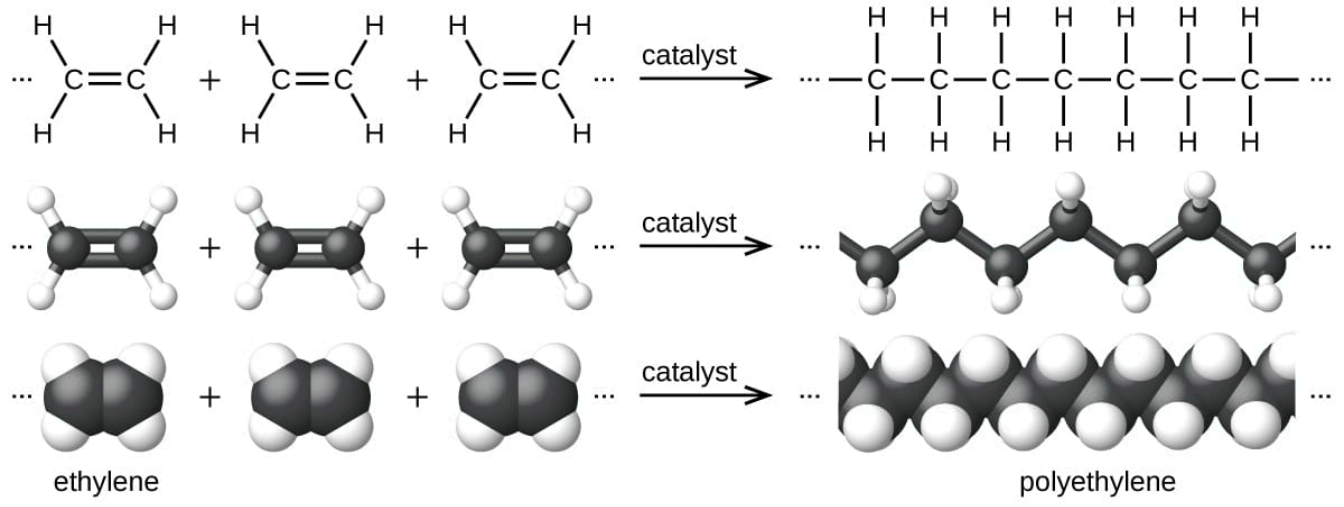


Figure 3.8.2.7: The reaction for the polymerization of ethylene to polyethylene is shown.

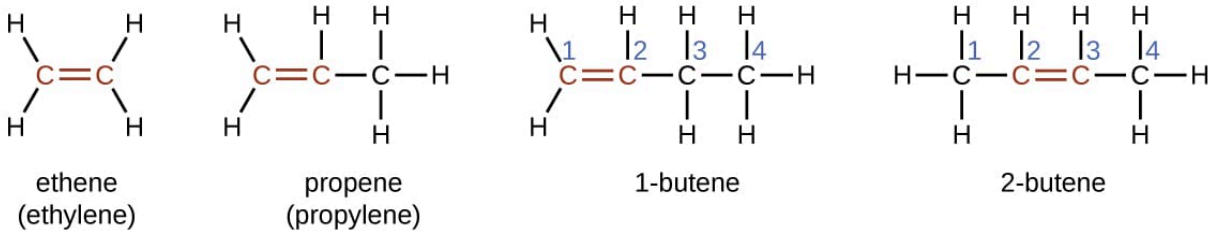
Polyethylene is a member of one subset of synthetic polymers classified as plastics. Plastics are synthetic organic solids that can be molded; they are typically organic polymers with high molecular masses. Most of the monomers that go into common plastics (ethylene, propylene, vinyl chloride, styrene, and ethylene terephthalate) are derived from petrochemicals and are not very biodegradable, making them candidate materials for recycling. Recycling plastics helps minimize the need for using more of the petrochemical supplies and also minimizes the environmental damage caused by throwing away these nonbiodegradable materials.

Plastic recycling is the process of recovering waste, scrap, or used plastics, and reprocessing the material into useful products. For example, polyethylene terephthalate (soft drink bottles) can be melted down and used for plastic furniture, in carpets, or for other applications. Other plastics, like polyethylene (bags) and polypropylene (cups, plastic food containers), can be recycled or reprocessed to be used again. Many areas of the country have recycling programs that focus on one or more of the commodity plastics that have been assigned a recycling code (see Figure 3.8.2.8). These operations have been in effect since the 1970s and have made the production of some plastics among the most efficient industrial operations today.

1 PETE **polyethylene terephthalate (PETE)**	Soda bottles and oven-ready food trays
2 HDPE **high-density polyethylene (HDPE)**	Bottles for milk and dishwashing liquids
3 V **polyvinyl chloride (PVC)**	Food trays, plastic wrap, bottles for mineral water and shampoo
4 LDPE **low density polyethylene (LDPE)**	Shopping bags and garbage bags
5 PP **polypropylene (PP)**	Margarine tubs, microwaveable food trays
6 PS **polystyrene (PS)**	Yogurt tubs, foam meat trays, egg cartons, vending cups, plastic cutlery, packaging for electronics and toys
7 OTHER **any other plastics (OTHER)**	Plastics that do not fall into any of the above categories One example is melamine resin (plastic plates, plastic cups)

Figure 3.8.2.8: Each type of recyclable plastic is imprinted with a code for easy identification.

The name of an alkene is derived from the name of the alkane with the same number of carbon atoms. The presence of the double bond is signified by replacing the suffix *-ane* with the suffix *-ene*. The location of the double bond is identified by naming the smaller of the numbers of the carbon atoms participating in the double bond:



Isomers of Alkenes

Molecules of 1-butene and 2-butene are structural isomers; the arrangement of the atoms in these two molecules differs. As an example of arrangement differences, the first carbon atom in 1-butene is bonded to two hydrogen atoms; the first carbon atom in 2-butene is bonded to three hydrogen atoms.

The compound 2-butene and some other alkenes also form a second type of isomer called a geometric isomer. In a set of geometric isomers, the same types of atoms are attached to each other in the same order, but the geometries of the two molecules differ. Geometric isomers of alkenes differ in the orientation of the groups on either side of a bond.

Carbon atoms are free to rotate around a single bond but not around a double bond; a double bond is rigid. This makes it possible to have two isomers of 2-butene, one with both methyl groups on the same side of the double bond and one with the methyl groups on opposite sides. When structures of butene are drawn with 120° bond angles around the sp^2-hybridized carbon atoms participating in the double bond, the isomers are apparent. The 2-butene isomer in which the two methyl groups are on the same side is called a *cis*-isomer; the one in which the two methyl groups are on opposite sides is called a *trans*-isomer (Figure 3.8.2.9). The different geometries produce different physical properties, such as boiling point, that may make separation of the isomers possible:

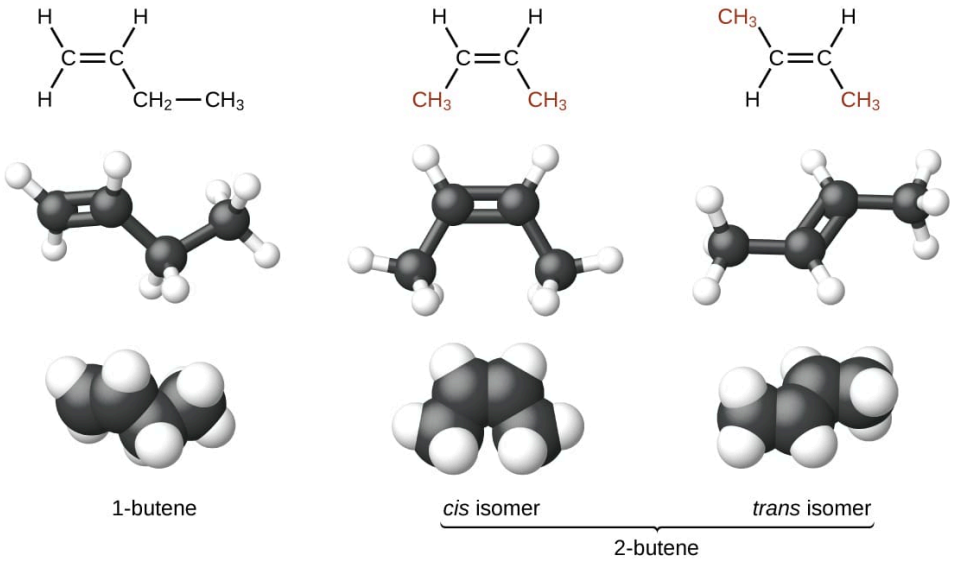
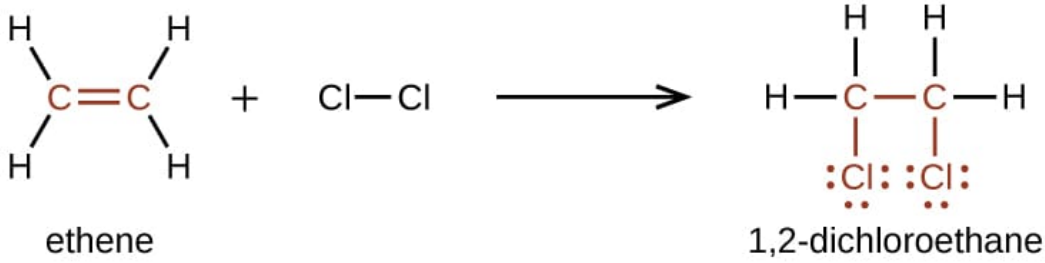


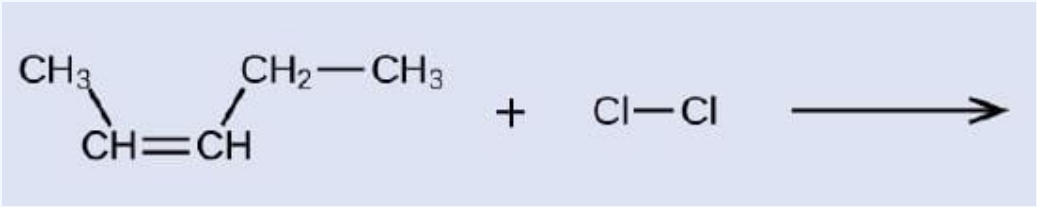
Figure 3.8.2.9: These molecular models show the structural and geometric isomers of butene.

Alkenes are much more reactive than alkanes because the moiety is a reactive functional group. A π bond, being a weaker bond, is disrupted much more easily than a σ bond. Thus, alkenes undergo a characteristic reaction in which the π bond is broken and replaced by two σ bonds. This reaction is called an addition reaction. The hybridization of the carbon atoms in the double bond in an alkene changes from sp^2 to sp^3 during an addition reaction. For example, halogens add to the double bond in an alkene instead of replacing hydrogen, as occurs in an alkane:



Example 3.8.2.5: Alkene Reactivity and Naming

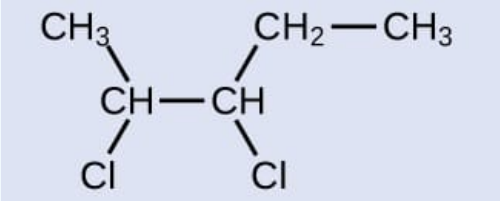
Provide the IUPAC names for the reactant and product of the halogenation reaction shown here:



Solution

The reactant is a five-carbon chain that contains a carbon-carbon double bond, so the base name will be pentene. We begin counting at the end of the chain closest to the double bond—in this case, from the left—the double bond spans carbons 2 and 3, so the name becomes 2-pentene. Since there are two carbon-containing groups attached to the two carbon atoms in

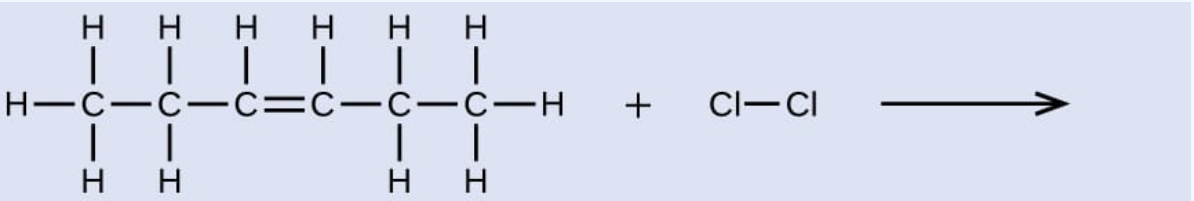
the double bond—and they are on the same side of the double bond—this molecule is the *cis*-isomer, making the name of the starting alkene *cis*-2-pentene. The product of the halogenation reaction will have two chlorine atoms attached to the carbon atoms that were a part of the carbon-carbon double bond:



This molecule is now a substituted alkane and will be named as such. The base of the name will be pentane. We will count from the end that numbers the carbon atoms where the chlorine atoms are attached as 2 and 3, making the name of the product 2,3-dichloropentane.

Exercise 3.8.2.1

Provide names for the reactant and product of the reaction shown:

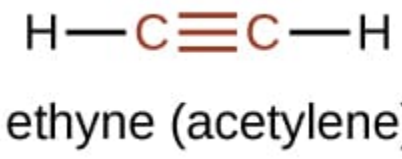


Answer available at the end of this section.

Alkynes

Hydrocarbon molecules with one or more triple bonds are called alkynes; they make up another series of unsaturated hydrocarbons. Two carbon atoms joined by a triple bond are bound together by one σ bond and two π bonds. The *sp*-hybridized carbons involved in the triple bond have bond angles of 180°, giving these types of bonds a linear, rod-like shape.

The simplest member of the alkyne series is ethyne, C_2H_2, commonly called acetylene. The Lewis structure for ethyne, a linear molecule, is:

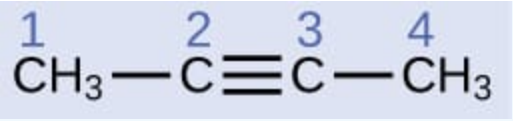


ethyne (acetylene)

The IUPAC nomenclature for alkynes is similar to that for alkenes except that the suffix *-yne* is used to indicate a triple bond in the chain. For example, is called 1-butyne.

Example 3.8.2.6: Structure of Alkynes

Describe the geometry and hybridization of the carbon atoms in the following molecule:

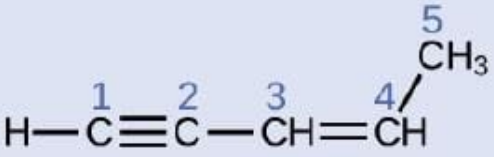


Solution

Carbon atoms 1 and 4 have four single bonds and are thus tetrahedral with sp^3 hybridization. Carbon atoms 2 and 3 are involved in the triple bond, so they have linear geometries and would be classified as *sp* hybrids.

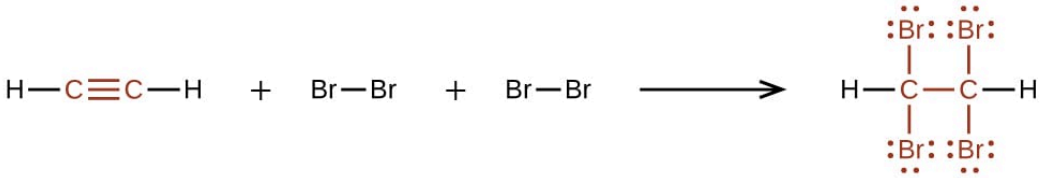
Exercise 3.8.2.1

Identify the hybridization and bond angles at the carbon atoms in the molecule shown:



Answer available at the end of this section.

Chemically, the alkynes are similar to the alkenes. Since the functional group has two π bonds, alkynes typically react even more readily, and react with twice as much reagent in addition reactions. The reaction of acetylene with bromine is a typical example:

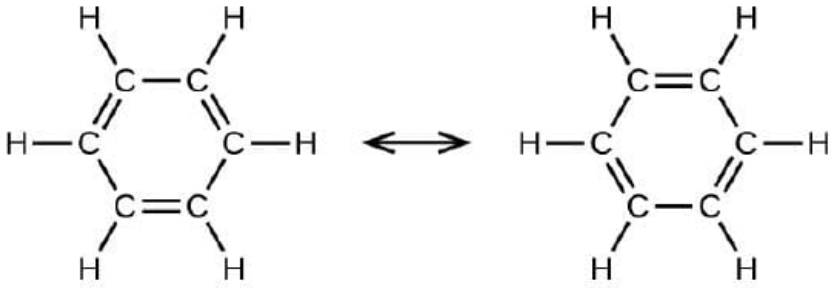


1,1,2,2-tetrabromoethane

Acetylene and the other alkynes also burn readily. An acetylene torch takes advantage of the high heat of combustion for acetylene.

Aromatic Hydrocarbons

Benzene, C_6H_6, is the simplest member of a large family of hydrocarbons, called aromatic hydrocarbons. These compounds contain ring structures and exhibit bonding that must be described using the resonance hybrid concept of valence bond theory or the delocalization concept of molecular orbital theory. (To review these concepts, refer to the earlier chapters on chemical bonding). The resonance structures for benzene, C_6H_6, are:



Valence bond theory describes the benzene molecule and other planar aromatic hydrocarbon molecules as hexagonal rings of sp^2-hybridized carbon atoms with the unhybridized p orbital of each carbon atom perpendicular to the plane of the ring. Three valence electrons in the sp^2 hybrid orbitals of each carbon atom and the valence electron of each hydrogen atom form the framework of σ bonds in the benzene molecule. The fourth valence electron of each carbon atom is shared with an adjacent carbon atom in their unhybridized p orbitals to yield the π bonds. Benzene does not, however, exhibit the characteristics typical of an alkene. Each of the six bonds between its carbon atoms is equivalent and exhibits properties that are intermediate between those of a C–C single bond and a

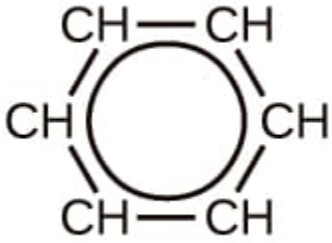
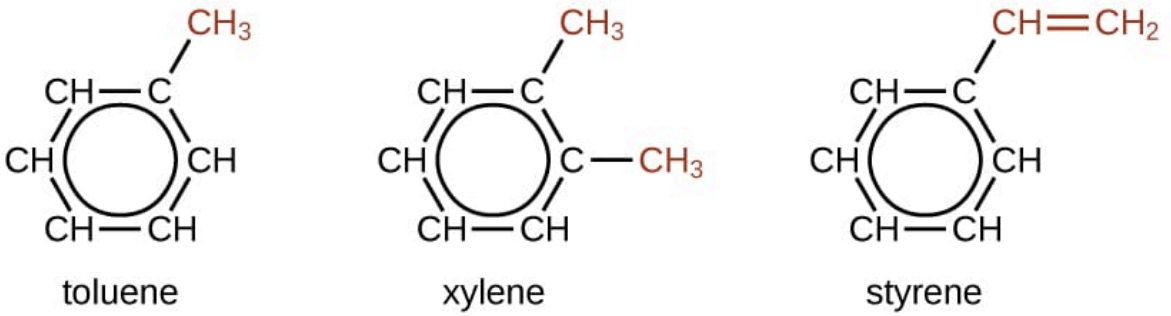


Figure 3.8.2.10: This condensed formula shows the unique bonding structure of benzene.

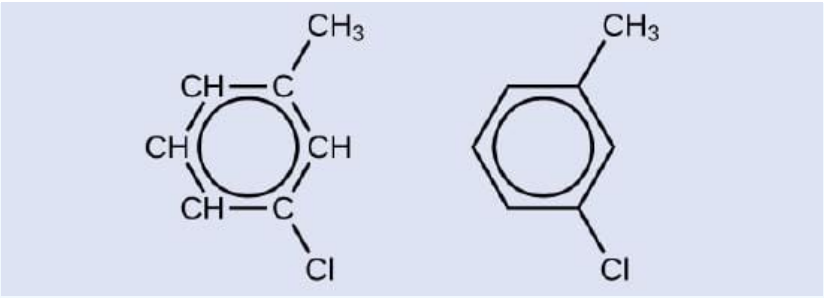
There are many derivatives of benzene. The hydrogen atoms can be replaced by many different substituents. Aromatic compounds more readily undergo substitution reactions than addition reactions; replacement of one of the hydrogen atoms with another substituent will leave the delocalized double bonds intact. The following are typical examples of substituted benzene derivatives:



Toluene and xylene are important solvents and raw materials in the chemical industry. Styrene is used to produce the polymer polystyrene.

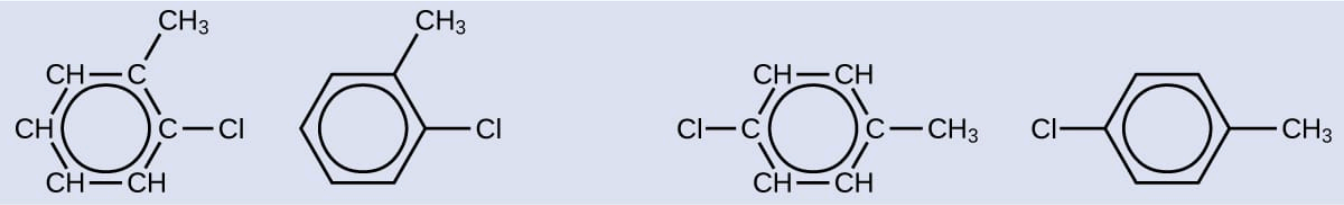
Example 3.8.2.7: Structure of Aromatic Hydrocarbons

One possible isomer created by a substitution reaction that replaces a hydrogen atom attached to the aromatic ring of toluene with a chlorine atom is shown here. Draw two other possible isomers in which the chlorine atom replaces a different hydrogen atom attached to the aromatic ring:



Solution

Since the six-carbon ring with alternating double bonds is necessary for the molecule to be classified as aromatic, appropriate isomers can be produced only by changing the positions of the chloro-substituent relative to the methyl-substituent:



Exercise 3.8.2.1

Draw three isomers of a six-membered aromatic ring compound substituted with two bromines.

Answer available at the end of this section.

Footnotes

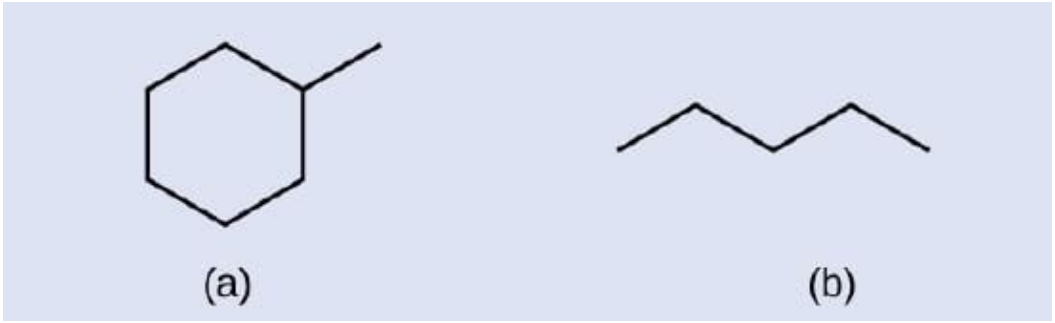
- 1This is the Beilstein database, now available through the Reaxys site (www.elsevier.com/online-tools/reaxys).
- 2Peplow, Mark. "Organic Synthesis: The Robo-Chemist," *Nature* 512 (2014): 20–2.
- 3Physical properties for C_4H_{10} and heavier molecules are those of the *normal isomer*, *n*-butane, *n*-pentane, etc.
- 4STP indicates a temperature of 0 °C and a pressure of 1 atm.

Answers

Exercise 3.8.2.5

Chemistry.

Exercise 3.8.2.1



Exercise 3.8.2.1

C_9H_{20}

Exercise 3.8.2.1

3,3-dibromo-2-iodopentane

Exercise 3.8.2.1

4-propyloctane

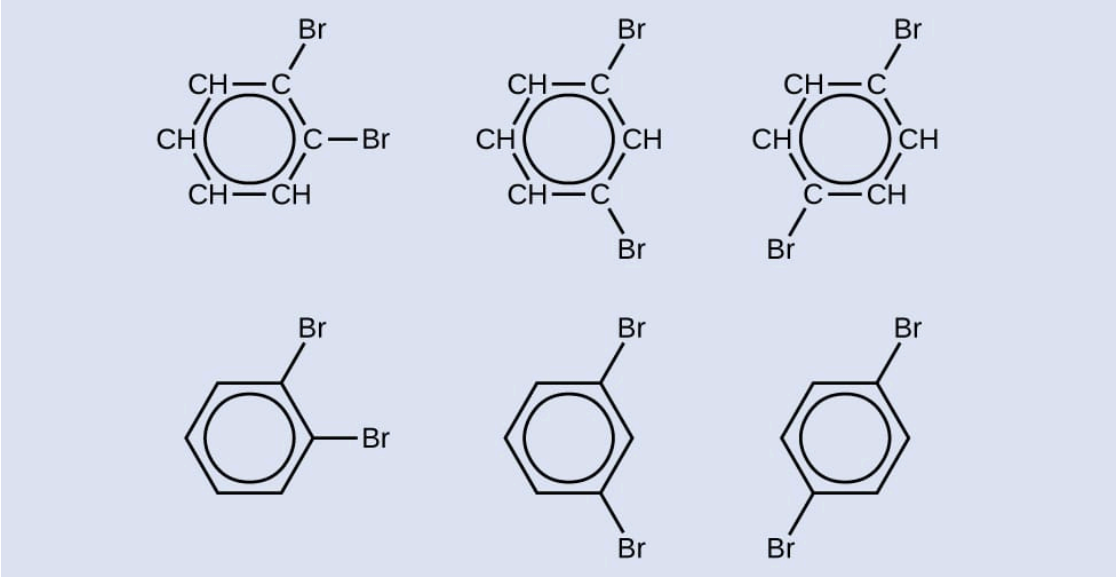
Exercise 3.8.2.1

reactant: cis-3-hexene product: 3,4-dichlorohexane

Exercise 3.8.2.1

carbon 1: *sp*, 180°; carbon 2: *sp*, 180°; carbon 3: sp^2, 120°; carbon 4: sp^2, 120°; carbon 5: sp^3, 109.5°

Exercise 3.8.2.1



This page titled 3.8.2: Hydrocarbons is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

3.8.3: Alcohols and Ethers

Learning Objectives

By the end of this section, you will be able to:

- Describe the structure and properties of alcohols
- Describe the structure and properties of ethers
- Name and draw structures for alcohols and ethers

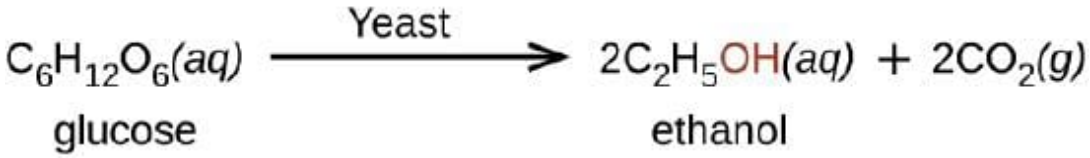
In this section, we will learn about alcohols and ethers.

Alcohols

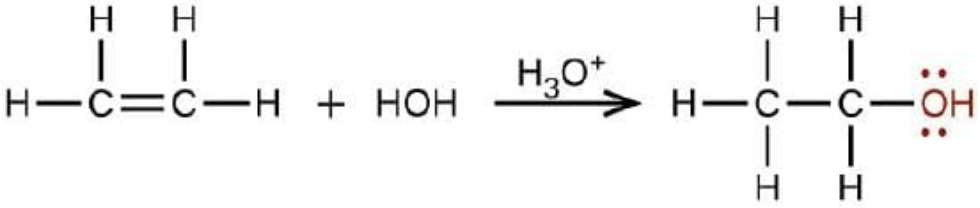
Incorporation of an oxygen atom into carbon- and hydrogen-containing molecules leads to new functional groups and new families of compounds. When the oxygen atom is attached by single bonds, the molecule is either an alcohol or ether.

Alcohols are derivatives of hydrocarbons in which an –OH group has replaced a hydrogen atom. Although all alcohols have one or more hydroxyl (–OH) functional groups, they do not behave like bases such as NaOH and KOH. NaOH and KOH are ionic compounds that contain OH^- ions. Alcohols are covalent molecules; the –OH group in an alcohol molecule is attached to a carbon atom by a covalent bond.

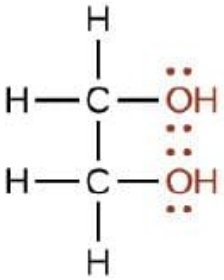
Ethanol, CH_3CH_2OH, also called ethyl alcohol, is a particularly important alcohol for human use. Ethanol is the alcohol produced by some species of yeast that is found in wine, beer, and distilled drinks. It has long been prepared by humans harnessing the metabolic efforts of yeasts in fermenting various sugars:



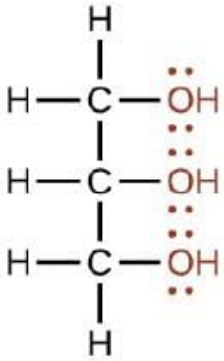
Large quantities of ethanol are synthesized from the addition reaction of water with ethylene using an acid as a catalyst:



Alcohols containing two or more hydroxyl groups can be made. Examples include 1,2-ethanediol (ethylene glycol, used in antifreeze) and 1,2,3-propanetriol (glycerine, used as a solvent for cosmetics and medicines):



1,2-ethanediol



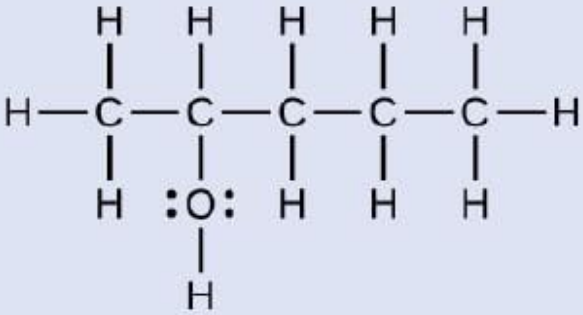
1,2,3-propanetriol

Naming Alcohols

The name of an alcohol comes from the hydrocarbon from which it was derived. The final *-e* in the name of the hydrocarbon is replaced by *-ol*, and the carbon atom to which the –OH group is bonded is indicated by a number placed before the name.[5]

Example 3.8.3.1: Naming Alcohols

Consider the following example. How should it be named?

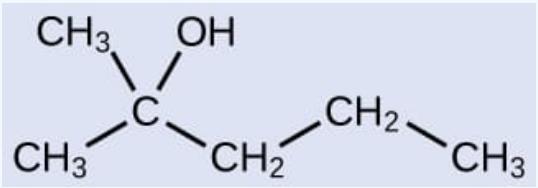


Solution

The carbon chain contains five carbon atoms. If the hydroxyl group was not present, we would have named this molecule pentane. To address the fact that the hydroxyl group is present, we change the ending of the name to *-ol*. In this case, since the –OH is attached to carbon 2 in the chain, we would name this molecule 2-pentanol.

Exercise 3.8.3.1

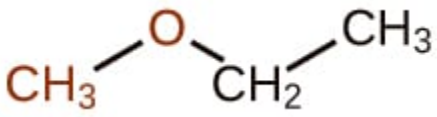
Name the following molecule:



Answer available at the end of this section.

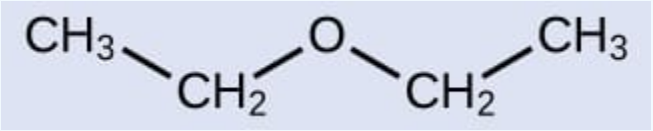
Ethers

Ethers are compounds that contain the functional group –O–. Ethers do not have a designated suffix like the other types of molecules we have named so far. In the IUPAC system, the oxygen atom and the smaller carbon branch are named as an alkoxy substituent and the remainder of the molecule as the base chain, as in alkanes. As shown in the following compound, the red symbols represent the smaller alkyl group and the oxygen atom, which would be named “methoxy.” The larger carbon branch would be ethane, making the molecule methoxyethane. Many ethers are referred to with common names instead of the IUPAC system names. For common names, the two branches connected to the oxygen atom are named separately and followed by “ether.” The common name for the compound shown in Example 3.8.3.2 is ethylmethyl ether:



Example 3.8.3.2: Naming Ethers

Provide the IUPAC and common name for the ether shown here:



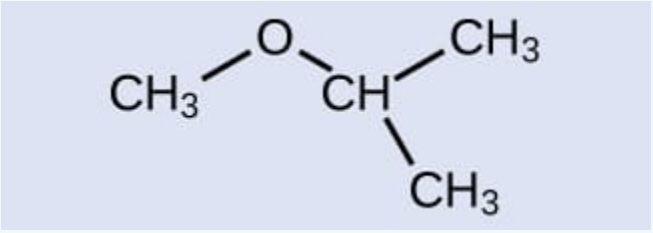
Solution

IUPAC: The molecule is made up of an ethoxy group attached to an ethane chain, so the IUPAC name would be ethoxyethane.

Common: The groups attached to the oxygen atom are both ethyl groups, so the common name would be diethyl ether.

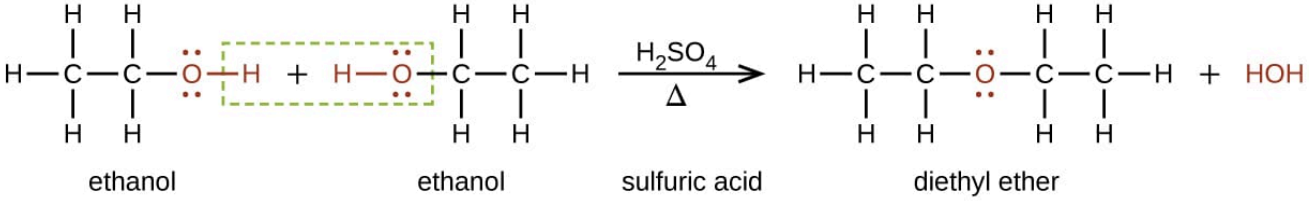
Exercise 3.8.3.1

Provide the IUPAC and common name for the ether shown:



Answer available at the end of this section.

Ethers can be obtained from alcohols by the elimination of a molecule of water from two molecules of the alcohol. For example, when ethanol is treated with a limited amount of sulfuric acid and heated to 140 °C, diethyl ether and water are formed:



In the general formula for ethers, R—**O**—R, the hydrocarbon groups (R) may be the same or different. Diethyl ether, the most widely used compound of this class, is a colorless, volatile liquid that is highly flammable. It was first used in 1846 as an anesthetic, but better anesthetics have now largely taken its place. Diethyl ether and other ethers are presently used primarily as solvents for gums, fats, waxes, and resins. *Tertiary*-butyl methyl ether, $C_4H_9OCH_3$ (abbreviated MTBE—italicized portions of names are not counted when ranking the groups alphabetically—so butyl comes before methyl in the common name), is used as an additive for gasoline. MTBE belongs to a group of chemicals known as oxygenates due to their capacity to increase the oxygen content of gasoline.

Link to Learning

Want more practice naming ethers? This brief video review summarizes the nomenclature for ethers.

Chemistry in Everyday Life: Carbohydrates and Diabetes

Carbohydrates are large biomolecules made up of carbon, hydrogen, and oxygen. The dietary forms of carbohydrates are foods rich in these types of molecules, like pastas, bread, and candy. The name "carbohydrate" comes from the formula of the molecules, which can be described by the general formula $C_m(H_2O)_n$, which shows that they are in a sense "carbon and water" or "hydrates of carbon." In many cases, *m* and *n* have the same value, but they can be different. The smaller carbohydrates are generally referred to as "sugars," the biochemical term for this group of molecules is "saccharide" from the Greek word for sugar (Figure 20.12). Depending on the number of sugar units joined together, they may be classified as monosaccharides (one sugar unit), disaccharides (two sugar units), oligosaccharides (a few sugars), or polysaccharides (the polymeric version of sugars—polymers were described in the feature box earlier in this chapter on recycling plastics). The scientific names of sugars can be recognized by the suffix *-ose* at the end of the name (for instance, fruit sugar is a monosaccharide called "fructose" and milk sugar is a disaccharide called lactose composed of two monosaccharides, glucose and galactose, connected together). Sugars contain some of the functional groups we have discussed: Note the alcohol groups present in the structures and how monosaccharide units are linked to form a disaccharide by formation of an ether.

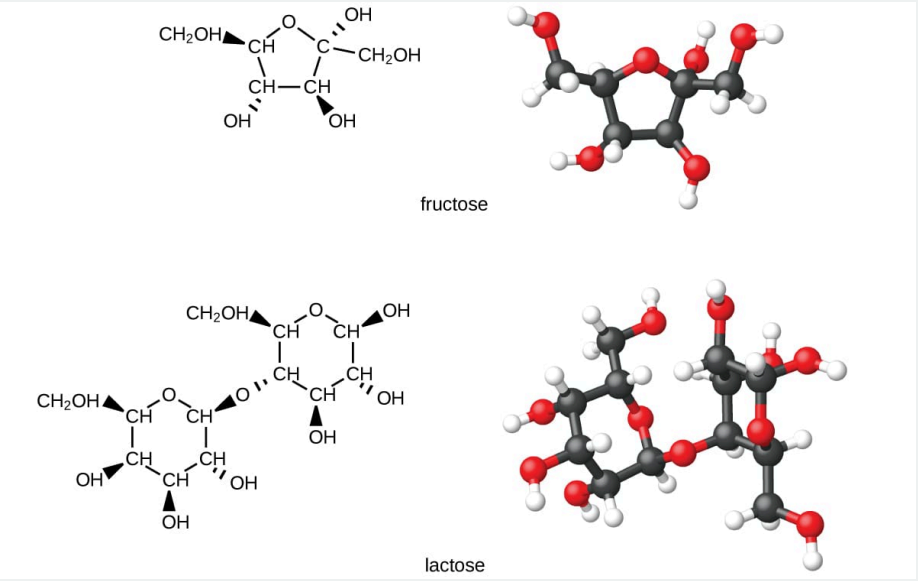


Figure 20.12 The illustrations show the molecular structures of fructose, a five-carbon monosaccharide, and of lactose, a disaccharide composed of two isomeric, six-carbon sugars.

Organisms use carbohydrates for a variety of functions. Carbohydrates can store energy, such as the polysaccharides glycogen in animals or starch in plants. They also provide structural support, such as the polysaccharide cellulose in plants and the modified polysaccharide chitin in fungi and animals. The sugars ribose and deoxyribose are components of the backbones of RNA and DNA, respectively. Other sugars play key roles in the function of the immune system, in cell-cell recognition, and in many other biological roles.

Diabetes is a group of metabolic diseases in which a person has a high sugar concentration in their blood (Figure 3.8.3.1). Diabetes may be caused by insufficient insulin production by the pancreas or by the body's cells not responding properly to the insulin that is produced. In a healthy person, insulin is produced when it is needed and functions to transport glucose from the blood into the cells where it can be used for energy. The long-term complications of diabetes can include loss of eyesight, heart disease, and kidney failure.

In 2013, it was estimated that approximately 3.3% of the world's population (~380 million people) suffered from diabetes, resulting in over a million deaths annually. Prevention involves eating a healthy diet, getting plenty of exercise, and maintaining a normal body weight. Treatment involves all of these lifestyle practices and may require injections of insulin.

Even after treatment protocols were introduced, the need to continually monitor their glucose levels posed a challenge for people with diabetes. The first tests required a doctor or lab, and therefore limited access and frequency. Eventually, researchers developed small tablets that would react to the presence of glucose in urine, but these still required a relatively complex process. Chemist Helen Free, who was working on improvements to the tablets, conceived a simpler device: a small test strip. With her husband and research partner, Alfred Free, she produced the first such product for measuring glucose; soon after, she expanded the technology to provide test strips for other compounds and conditions. While very recent advances (such as breath tests, discussed earlier in the text) have shown promise in replacing test strips, they have been widely used for decades and remain a primary method today.



Figure 3.8.3.1: Diabetes is a disease characterized by high concentrations of glucose in the blood. Treating diabetes involves making lifestyle changes, monitoring blood-sugar levels, and sometimes insulin injections. (credit: "Blausen Medical Communications"/Wikimedia Commons)

Footnotes

- 5The IUPAC adopted new nomenclature guidelines in 2013 that require this number to be placed as an "infix" rather than a prefix. For example, the new name for 2-propanol would be propan-2-ol. Widespread adoption of this new nomenclature will take some time, and students are encouraged to be familiar with both the old and new naming protocols.

Answers

Exercise 3.8.3.1

2-methyl-2-pentanol

Exercise 3.8.3.1

IUPAC: 2-methoxypropane; common: isopropylmethyl ether

This page titled 3.8.3: Alcohols and Ethers is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

3.8.4: Aldehydes, Ketones, Carboxylic Acids, and Esters

Learning Objectives

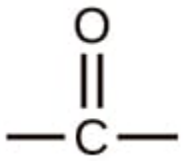
By the end of this section, you will be able to:

- Describe the structure and properties of aldehydes, ketones, carboxylic acids and esters

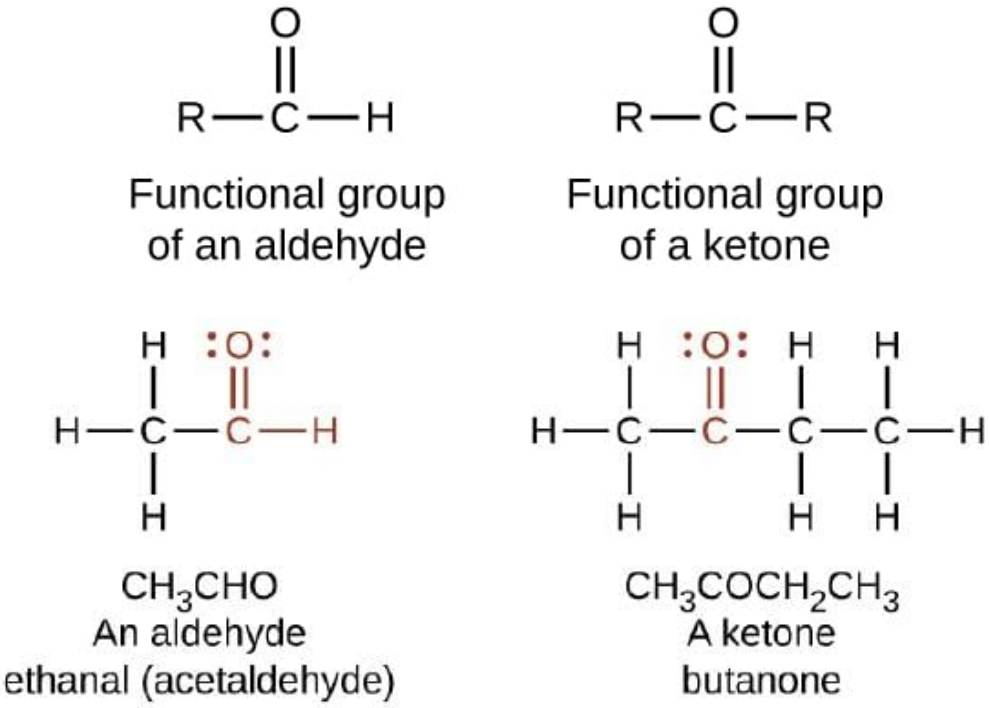
Another class of organic molecules contains a carbon atom connected to an oxygen atom by a double bond, commonly called a carbonyl group. The trigonal planar carbon in the carbonyl group can attach to two other substituents leading to several subfamilies (aldehydes, ketones, carboxylic acids and esters) described in this section.

Aldehydes and Ketones

Both aldehydes and ketones contain a carbonyl group, a functional group with a carbon-oxygen double bond. The names for aldehyde and ketone compounds are derived using similar nomenclature rules as for alkanes and alcohols, and include the class-identifying suffixes *-al* and *-one*, respectively:



In an aldehyde, the carbonyl group is bonded to at least one hydrogen atom. In a ketone, the carbonyl group is bonded to two carbon atoms:



As text, an aldehyde group is represented as –CHO; a ketone is represented as –C(O)– or –CO–.

In both aldehydes and ketones, the geometry around the carbon atom in the carbonyl group is trigonal planar; the carbon atom exhibits sp^2 hybridization. Two of the sp^2 orbitals on the carbon atom in the carbonyl group are used to form σ bonds to the other carbon or hydrogen atoms in a molecule. The remaining sp^2 hybrid orbital forms a σ bond to the oxygen atom. The unhybridized *p* orbital on the carbon atom in the carbonyl group overlaps a *p* orbital on the oxygen atom to form the π bond in the double bond.

Like the

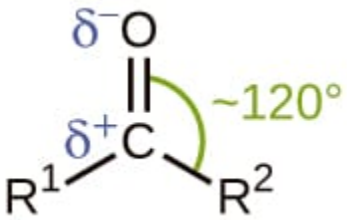
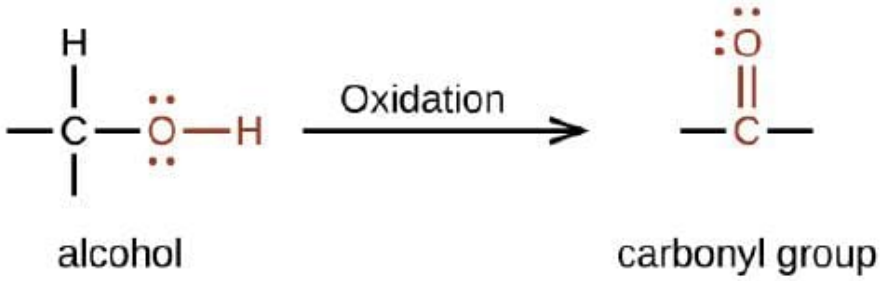


Figure 3.8.4.1: The carbonyl group is polar, and the geometry of the bonds around the central carbon is trigonal planar.

The importance of molecular structure in the reactivity of organic compounds is illustrated by the reactions that produce aldehydes and ketones. We can prepare a carbonyl group by oxidation of an alcohol—for organic molecules, oxidation of a carbon atom is said to occur when a carbon-hydrogen bond is replaced by a carbon-oxygen bond. The reverse reaction—replacing a carbon-oxygen bond by a carbon-hydrogen bond—is a reduction of that carbon atom. Recall that oxygen is generally assigned a –2 oxidation number unless it is elemental or attached to a fluorine. Hydrogen is generally assigned an oxidation number of +1 unless it is attached to a metal. Since carbon does not have a specific rule, its oxidation number is determined algebraically by factoring the atoms it is attached to and the overall charge of the molecule or ion. In general, a carbon atom attached to an oxygen atom will have a more positive oxidation number and a carbon atom attached to a hydrogen atom will have a more negative oxidation number. This should fit nicely with your understanding of the polarity of C–O and C–H bonds. The other reagents and possible products of these reactions are beyond the scope of this chapter, so we will focus only on the changes to the carbon atoms:



Example 3.8.4.1: Oxidation and Reduction in Organic Chemistry

Methane represents the completely reduced form of an organic molecule that contains one carbon atom. Sequentially replacing each of the carbon-hydrogen bonds with a carbon-oxygen bond would lead to an alcohol, then an aldehyde, then a carboxylic acid (discussed later), and, finally, carbon dioxide:

What are the oxidation numbers for the carbon atoms in the molecules shown here?

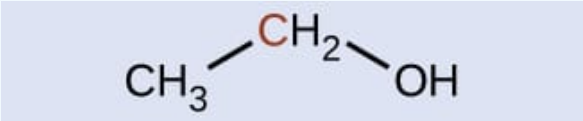
Solution

In this example, we can calculate the oxidation number (review the chapter on oxidation-reduction reactions if necessary) for the carbon atom in each case (note how this would become difficult for larger molecules with additional carbon atoms and hydrogen atoms, which is why organic chemists use the definition dealing with replacing C–H bonds with C–O bonds described). For CH_4, the carbon atom carries a –4 oxidation number (the hydrogen atoms are assigned oxidation numbers of +1 and the carbon atom balances that by having an oxidation number of –4). For the alcohol (in this case, methanol), the carbon atom has an oxidation number of –2 (the oxygen atom is assigned –2, the four hydrogen atoms each are assigned +1, and the carbon atom balances the sum by having an oxidation number of –2; note that compared to the carbon atom in CH_4, this carbon atom has lost two electrons so it was oxidized); for the aldehyde, the carbon atom's oxidation number is 0 (–2 for

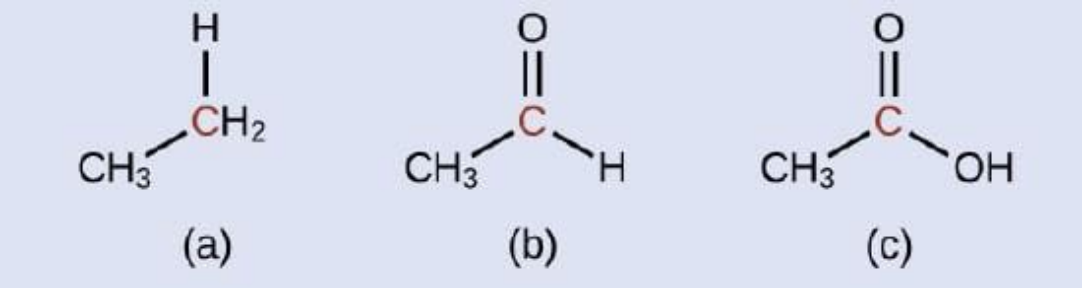
the oxygen atom and +1 for each hydrogen atom already balances to 0, so the oxidation number for the carbon atom is 0); for the carboxylic acid, the carbon atom's oxidation number is +2 (two oxygen atoms each at –2 and two hydrogen atoms at +1); and for carbon dioxide, the carbon atom's oxidation number is +4 (here, the carbon atom needs to balance the –4 sum from the two oxygen atoms).

Exercise 3.8.4.1

Indicate whether the marked carbon atoms in the three molecules here are oxidized or reduced relative to the marked carbon atom in ethanol:

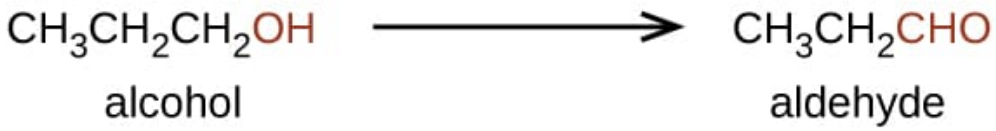


There is no need to calculate oxidation states in this case; instead, just compare the types of atoms bonded to the marked carbon atoms:

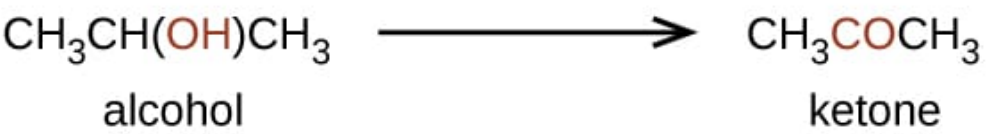


Answer available at the end of this section.

Aldehydes are commonly prepared by the oxidation of alcohols whose –OH functional group is located on the carbon atom at the end of the chain of carbon atoms in the alcohol:



Alcohols that have their –OH groups in the middle of the chain are necessary to synthesize a ketone, which requires the carbonyl group to be bonded to two other carbon atoms:



An alcohol with its –OH group bonded to a carbon atom that is bonded to no or one other carbon atom will form an aldehyde. An alcohol with its –OH group attached to two other carbon atoms will form a ketone. If three carbons are attached to the carbon bonded to the –OH, the molecule will not have a C–H bond to be replaced, so it will not be susceptible to oxidation.

Formaldehyde, an aldehyde with the formula HCHO, is a colorless gas with a pungent and irritating odor. It is sold in an aqueous solution called formalin, which contains about 37% formaldehyde by weight. Formaldehyde causes coagulation of proteins, so

it kills bacteria (and any other living organism) and stops many of the biological processes that cause tissue to decay. Thus, formaldehyde is used for preserving tissue specimens and embalming bodies. It is also used to sterilize soil or other materials. Formaldehyde is used in the manufacture of Bakelite, a hard plastic having high chemical and electrical resistance.

Dimethyl ketone, CH_3COCH_3, commonly called acetone, is the simplest ketone. It is made commercially by fermenting corn or molasses, or by oxidation of 2-propanol. Acetone is a colorless liquid. Among its many uses are as a solvent for lacquer (including fingernail polish), cellulose acetate, cellulose nitrate, acetylene, plastics, and varnishes; as a paint and varnish remover; and as a solvent in the manufacture of pharmaceuticals and chemicals.

Carboxylic Acids and Esters

The odor of vinegar is caused by the presence of acetic acid, a carboxylic acid, in the vinegar. The odor of ripe bananas and many other fruits is due to the presence of esters, compounds that can be prepared by the reaction of a carboxylic acid with an alcohol. Because esters do not have hydrogen bonds between molecules, they have lower vapor pressures than the alcohols and carboxylic acids from which they are derived (see Figure 3.8.4.2).

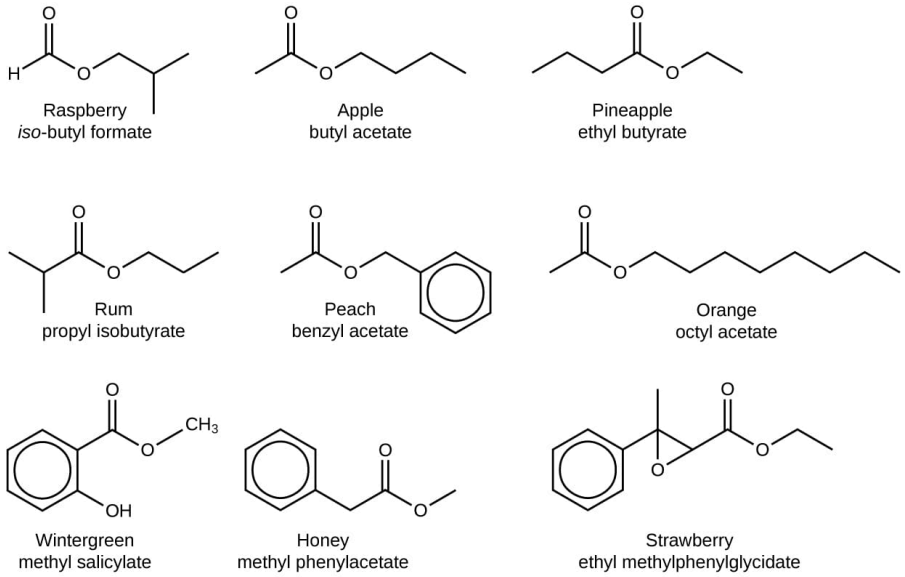
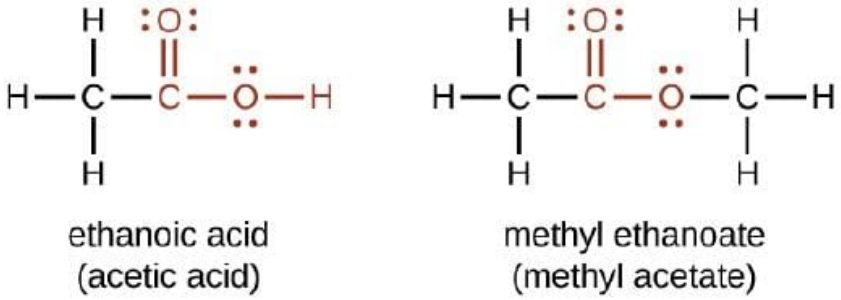


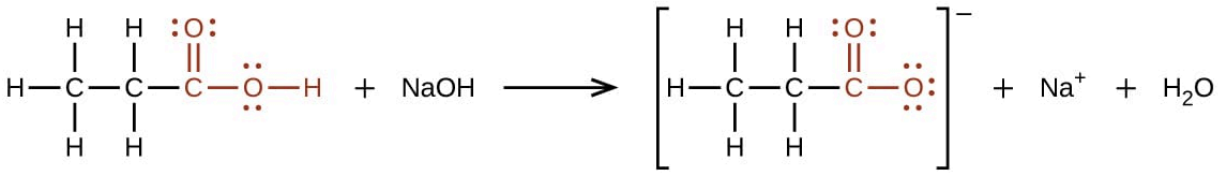
Figure 3.8.4.2: Esters are responsible for the odors associated with various plants and their fruits.

Both carboxylic acids and esters contain a carbonyl group with a second oxygen atom bonded to the carbon atom in the carbonyl group by a single bond. In a carboxylic acid, the second oxygen atom also bonds to a hydrogen atom. In an ester, the second oxygen atom bonds to another carbon atom. The names for carboxylic acids and esters include prefixes that denote the lengths of the carbon chains in the molecules and are derived following nomenclature rules similar to those for inorganic acids and salts (see these examples):



The functional groups for an acid and for an ester are shown in red in these formulas.

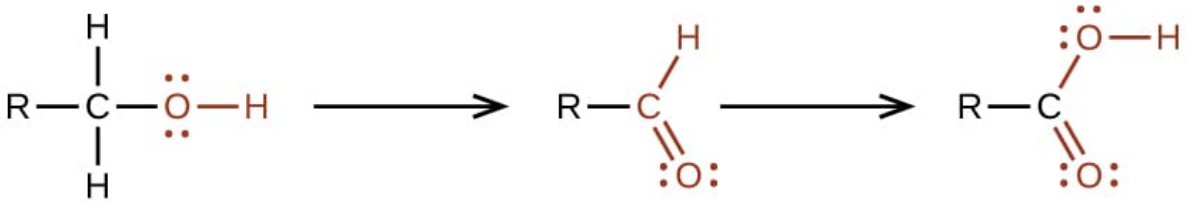
The hydrogen atom in the functional group of a carboxylic acid will react with a base to form an ionic salt:



propionic acid proprionate ion

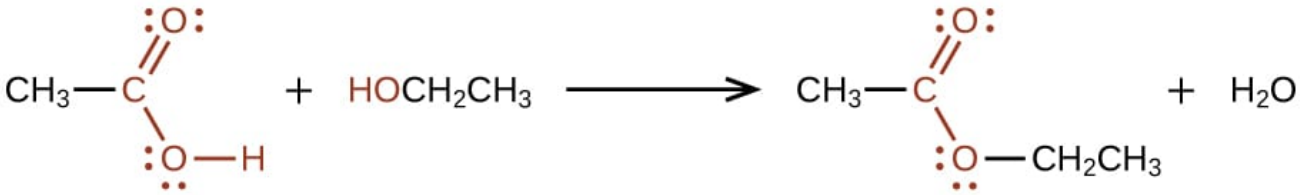
Carboxylic acids are weak acids (see the chapter on acids and bases), meaning they are not 100% ionized in water. Generally only about 1% of the molecules of a carboxylic acid dissolved in water are ionized at any given time. The remaining molecules are undissociated in solution.

We prepare carboxylic acids by the oxidation of aldehydes or alcohols whose –OH functional group is located on the carbon atom at the end of the chain of carbon atoms in the alcohol:



alcohol aldehyde carboxylic acid

Esters are produced by the reaction of acids with alcohols. For example, the ester ethyl acetate, $CH_3CO_2CH_2CH_3$, is formed when acetic acid reacts with ethanol:



acetic acid ethanol ethyl acetate

The simplest carboxylic acid is formic acid, HCO_2H, known since 1670. Its name comes from the Latin word *formicus*, which means “ant”; it was first isolated by the distillation of red ants. It is partially responsible for the pain and irritation of ant and wasp stings, and is responsible for a characteristic odor of ants that can be sometimes detected in their nests.

Acetic acid, CH_3CO_2H, constitutes 3–6% vinegar. Cider vinegar is produced by allowing apple juice to ferment without oxygen present. Yeast cells present in the juice carry out the fermentation reactions. The fermentation reactions change the sugar present in the juice to ethanol, then to acetic acid. Pure acetic acid has a penetrating odor and produces painful burns. It is an excellent solvent for many organic and some inorganic compounds, and it is essential in the production of cellulose acetate, a component of many synthetic fibers such as rayon.

The distinctive and attractive odors and flavors of many flowers, perfumes, and ripe fruits are due to the presence of one or more esters (Figure 3.8.4.3). Among the most important of the natural esters are fats (such as lard, tallow, and butter) and oils (such as linseed, cottonseed, and olive oils), which are esters of the trihydroxyl alcohol glycerine, $C_3H_5(OH)_3$, with large carboxylic acids, such as palmitic acid, $CH_3(CH_2)_{14}CO_2H$, stearic acid, $CH_3(CH_2)_{16}CO_2H$, and oleic acid, Oleic acid is an unsaturated acid; it contains a double bond. Palmitic and stearic acids are saturated acids that contain no double or triple bonds.



Figure 3.8.4.3: Over 350 different volatile molecules (many members of the ester family) have been identified in strawberries. (credit: Rebecca Siegel)

Answers

Exercise 3.8.4.1

(a) reduced (bond to oxygen atom replaced by bond to hydrogen atom); (b) oxidized (one bond to hydrogen atom replaced by one bond to oxygen atom); (c) oxidized (2 bonds to hydrogen atoms have been replaced by bonds to an oxygen atom)

This page titled 3.8.4: Aldehydes, Ketones, Carboxylic Acids, and Esters is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

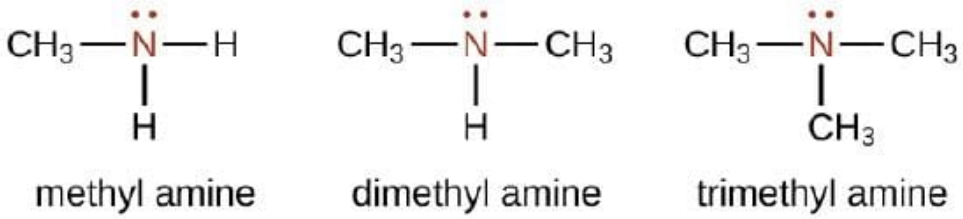
3.8.5: Amines and Amides

Learning Objectives

By the end of this section, you will be able to:

- Describe the structure and properties of an amine
- Describe the structure and properties of an amide

Amines are molecules that contain carbon-nitrogen bonds. The nitrogen atom in an amine has a lone pair of electrons and three bonds to other atoms, either carbon or hydrogen. Various nomenclatures are used to derive names for amines, but all involve the class-identifying suffix *–ine* as illustrated here for a few simple examples:



In some amines, the nitrogen atom replaces a carbon atom in an aromatic hydrocarbon. Pyridine (Figure 3.8.5.1) is one such heterocyclic amine. A heterocyclic compound contains atoms of two or more different elements in its ring structure.

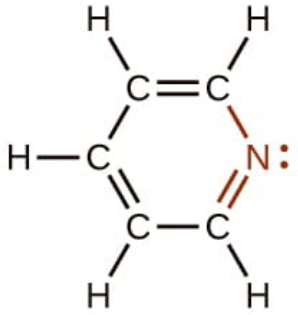


Figure 3.8.5.1: The illustration shows one of the resonance structures of pyridine.

How Sciences Interconnect: DNA in Forensics and Paternity

The genetic material for all living things is a polymer of four different molecules, which are themselves a combination of three subunits. The genetic information, the code for developing an organism, is contained in the specific sequence of the four molecules, similar to the way the letters of the alphabet can be sequenced to form words that convey information. The information in a DNA sequence is used to form two other types of polymers, one of which are proteins. The proteins interact to form a specific type of organism with individual characteristics.

A genetic molecule is called DNA, which stands for deoxyribonucleic acid. The four molecules that make up DNA are called nucleotides. Each nucleotide consists of a single- or double-ringed molecule containing nitrogen, carbon, oxygen, and hydrogen called a nitrogenous base. Each base is bonded to a five-carbon sugar called deoxyribose. The sugar is in turn bonded to a phosphate group PO43–)Figure 3.8.5.2).

It probably makes sense that the sequence of nucleotides in the DNA of a cat differs from those of a dog. But it is also true that the sequences of the DNA in the cells of two individual pugs differ. Likewise, the sequences of DNA in you and a sibling differ (unless your sibling is an identical twin), as do those between you and an unrelated individual. However, the DNA

sequences of two related individuals are more similar than the sequences of two unrelated individuals, and these similarities in sequence can be observed in various ways. This is the principle behind DNA fingerprinting, which is a method used to determine whether two DNA samples came from related (or the same) individuals or unrelated individuals.

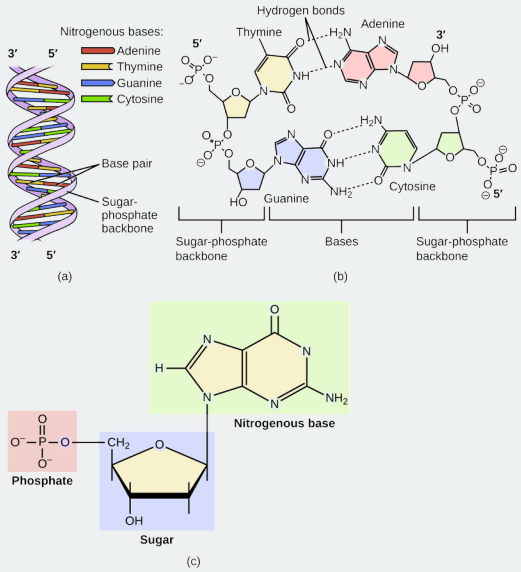


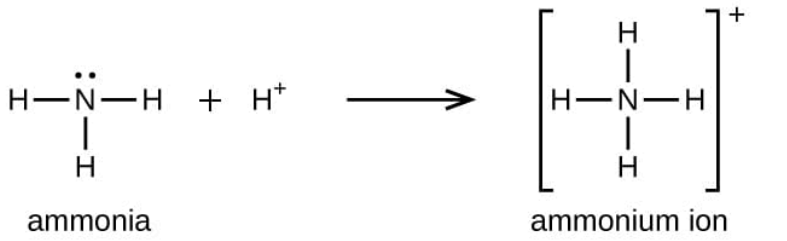
Figure 3.8.5.2: DNA is an organic molecule and the genetic material for all living organisms. (a) DNA is a double helix consisting of two single DNA strands hydrogen bonded together at each nitrogenous base. (b) This detail shows the hydrogen bonding (dotted lines) between nitrogenous bases on each DNA strand and the way in which each nucleotide is joined to the next, forming a backbone of sugars and phosphate groups along each strand. (c) This detail shows the structure of one of the four nucleotides that makes up the DNA polymer. Each nucleotide consists of a nitrogenous base (a double-ring molecule, in this case), a five-carbon sugar (deoxyribose), and a phosphate group.

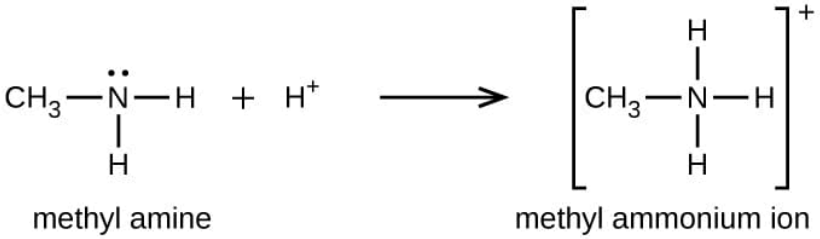
Using similarities in sequences, technicians can determine whether a man is the father of a child (the identity of the mother is rarely in doubt, except in the case of an adopted child and a potential birth mother). Likewise, forensic geneticists can determine whether a crime scene sample of human tissue, such as blood or skin cells, contains DNA that matches exactly the DNA of a suspect.

Link to Learning

Watch this video animation of how DNA is packaged for a visual lesson in its structure.

Like ammonia, amines are weak bases due to the lone pair of electrons on their nitrogen atoms:





The basicity of an amine's nitrogen atom plays an important role in much of the compound's chemistry. Amine functional groups are found in a wide variety of compounds, including natural and synthetic dyes, polymers, vitamins, and medications such as penicillin and codeine. They are also found in many molecules essential to life, such as amino acids, hormones, neurotransmitters, and DNA.

How Sciences Interconnect: Addictive Alkaloids

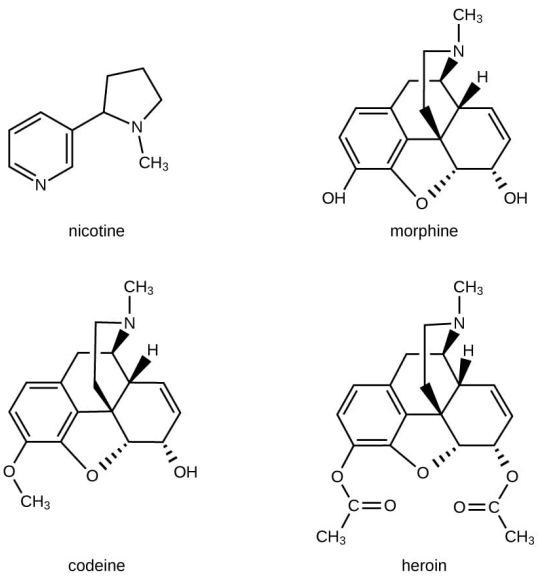
Since ancient times, plants have been used for medicinal purposes. One class of substances, called *alkaloids*, found in many of these plants has been isolated and found to contain cyclic molecules with an amine functional group. These amines are bases. They can react with H_3O^+ in a dilute acid to form an ammonium salt, and this property is used to extract them from the plant:

R3N+H3O++Cl–⟡[R3NH+]Cl–+H2OR3N+H3O++Cl–⟡[R3NH+]Cl–+H2O

The name alkaloid means "like an alkali." Thus, an alkaloid reacts with acid. The free compound can be recovered after extraction by reaction with a base:

R3NH+]Cl–+OH–⟡R3N+H2O+Cl–[R3NH+]Cl–+OH–⟡R3N+H2O+Cl–

The structures of many naturally occurring alkaloids have profound physiological and psychotropic effects in humans. Examples of these drugs include nicotine, morphine, codeine, and heroin. The plant produces these substances, collectively called secondary plant compounds, as chemical defenses against the numerous pests that attempt to feed on the plant:

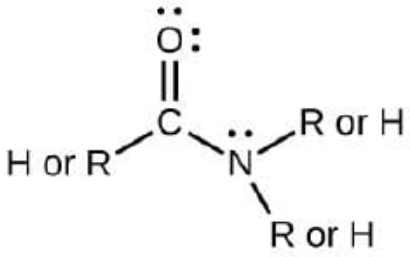


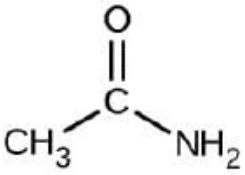
In these diagrams, as is common in representing structures of large organic compounds, carbon atoms in the rings and the hydrogen atoms bonded to them have been omitted for clarity. The solid wedges indicate bonds that extend out of the page. The dashed wedges indicate bonds that extend into the page. Notice that small changes to a part of the molecule change the properties of morphine, codeine, and heroin. Morphine, a strong narcotic used to relieve pain, contains two hydroxyl functional groups, located at the bottom of the molecule in this structural formula. Changing one of these hydroxyl groups to a methyl ether group forms codeine, a less potent drug used as a local anesthetic. If both hydroxyl groups are converted to esters of acetic acid, the powerfully addictive drug heroin results (Figure 3.8.5.3).



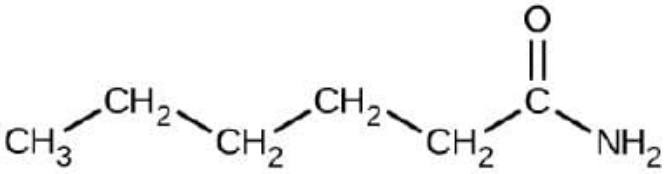
Figure 3.8.5.3: Poppies can be used in the production of opium, a plant latex that contains morphine from which other opiates, such as heroin, can be synthesized. (credit: Karen Roe)

Amides are molecules that contain nitrogen atoms connected to the carbon atom of a carbonyl group. Like amines, various nomenclature rules may be used to name amides, but all include use of the class-specific suffix *-amide*:



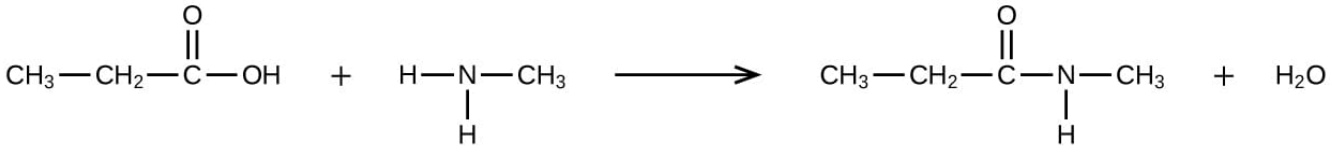


acetamide



hexanamide

Amides can be produced when carboxylic acids react with amines or ammonia in a process called amidation. A water molecule is eliminated from the reaction, and the amide is formed from the remaining pieces of the carboxylic acid and the amine (note the similarity to formation of an ester from a carboxylic acid and an alcohol discussed in the previous section):



carboxylic acid amine amide water

The reaction between amines and carboxylic acids to form amides is biologically important. It is through this reaction that amino acids (molecules containing both amine and carboxylic acid substituents) link together in a polymer to form proteins.

How Sciences Interconnect: Proteins and Enzymes

Proteins are large biological molecules made up of long chains of smaller molecules called amino acids. Organisms rely on proteins for a variety of functions—proteins transport molecules across cell membranes, replicate DNA, and catalyze metabolic reactions, to name only a few of their functions. The properties of proteins are functions of the combination of amino acids that compose them and can vary greatly. Interactions between amino acid sequences in the chains of proteins result in the folding of the chain into specific, three-dimensional structures that determine the protein's activity.

Amino acids are organic molecules that contain an amine functional group ($-NH_2$), a carboxylic acid functional group (–COOH), and a side chain (that is specific to each individual amino acid). Most living things build proteins from the same 20 different amino acids. Amino acids connect by the formation of a peptide bond, which is a covalent bond formed between two amino acids when the carboxylic acid group of one amino acid reacts with the amine group of the other amino acid. The formation of the bond results in the production of a molecule of water (in general, reactions that result in the production of water when two other molecules combine are referred to as condensation reactions). The resulting bond—between the carbonyl group carbon atom and the amine nitrogen atom is called a peptide link or peptide bond. Since each of the original amino acids has an unreacted group (one has an unreacted amine and the other an unreacted carboxylic acid), more peptide bonds can form to other amino acids, extending the structure. (Figure 3.8.5.4) A chain of connected amino acids is called a polypeptide. Proteins contain at least one long polypeptide chain.

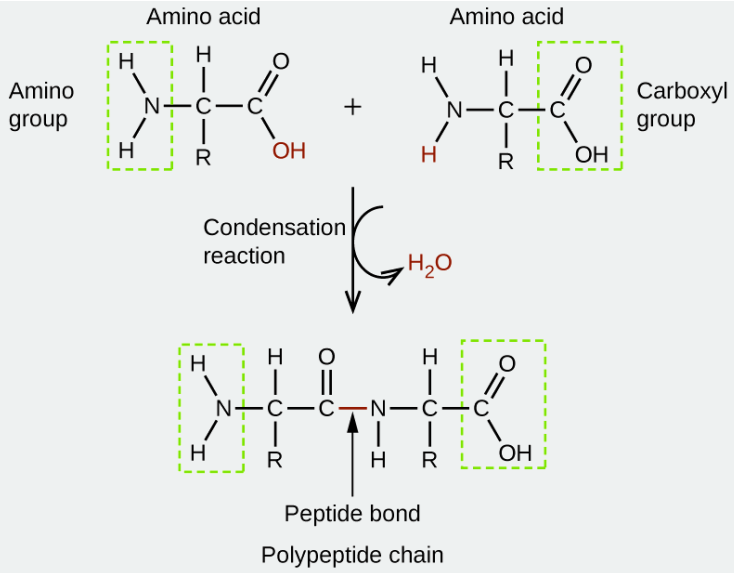


Figure 3.8.5.4: This condensation reaction forms a dipeptide from two amino acids and leads to the formation of water.

Enzymes are large biological molecules, mostly composed of proteins, which are responsible for the thousands of metabolic processes that occur in living organisms. Enzymes are highly specific catalysts; they speed up the rates of certain reactions. Enzymes function by lowering the activation energy of the reaction they are catalyzing, which can dramatically increase the rate of the reaction. Most reactions catalyzed by enzymes have rates that are millions of times faster than the noncatalyzed version. Like all catalysts, enzymes are not consumed during the reactions that they catalyze. Enzymes do differ from other catalysts in how specific they are for their substrates (the molecules that an enzyme will convert into a different product). Each enzyme is only capable of speeding up one or a few very specific reactions or types of reactions. Since the function of enzymes is so specific, the lack or malfunctioning of an enzyme can lead to serious health consequences. One disease that is the result of an enzyme malfunction is phenylketonuria. In this disease, the enzyme that catalyzes the first step in the degradation of the amino acid phenylalanine is not functional (Figure 3.8.5.5). Untreated, this can lead to an accumulation of phenylalanine, which can lead to intellectual disabilities.

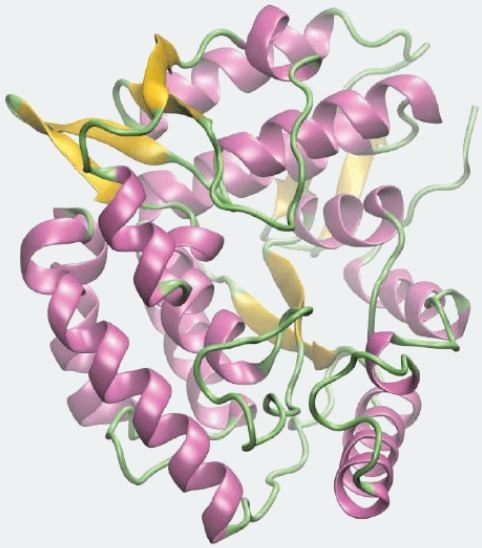


Figure 3.8.5.5: A computer rendering shows the three-dimensional structure of the enzyme phenylalanine hydroxylase. In the disease phenylketonuria, a defect in the shape of phenylalanine hydroxylase causes it to lose its function in breaking down phenylalanine.

Chemistry in Everyday Life: Kevlar

Kevlar (Figure 3.8.5.6) is a synthetic polymer made from two monomers 1,4-phenylene-diamine and terephthaloyl chloride (Kevlar is a registered trademark of DuPont). The material was developed by Susan Kwolek while she worked to find a replacement for steel in tires. Kwolek's work involved synthesizing polyamides and dissolving them in solvents, then spinning the resulting solution into fibers. One of her solutions proved to be quite different in initial appearance and structure. And once spun, the resulting fibers were particularly strong. From this initial discovery, Kevlar was created. The material has a high tensile strength-to-weight ratio (it is about 5 times stronger than an equal weight of steel), making it useful for many applications from bicycle tires to sails to body armor.

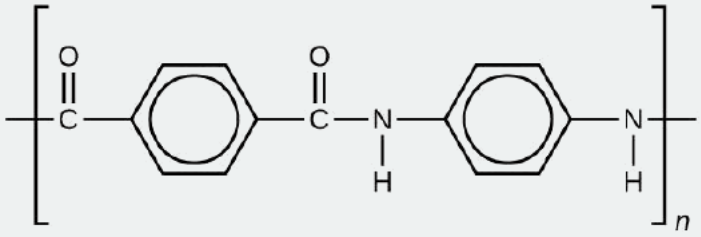


Figure 3.8.5.6: This illustration shows the formula for polymeric Kevlar.

The material owes much of its strength to hydrogen bonds between polymer chains (refer back to the chapter on intermolecular interactions). These bonds form between the carbonyl group oxygen atom (which has a partial negative charge due to oxygen's electronegativity) on one monomer and the partially positively charged hydrogen atom in the N–H bond of an adjacent monomer in the polymer structure (see dashed lines in Figure 3.8.5.7). There is additional strength derived from the interaction between the unhybridized *p* orbitals in the six-membered rings, called aromatic stacking.

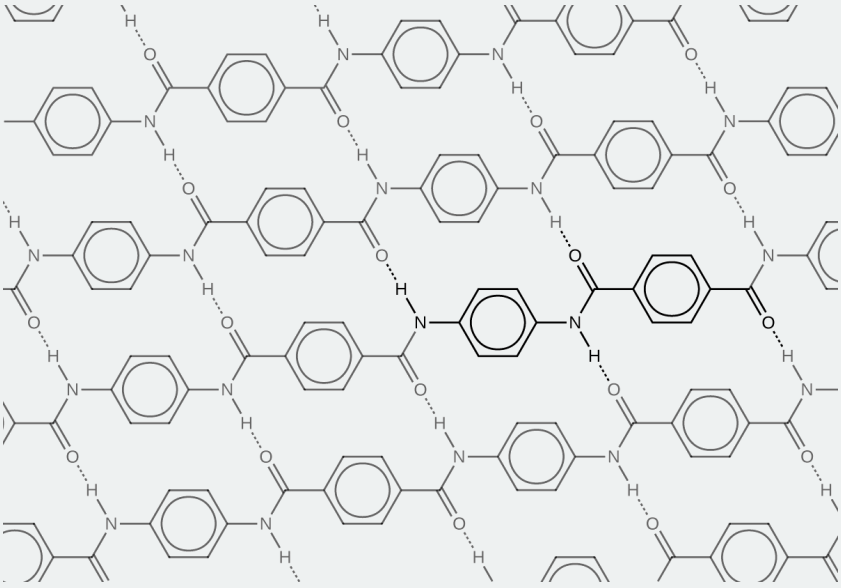


Figure 3.8.5.7: The diagram shows the polymer structure of Kevlar, with hydrogen bonds between polymer chains represented by dotted lines.

Kevlar may be best known as a component of body armor, combat helmets, and face masks. Since the 1980s, the US military has used Kevlar as a component of the PASGT (personal armor system for ground troops) helmet and vest. Kevlar is also used to protect armored fighting vehicles and aircraft carriers. Civilian applications include protective gear for emergency service personnel such as body armor for police officers and heat-resistant clothing for fire fighters. Kevlar based clothing is considerably lighter and thinner than equivalent gear made from other materials (Figure 3.8.5.8). Beyond Kevlar, Susan Kwolek was instrumental in the development of Nomex, a fireproof material, and was also involved in the creation of Lycra. She became just the fourth woman inducted into the National Inventors Hall of Fame, and received a number of other awards for her significant contributions to science and society.



(a)



(b)



(c)

Figure 3.8.5.8: (a) These soldiers are sorting through pieces of a Kevlar helmet that helped absorb a grenade blast. Kevlar is also used to make (b) canoes and (c) marine mooring lines. (credit a: modification of work by “Cla68”/Wikimedia Commons; credit b: modification of work by “OakleyOriginals”/Flickr; credit c: modification of work by Casey H. Kyhl)

In addition to its better-known uses, Kevlar is also often used in cryogenics for its very low thermal conductivity (along with its high strength). Kevlar maintains its high strength when cooled to the temperature of liquid nitrogen (–196 °C).

The table here summarizes the structures discussed in this chapter:

Compound Name	Structure of Compound and Functional Group (red)	Example	
		Formula	Name
alkene	C=C	C_2H_4	ethene
alkyne	C≡C	C_2H_2	ethyne
alcohol	R—Ö—H	CH_3CH_2OH	ethanol
ether	R—Ö—R′	$(C_2H_5)_2O$	diethyl ether
aldehyde	R—C(=O)—H	CH_3CHO	ethanal
ketone	R—C(=O)—R′	$CH_3COCH_2CH_3$	methyl ethyl ketone
carboxylic acid	R—C(=O)—Ö—H	CH_3COOH	acetic acid
ester	R—C(=O)—Ö—R′	$CH_3CO_2CH_2CH_3$	ethyl acetate
amine	R—N(—H)—H; R—N(—R′)—H; R—N(—R′)—R″	$C_2H_5NH_2$	ethylamine
amide	R—C(=O)—N(—H)—R′	CH_3CONH_2	acetamide

This page titled 3.8.5: Amines and Amides is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

3.8.6: Key Terms

Example and Directions

Words (or words that have the same definition)	The definition is case sensitive	(Optional) Image to display with the definition [Not displayed in Glossary, only in pop-up on pages]	(Optional) Caption for Image	(Optional) External or Internal Link	(Optional) Source for Definition
(Eg. "Genetic, Hereditary, DNA ...")	(Eg. "Relating to genes or heredity")		The infamous double helix	https://bio.libretexts.org/	CC-BY-SA; Delmar Larsen

Glossary Entries

Word(s)	Definition	Image	Caption	Link	Source
addition reaction	reaction in which a double carbon-carbon bond forms a single carbon-carbon bond by the addition of a reactant. Typical reaction for an alkene.				
alcohol	organic compound with a hydroxyl group (–OH) bonded to a carbon atom				
aldehyde	organic compound containing a carbonyl group bonded to two hydrogen atoms or a hydrogen atom and a carbon substituent				
alkane	molecule consisting of only carbon and hydrogen atoms connected by single (σ) bonds				
alkene	molecule consisting of carbon and hydrogen containing at least one carbon-carbon double bond				
alkyl group	substituent, consisting of an alkane missing one hydrogen atom, attached to a larger structure				
alkyne	molecule consisting of carbon and hydrogen containing at least one carbon-carbon triple bond				
amide	organic molecule that features a nitrogen atom connected to the carbon atom in a carbonyl group				
amine	organic molecule in which a nitrogen atom is bonded to one or more alkyl group				
aromatic hydrocarbon	cyclic molecule consisting of carbon and hydrogen with delocalized alternating carbon-carbon single and double bonds, resulting in enhanced stability				
carbonyl group	carbon atom double bonded to an oxygen atom				
carboxylic acid	organic compound containing a carbonyl group with an attached hydroxyl group				
ester	organic compound containing a carbonyl group with an attached oxygen atom that is bonded to a carbon substituent				
ether	organic compound with an oxygen atom that is bonded to two carbon atoms				
functional group	part of an organic molecule that imparts a specific chemical reactivity to the molecule				
ketone	organic compound containing a carbonyl group with two carbon substituents attached to it				

Word(s)	Definition	Image	Caption	Link	Source
organic compound	natural or synthetic compound that contains carbon				
saturated hydrocarbon	molecule containing carbon and hydrogen that has only single bonds between carbon atoms				
skeletal structure	shorthand method of drawing organic molecules in which carbon atoms are represented by the ends of lines and bends in between lines, and hydrogen atoms attached to the carbon atoms are not shown (but are understood to be present by the context of the structure)				
substituent	branch or functional group that replaces hydrogen atoms in a larger hydrocarbon chain				
substitution reaction	reaction in which one atom replaces another in a molecule				

This page titled 3.8.6: Key Terms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

3.8.7: Summary

Strong, stable bonds between carbon atoms produce complex molecules containing chains, branches, and rings. The chemistry of these compounds is called organic chemistry. Hydrocarbons are organic compounds composed of only carbon and hydrogen. The alkanes are saturated hydrocarbons—that is, hydrocarbons that contain only single bonds. Alkenes contain one or more carbon-carbon double bonds. Alkynes contain one or more carbon-carbon triple bonds. Aromatic hydrocarbons contain ring structures with delocalized π electron systems.

Many organic compounds that are not hydrocarbons can be thought of as derivatives of hydrocarbons. A hydrocarbon derivative can be formed by replacing one or more hydrogen atoms of a hydrocarbon by a functional group, which contains at least one atom of an element other than carbon or hydrogen. The properties of hydrocarbon derivatives are determined largely by the functional group. The –OH group is the functional group of an alcohol. The –R–O–R– group is the functional group of an ether.

Functional groups related to the carbonyl group include the –CHO group of an aldehyde, the –CO– group of a ketone, the $-CO_2H$ group of a carboxylic acid, and the $-CO_2R$ group of an ester. The carbonyl group, a carbon-oxygen double bond, is the key structure in these classes of organic molecules: Aldehydes contain at least one hydrogen atom attached to the carbonyl carbon atom, ketones contain two carbon groups attached to the carbonyl carbon atom, carboxylic acids contain a hydroxyl group attached to the carbonyl carbon atom, and esters contain an oxygen atom attached to another carbon group connected to the carbonyl carbon atom. All of these compounds contain oxidized carbon atoms relative to the carbon atom of an alcohol group.

The addition of nitrogen into an organic framework leads to two families of molecules. Compounds containing a nitrogen atom bonded in a hydrocarbon framework are classified as amines. Compounds that have a nitrogen atom bonded to one side of a carbonyl group are classified as amides. Amines are a basic functional group. Amines and carboxylic acids can combine in a condensation reaction to form amides.

This page titled 3.8.7: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

3.8.8: Exercises

1.

Write the chemical formula and Lewis structure of the following, each of which contains five carbon atoms:

a. an alkane
b. an alkene
c. an alkyne

2.

What is the difference between the hybridization of carbon atoms' valence orbitals in saturated and unsaturated hydrocarbons?

3.

On a microscopic level, how does the reaction of bromine with a saturated hydrocarbon differ from its reaction with an unsaturated hydrocarbon? How are they similar?

4.

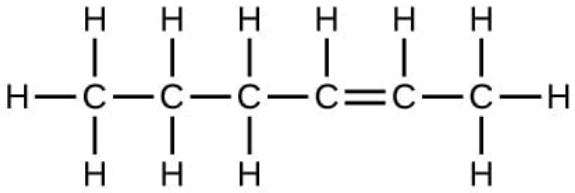
On a microscopic level, how does the reaction of bromine with an alkene differ from its reaction with an alkyne? How are they similar?

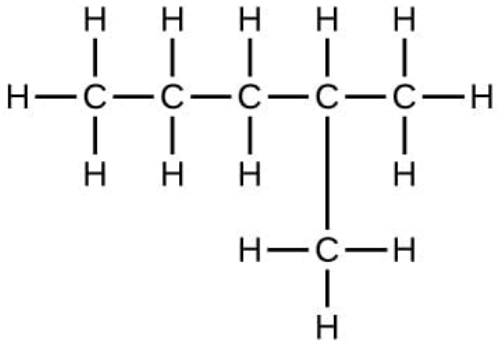
5.

Explain why unbranched alkenes can form geometric isomers while unbranched alkanes cannot. Does this explanation involve the macroscopic domain or the microscopic domain?

6.

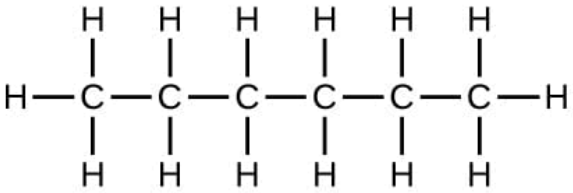
Explain why these two molecules are not isomers:

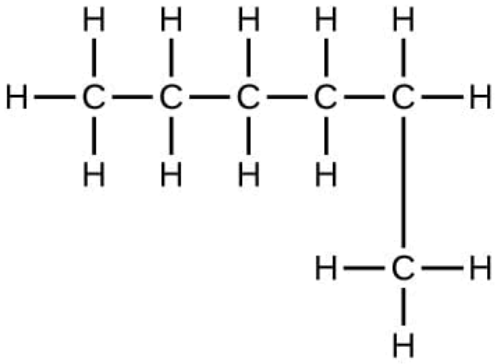




7.

Explain why these two molecules are not isomers:





8.

How does the carbon-atom hybridization change when polyethylene is prepared from ethylene?

9.

Write the Lewis structure and molecular formula for each of the following hydrocarbons:

a. (f) 4-methyl-2-pentyne

10.

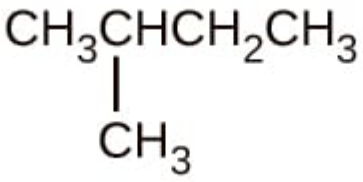
Write the chemical formula, condensed formula, and Lewis structure for each of the following hydrocarbons:

a. (f) 3,4-dimethyl-1-pentyne

11.

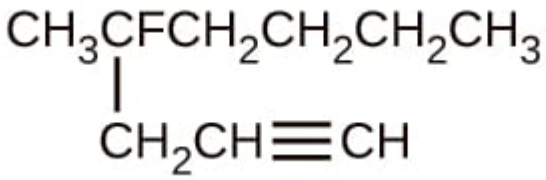
Give the complete IUPAC name for each of the following compounds:

a. (c)

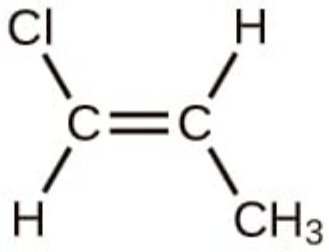


(d)

(e)



(f)

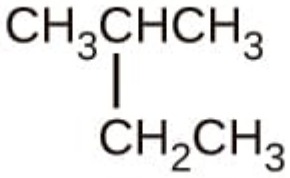


(g)

12.

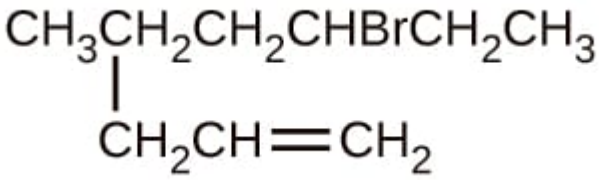
Give the complete IUPAC name for each of the following compounds:

a. (c)



(d)

(e)



(f)

13.

Butane is used as a fuel in disposable lighters. Write the Lewis structure for each isomer of butane.

14.

Write Lewis structures and name the five structural isomers of hexane.

15.

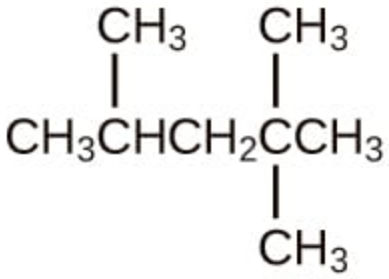
Write Lewis structures for the *cis*–*trans* isomers of

16.

Write structures for the three isomers of the aromatic hydrocarbon xylene, $C_6H_4(CH_3)_2$.

17.

Isooctane is the common name of the isomer of C_8H_{18} used as the standard of 100 for the gasoline octane rating:



a. What is the IUPAC name for the compound?
b. Name the other isomers that contain a five-carbon chain with three methyl substituents.

18.

Write Lewis structures and IUPAC names for the alkyne isomers of C_4H_6.

19.

Write Lewis structures and IUPAC names for all isomers of C_4H_9Cl.

20.

Name and write the structures of all isomers of the propyl and butyl alkyl groups.

21.

Write the structures for all the isomers of the $-C_5H_{11}$ alkyl group.

22.

Write Lewis structures and describe the molecular geometry at each carbon atom in the following compounds:

a. *cis*-3-hexene
b. *cis*-1-chloro-2-bromoethene
c. 2-pentyne
d. *trans*-6-ethyl-7-methyl-2-octene

23.

Benzene is one of the compounds used as an octane enhancer in unleaded gasoline. It is manufactured by the catalytic conversion of acetylene to benzene:

Draw Lewis structures for these compounds, with resonance structures as appropriate, and determine the hybridization of the carbon atoms in each.

24.

Teflon is prepared by the polymerization of tetrafluoroethylene. Write the equation that describes the polymerization using Lewis symbols.

25.

Write two complete, balanced equations for each of the following reactions, one using condensed formulas and one using Lewis structures.

a. 1 mol of 1-butyne reacts with 2 mol of iodine.
b. Pentane is burned in air.

26.

Write two complete, balanced equations for each of the following reactions, one using condensed formulas and one using Lewis structures.

a. 2-butene reacts with chlorine.
b. benzene burns in air.

27.

What mass of 2-bromopropane could be prepared from 25.5 g of propene? Assume a 100% yield of product.

28.

Acetylene is a very weak acid; however, it will react with moist silver(I) oxide and form water and a compound composed of silver and carbon. Addition of a solution of HCl to a 0.2352-g sample of the compound of silver and carbon produced acetylene and 0.2822 g of AgCl.

a. What is the empirical formula of the compound of silver and carbon?
b. The production of acetylene on addition of HCl to the compound of silver and carbon suggests that the carbon is present as the acetylide ion, . Write the formula of the compound showing the acetylide ion.

29.

Ethylene can be produced by the pyrolysis of ethane:

How many kilograms of ethylene is produced by the pyrolysis of $1.000\ 10^3$ kg of ethane, assuming a 100.0% yield?

30.

Why do the compounds hexane, hexanol, and hexene have such similar names?

31.

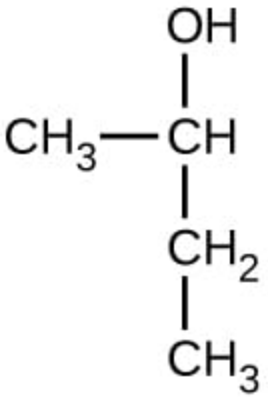
Write condensed formulas and provide IUPAC names for the following compounds:

a. ethyl alcohol (in beverages)
b. methyl alcohol (used as a solvent, for example, in shellac)
c. ethylene glycol (antifreeze)
d. isopropyl alcohol (used in rubbing alcohol)
e. glycerine

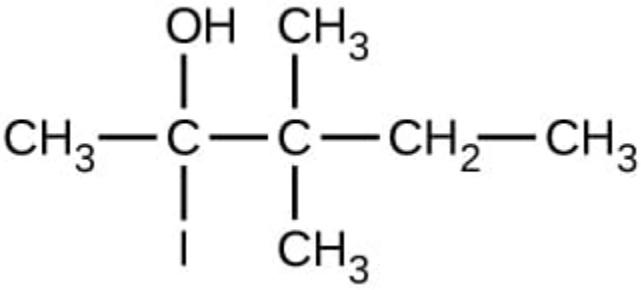
32.

Give the complete IUPAC name for each of the following compounds:

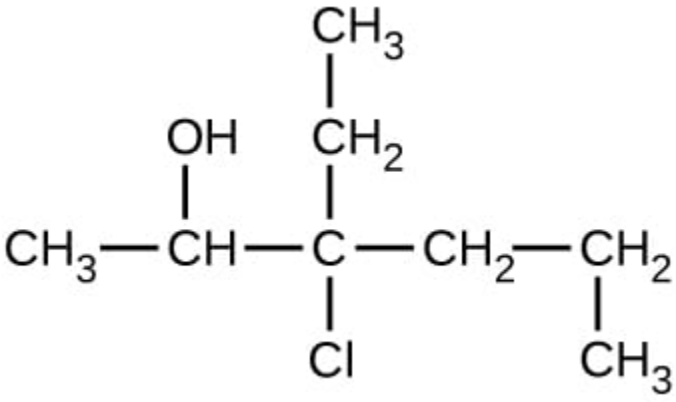
(a)



(b)



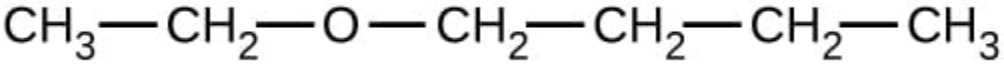
(c)



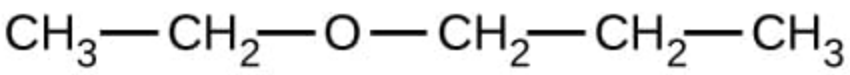
33.

Give the complete IUPAC name and the common name for each of the following compounds:

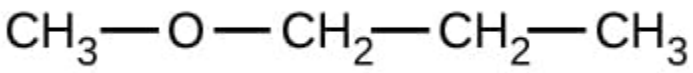
(a)



(b)



(c)



34.

Write the condensed structures of both isomers with the formula C_2H_6O. Label the functional group of each isomer.

35.

Write the condensed structures of all isomers with the formula $C_2H_6O_2$. Label the functional group (or groups) of each isomer.

36.

Draw the condensed formulas for each of the following compounds:

a. dipropyl ether
b. 2,2-dimethyl-3-hexanol
c. 2-ethoxybutane

37.

MTBE, Methyl *tert*-butyl ether, $CH_3OC(CH_3)_3$, is used as an oxygen source in oxygenated gasolines. MTBE is manufactured by reacting 2-methylpropene with methanol.

a. Using Lewis structures, write the chemical equation representing the reaction.
b. What volume of methanol, density 0.7915 g/mL, is required to produce exactly 1000 kg of MTBE, assuming a 100% yield?

38.

Write two complete balanced equations for each of the following reactions, one using condensed formulas and one using Lewis structures.

a. propanol is converted to dipropyl ether
b. propene is treated with water in dilute acid.

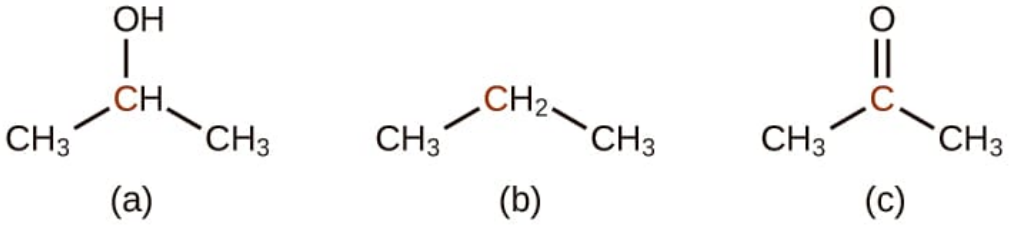
39.

Write two complete balanced equations for each of the following reactions, one using condensed formulas and one using Lewis structures.

a. 2-butene is treated with water in dilute acid
b. ethanol is dehydrated to yield ethene

40.

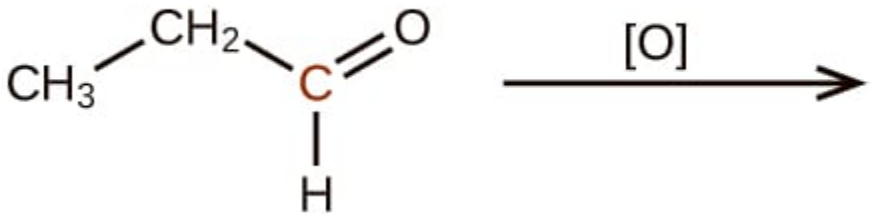
Order the following molecules from least to most oxidized, based on the marked carbon atom:



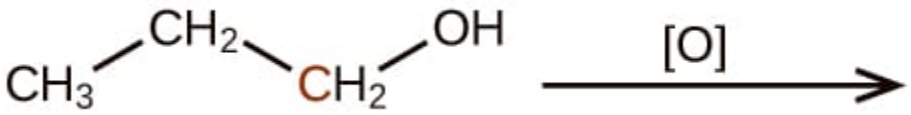
41.

Predict the products of oxidizing the molecules shown in this problem. In each case, identify the product that will result from the minimal increase in oxidation state for the highlighted carbon atom:

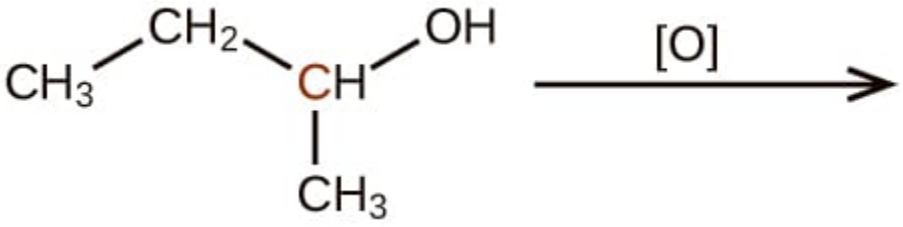
(a)



(b)



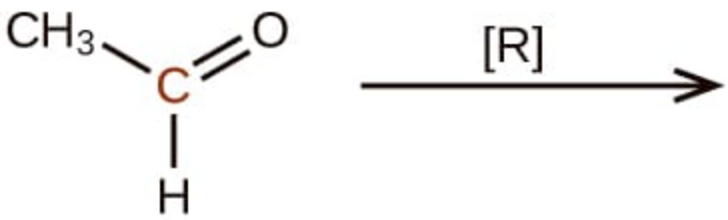
(c)



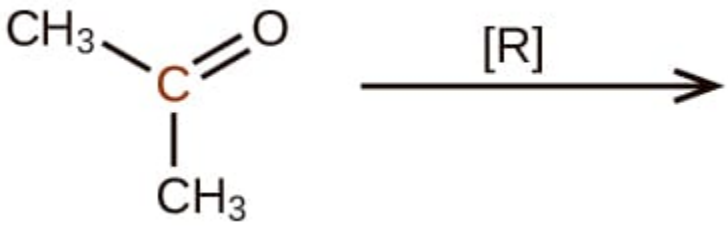
42.

Predict the products of reducing the following molecules. In each case, identify the product that will result from the minimal decrease in oxidation state for the highlighted carbon atom:

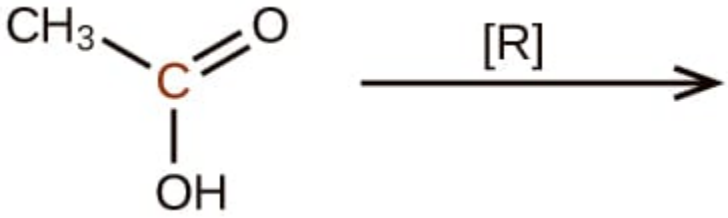
(a)



(b)



(c)



43.

Explain why it is not possible to prepare a ketone that contains only two carbon atoms.

44.

How does hybridization of the substituted carbon atom change when an alcohol is converted into an aldehyde? An aldehyde to a carboxylic acid?

45.

Fatty acids are carboxylic acids that have long hydrocarbon chains attached to a carboxylate group. How does a saturated fatty acid differ from an unsaturated fatty acid? How are they similar?

46.

Write a condensed structural formula, such as CH_3CH_3, and describe the molecular geometry at each carbon atom.

a. (f) formaldehyde

47.

Write a condensed structural formula, such as CH_3CH_3, and describe the molecular geometry at each carbon atom.

a. 2-propanol
b. acetone
c. dimethyl ether
d. acetic acid

e. 3-methyl-1-hexene

48.

The foul odor of rancid butter is caused by butyric acid, $CH_3CH_2CH_2CO_2H$.

a. Draw the Lewis structure and determine the oxidation number and hybridization for each carbon atom in the molecule.
b. The esters formed from butyric acid are pleasant-smelling compounds found in fruits and used in perfumes. Draw the Lewis structure for the ester formed from the reaction of butyric acid with 2-propanol.

49.

Write the two-resonance structures for the acetate ion.

50.

Write two complete, balanced equations for each of the following reactions, one using condensed formulas and one using Lewis structures:

a. ethanol reacts with propionic acid
b. benzoic acid, $C_6H_5CO_2H$, is added to a solution of sodium hydroxide

51.

Write two complete balanced equations for each of the following reactions, one using condensed formulas and one using Lewis structures.

a. 1-butanol reacts with acetic acid
b. propionic acid is poured onto solid calcium carbonate

52.

Yields in organic reactions are sometimes low. What is the percent yield of a process that produces 13.0 g of ethyl acetate from 10.0 g of CH_3CO_2H?

53.

Alcohols A, B, and C all have the composition $C_4H_{10}O$. Molecules of alcohol A contain a branched carbon chain and can be oxidized to an aldehyde; molecules of alcohol B contain a linear carbon chain and can be oxidized to a ketone; and molecules of alcohol C can be oxidized to neither an aldehyde nor a ketone. Write the Lewis structures of these molecules.

54.

Write the Lewis structures of both isomers with the formula C_2H_7N.

55.

What is the molecular structure about the nitrogen atom in trimethyl amine and in the trimethyl ammonium ion, $(CH_3)_3NH^+$? What is the hybridization of the nitrogen atom in trimethyl amine and in the trimethyl ammonium ion?

56.

Write the two resonance structures for the pyridinium ion, $C_5H_5NH^+$.

57.

Draw Lewis structures for pyridine and its conjugate acid, the pyridinium ion, $C_5H_5NH^+$. What are the hybridizations, electron domain geometries, and molecular geometries about the nitrogen atoms in pyridine and in the pyridinium ion?

58.

Write the Lewis structures of all isomers with the formula C_3H_7ON that contain an amide linkage.

59.

Write two complete balanced equations for the following reaction, one using condensed formulas and one using Lewis structures.

Methyl amine is added to a solution of HCl.

60.

Write two complete, balanced equations for each of the following reactions, one using condensed formulas and one using Lewis structures.

Ethylammonium chloride is added to a solution of sodium hydroxide.

61.

Identify any carbon atoms that change hybridization and the change in hybridization during the reactions in Exercise 20.26.

62.

Identify any carbon atoms that change hybridization and the change in hybridization during the reactions in Exercise 20.39.

63.

Identify any carbon atoms that change hybridization and the change in hybridization during the reactions in Exercise 20.51.

This page titled 3.8.8: Exercises is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 3.9: NUCLEAR CHEMISTRY

3.9.1: Introduction to nuclear chemistry
3.9.2: Radioactivity
3.9.3: Half-life of radioisotopes
3.9.4: Radiation measurements
3.9.5: Ionizing radiation exposures
3.9.6: Medical uses of radioisotopes
3.9.7: Making radioisotopes for medical uses
3.9.8: Nuclear fusion and fission

This page titled Chapter 3.9: Nuclear chemistry is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.9.1: Introduction to nuclear chemistry

What is a nuclear reaction?

Unlike chemical reactions that involve valence electrons, **nuclear reactions involve changes in the nucleus of an atom,** as shown in Fig. 8.1.1.

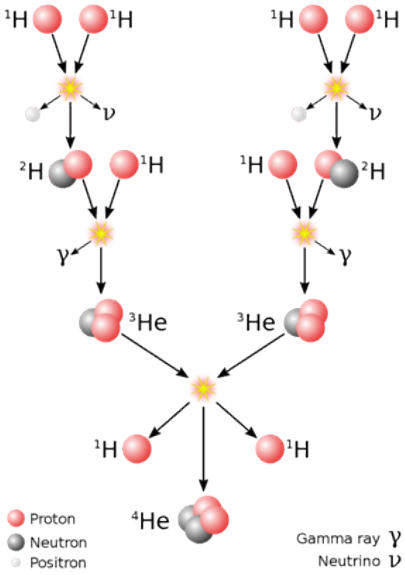


Figure 3.9.1.1: The proton-proton chain reaction, branch I, dominates in stars the size of the Sun or smaller. Source: Sarang, Public domain, via Wikimedia Commons

Nuclear reaction

Nuclear reactions involve changes in the nucleus of an atom. A nuclear reaction may result in one or more of the following: i) conversion of an atom to its isotope or an atom of another element, ii) conversion of mass into energy or vice versa, and iii) release of nuclear radiations.

Although nuclear reactions are less numerous than chemical reactions, they are essential in many aspects, e.g., they are the source of energy in the sun and stars and the synthesis of elements in the universe. Nuclear reactions are becoming essential in human life in the form of electricity production from nuclear power plants, a source of radioisotopes for medical imaging to visualize organs and diagnose diseases, to treat tumors, and cancerous cells, as shown in Fig. 8.1.2.

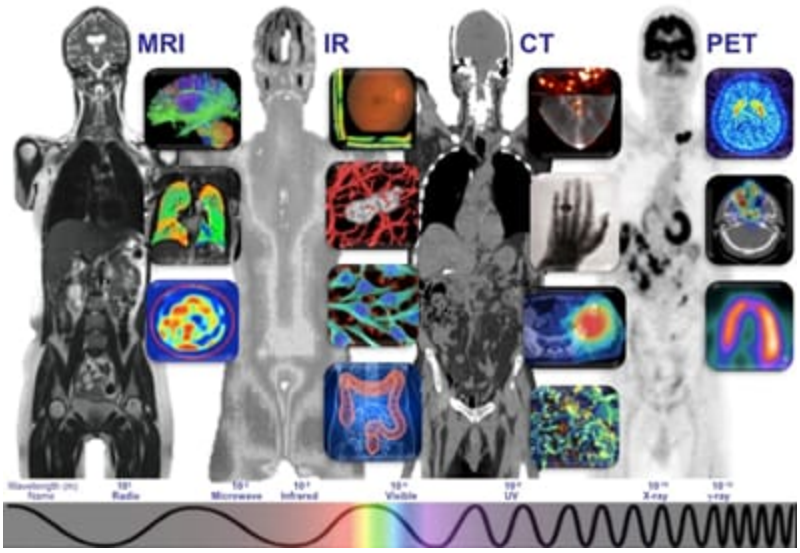




Figure 3.9.1.2: Spectrum of medical imaging (left) and radiation therapy of the pelvis. (right). Source: Martin Tornai, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>, via Wikimedia Commons, and Dina Wakulchik from Indianapolis, Indiana, USA / CC BY (https://creativecommons.org/licenses/by/2.0)

Nuclear nomenclature and symbols

Nucleoid

> Nucleoid is another name for the nucleus of an atom, that is often used in nuclear chemistry

The composition of a nucleoid is represented by the same symbol that represents the isotopes of elements, as ${}^{A}_{Z}\mathrm{X}$, where X is the element symbol, Z is the number of protons, and A is the number of protons and neutron in the nucleus. For example, carbon exists as a mixture of ${}^{12}_{6}\mathrm{C}$, and ${}^{13}_{6}\mathrm{C}$ isotopes. The name of the element, followed by the number of nucleons separated by a hyphen, is another way of representing a nucleoid. For example, carbon-12, carbon-13, and carbon-14 represent the carbon nucleoids having 6 protons each, but 6, 7, and 8 neutrons, respectively. Similarly, hydrogen exists as a mixture of ${}^{1}_{1}\mathrm{H}$, ${}^{2}_{1}\mathrm{H}$, and ${}^{3}_{1}\mathrm{H}$, that can also be represented as hydrogen-1, hydrogen-2, and hydrogen-3, respectively.

Nucleons

The protons and neutrons are also called nucleons.

Nuclear reaction

A nuclear reaction is a process in which two nuclei, or a nucleus and an external subatomic particle, collide to produce one or more new nuclides.

The nuclear reaction is a reaction that involves nucleoids. The reactant nucleoid, called the **parent nucleoid**, usually transforms into a different nucleoid called the **daughter nucleoid**. The daughter nucleoid may be an isotope of the parent nucleoid, or it may be a different element. The conversion of an isotope to another isotope of the same or a different element is a nuclear reaction that is called **transmutation** or a nuclear transformation, as shown in Fig. 8.1.3.

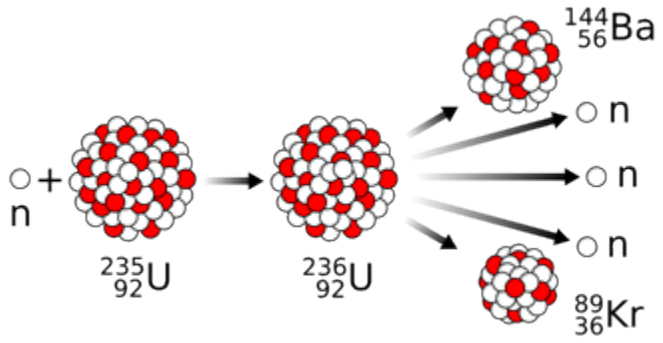


Figure $3.9.1.3$: A nuclear fission reaction in which a partent nucleoid $^{235}_{92}\mathrm{U}$ absorbs a neutron and transforms to a daughter nucleoid $^{236}_{92}\mathrm{U}$, which later on transforms to two daughter nucleoids $^{144}_{56}\mathrm{Ba}$ and $^{89}_{36}\mathrm{Kr}$ along with emission three protons. Copy and Paste Caption here. Source: MikeRun, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

Nuclear radiations

Nuclearradiation or radioactivity is the particles and energy emitted by the nucleus during a nuclear reaction.

The nuclear reaction is accompanied by the emission of nuclear radiations including high-energy electromagnetic traditions called gamma-rays (γ-rays), subatomic particles like electrons, positrons, protons, neutrons, or a small nucleus, like $^{4}_{2}\mathrm{He}^{2+}$ called alpha-particles (α-particles). Nuclear radiations are ionizing radiations, i.e., they can knock off electrons from the atoms they come in contact with.

Radioactive

The nucleoids that are capable of spontaneous disintegration, causing the emission of nuclear radiation, are called radioactive.

The process of emission of nuclear radiation by a spontaneous disintegration of radioactive nucleoids is called radioactivity, as illustrated in Fig. 8.1.4.

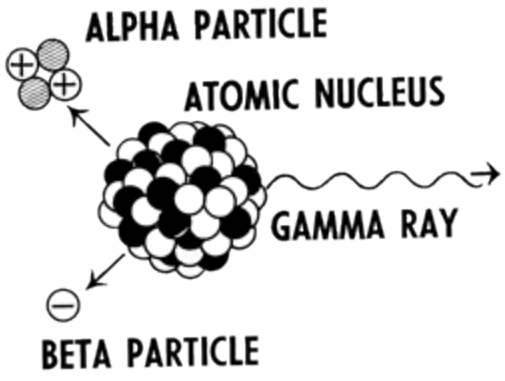


Figure 3.9.1.4: Illustration of radioactive decay of a nucleoid. Source: Pearson Scott Foresman, Public domain, via Wikimedia Commons

The nuclear radiations include gamma-rays (γ-rays), alpha-particles (α-particles), beta-particle (β-particles), neutrons (n), and positron (β^+-particles).

Gamma-rays

The gamma-rays are electromagnetic radiations that have no mass and have energy higher than that of X-rays. The symbol γ, ${}^{0}_{0}\gamma$, or γ-ray represents a gamma-ray.

Alpha-particles

The alpha-particles (α-particles) are helium nuclei with two protons, two neutrons, and without electrons, i.e., ${}^{4}_{2}\mathrm{He}^{2+}$. The α -particles are also represented as ${}^{4}_{2}\mathrm{He}$ or Helium-4.

Beta-particles

The beta-particles (β-particles) are fast-moving electrons that have atomic number -1, charge -1, and negligible mass. The symbol β, β^-, ${}^{0}_{-1}\beta$, or ${}^{0}_{-1}\mathrm{e}$ also represents a β-particle.

Positrons

Positrons are anti-particle of electron, i.e., they have the same mass but opposite charge than that of an electron. The symbol $+\beta$, β^+, ${}^{0}_{+1}\beta$, or ${}^{0}_{+1}\mathrm{e}$ represents a positron.

This page titled 3.9.1: Introduction to nuclear chemistry is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.9.2: Radioactivity

The cause of radioactivity

Only a particular combination of protons and neutrons forms stable nucleoids; the rest are unstable nucleoids, as illustrated in Fig. 8.2.1. The observations on the stable nucleoids are the following.

Observations on the stable nucleoids

1. Hydrogen with one proton and no neutrons is a stable nucleoid.
2. Other light nucleoids, up to atomic number 20 are usually stable when the number of protons is equal to the number of neutrons
3. The medium nucleoids from atomic numbers 20 to 82 are generally stable when the number of neutrons is more than the number of protons.
4. The heavier nucleoids with atomic number 84 or more are unstable.

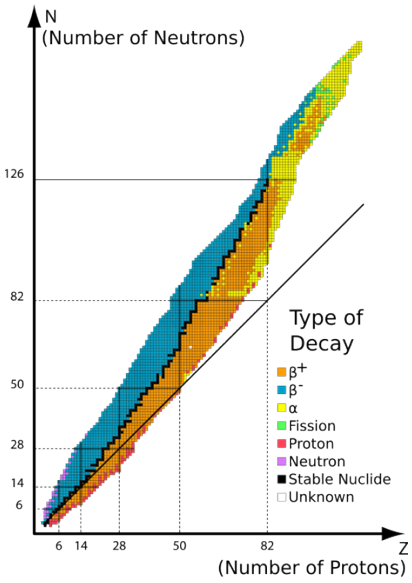


Figure 3.9.2.1: Graph of isotopes by type of nuclear decay. Orange and blue nuclides are unstable, with the black squares between these regions representing stable nuclides. The unbroken line passing below many of the nuclides represents the theoretical position on the graph of nuclides for which the proton number is the same as the neutron number. The graph shows that elements with more than 20 protons must have more neutrons than protons, in order to be stable. Source: Table_isotopes.svg: Napy1kenobiderivative work: Sjlegg / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

The nucleoids shown as black dots in Fig. 8.2.1 are naturally occurring stable nucleoids. All nucleoids shown in colored dots other than black are unstable.

Note

The unstable nucleoids are radioactive -they spontaneously disintegrate, i.e., they tend to re-arrange the nucleon composition in the nucleus to become a more stable nucleoid.

The radioactivity releases energy and particles, i.e., nuclear radiation along with the re-arrangement of the nucleus, as explained in the following sections.

Alpha-decay

The α-emission is one of the important processes for stabilizing heavy unstable nucleoids shown with yellow dots in Fig. 8.2.1. Since the α-particle has two protons and two neutrons, the daughter nucleoid produced along with the α-decay has two fewer protons and four fewer mass numbers in it than the parent nucleoid, as illustrated in Fig. 8.2.2.

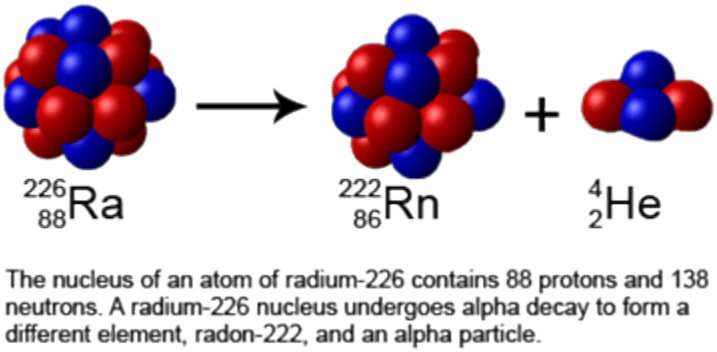


Figure 3.9.2.2: Illustration of alpha-decay process. Source: PerOX / CC0

For example:

$$^{238}_{92}\mathrm{U} \longrightarrow {}^{234}_{90}\mathrm{Th} + {}^{4}_{2}\mathrm{He}$$

Some times the α-decay is accompanied by γ-emission, e.g.:

$$^{210}_{84}\mathrm{Po} \longrightarrow {}^{206}_{82}\mathrm{Pb} + {}^{4}_{2}\mathrm{He} + \gamma$$

Note that the γ-rays have zero mass, so they do not change the atomic number and mass number of the parent nucleoid.

Steps of writing a nuclear reaction equation

1. Write the symbols of the know nucleoids, particles, and radiations in the reactants and products, separated by an arrow. Leave a question mark for the unknown pieces of information.
2. Balance the mass number on the two sides of the equation.
3. Balance the atomic number on the two sides of the equation.
4. Wright the symbols of the unknown nucleoid or particles by finding them in the periodic table, based on the atomic numbers.

α-decay in smoke detectors

Smoke detectors used in homes need α-particles for their function. Americium-241 is the α-decay-emitter used in the smoke detectors.

Example 3.9.2.1

Write the nuclear reaction equation for the α-decay of americium-241.

Solution

Step 1. The symbol and the atomic number of americium in the periodic table are Am and 95, respectively. So the initial equation is:

$$^{241}_{95}\mathrm{Am} \longrightarrow {}^{?}_{?}? + {}^{4}_{2}\mathrm{He}$$

Step 2. Balance the mass number on the two sides of the equation, i.e., the mass number of unknown nucleoid is 241-4 = 237:

$$^{241}_{95}\mathrm{Am} \longrightarrow {}^{237}_{?}? + {}^{4}_{2}\mathrm{He}$$

Step 3. Balance the atomic number on the two sides of the equation, i.e., the atomic number of the unknown nucleoid is 95-2 = 93:

$$^{241}_{95}\mathrm{Am} \longrightarrow {}^{237}_{93}? + {}^{4}_{2}\mathrm{He}$$

Step 4. Find the symbol of the unknown nucleoid from the periodic table of elements, i.e., the element at atomic number 93 is neptunium symbol Np:

$$^{241}_{95}\mathrm{Am} \longrightarrow {}^{237}_{93}\mathrm{Np} + {}^{4}_{2}\mathrm{He}$$

This is the balanced nuclear equation for the α-decay of americium-241 in smoke detectors.

Radium-226, present in many types of rocks and soils, is an α-emitter producing radon-226 in the process. The radon-226 is also an α-emitter that can diffuse into houses from the rocks and soil underneath the buildings. Radon is an environmental health issue in the buildings when its concentration becomes above a certain level. The nuclear equation for the α-decay of radon-222 is the following.

$$^{222}_{86}\mathrm{Rn} \longrightarrow {}^{218}_{84}\mathrm{Po} + {}^{4}_{2}\mathrm{He}$$

Example 3.9.2.2

Write the nuclear equation for the α-decay of radium-226.

Solution

Step 1. The initial equation is:

$$^{226}_{88}\mathrm{Rn} \longrightarrow {}^{?}_{?}? + {}^{4}_{2}\mathrm{He}$$

Step 2. Balance the mass number on the two sides of the equation, i.e., is 226-4 = 222:

$$^{226}_{88}\mathrm{Rn} \longrightarrow {}^{222}_{?}? + {}^{4}_{2}\mathrm{He}$$

Step 3. Balance the atomic number on the two sides of the equation, i.e., 88-2 = 86:

$$^{226}_{88}\mathrm{Rn} \longrightarrow {}^{222}_{86?}? + {}^{4}_{2}\mathrm{He}$$

Step 4. Find the symbol of the unknown nucleoid from the periodic table of elements, i.e., the element at atomic number 86 is radon symbol Rn:

$$^{226}_{88}\mathrm{Rn} \longrightarrow {}^{222}_{?}\mathrm{Rn} + {}^{4}_{2}\mathrm{He}$$

This is the balanced nuclear equation for the α-decay of radium-226 in rocks and soil.

Beta-decay

The nucleoids marked blue in Fig. 8.2.1. have more neutrons than needed for stability. They usually stabilize them by converting one of the neutrons (n) into a proton (p) and an electron (e) by the following nuclear process:

$$^{1}_{0}\mathrm{n} \longrightarrow {}^{1}_{1}\mathrm{p} + {}^{0}_{-1}\mathrm{e}$$

The proton stays in the nucleus, but the electron emits from the nucleus, as illustrated in Fig. 8.2.3. The emitted electron is called

β-particle. The process of the β-particle emission is called beta-decay. Note that the neutron has zero atomic number as there is no proton in it, and the electron has a -1 atomic number to balance the +1 atomic number of the proton. Proton has +1 and electron has -1 charge, which is also balanced. The mass number of an electron is zero as it has negligible mass compared to the mass of a proton or a neutron. The electrical charges and the emission of another particle called neutrino are ignored in this equation.

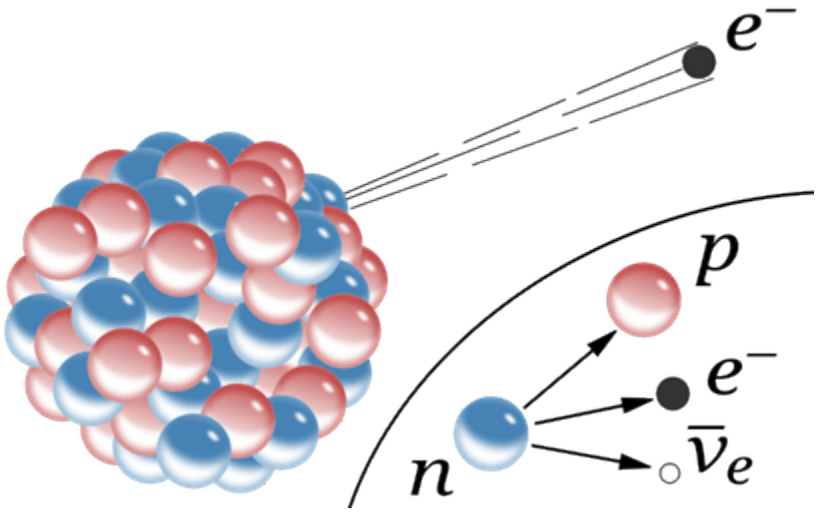


Figure 3.9.2.3: Illustration of beta-decay. A neutron (n) is transformed into a proton (p) in the nucleus, ejecting an electron (e^-) and some energy as radiation. Source: Inductiveload/ Public domain

An example of β-decay is the transformation of nitrogen-16 to oxygen-17:

$$^{16}_{7}\mathrm{N} \longrightarrow {}^{16}_{8}\mathrm{O} + {}^{0}_{-1}\mathrm{e}$$

Note that in the β-decay process, the mass number remains the same, but the atomic number increases by one in the daughter nucleus. The nuclear equation is balanced because the mass number is the same (16 = 16+0), and the atomic number is also the same (7 = 8-1) on the two sides of the equation.

Uses of some beta-emitters

Iodine-131 is used for radiation therapy of an overactive thyroid gland.

Yttrium-90 is used to treat cancer and is also injected into large joints to relieve the pain due to arthritis.

Phosphorous-32 is used to treat leukemia and other blood disorders.

Carbon-14 is used to determine the age of a fossil or an old object.

Example 3.9.2.3

Write the nuclear equation for the β-decay of iodine-131.

Solution

Step 1. The symbol and the atomic number of iodine in the periodic table are I and 53, respectively. So the initial equation is:

$$^{131}_{53}\mathrm{I} \longrightarrow {}^{?}_{?}? + {}^{0}_{-1}\mathrm{e}$$

Step 2. Balance the mass number on the two sides of the equation, i.e., the mass number of unknown nucleoid is 131-0 = 131:

$$^{131}_{53}\mathrm{I} \longrightarrow {}^{131}_{?}? + {}^{0}_{-1}\mathrm{e}$$

Step 3. Balance the atomic number on the two sides of the equation, i.e., the atomic number of the unknown nucleoid is 53-(-1) = 54:

$$^{131}_{53}\mathrm{I} \longrightarrow {}^{131}_{54}? + {}^{0}_{-1}\mathrm{e}$$

Step 4. Find the symbol of the unknown nucleoid from the periodic table of elements, i.e., the element at atomic number 54 is xenon symbol Xe:

$$^{131}_{53}\mathrm{I} \longrightarrow {}^{131}_{54}\mathrm{Xe} + {}^{0}_{-1}\mathrm{e}$$

This is the balanced nuclear equation for the α-decay of iodine-131 that is used to treat over-active thyroid glands.

Example 3.9.2.4

Write the nuclear equation for the β-decay of yttrium-90.

Solution

Step 1. The symbol and the atomic number of yttrium in the periodic table are Y and 39, respectively. So the initial equation is:

$$^{90}_{39}\mathrm{Y} \longrightarrow {}^{?}_{?}? + {}^{0}_{-1}\mathrm{e}$$

Step 2. Balance the mass number on the two sides of the equation, i.e., the mass number of unknown nucleoid is 90-0 = 90:

$$^{90}_{39}\mathrm{Y} \longrightarrow {}^{90}_{?}? + {}^{0}_{-1}\mathrm{e}$$

Step 3. Balance the atomic number on the two sides of the equation, i.e., the atomic number of the unknown nucleoid is 39-(-1) = 40:

$$^{90}_{39}\mathrm{Y} \longrightarrow {}^{90}_{40}? + {}^{0}_{-1}\mathrm{e}$$

Step 4. Find the symbol of the unknown nucleoid from the periodic table of elements, i.e., the element at atomic number 40 is zirconium symbol Z:

$$^{90}_{39}\mathrm{Y} \longrightarrow {}^{90}_{40}\mathrm{Z} + {}^{0}_{-1}\mathrm{e}$$

This is the balanced nuclear equation for the β-decay of yttrium-90.

Positron emission

The nucleoids marked orange in Fig. 8.2.1. have more protons than needed for stability. They usually stabilize them by converting one of the protons (p) into a neutron (n) and a positron β^+ by the following nuclear process:

$$^{1}_{1}\mathrm{p} \longrightarrow {}^{1}_{0}\mathrm{n} + {}^{0}_{1}\mathrm{e}$$

The neutron stays in the nucleus, but the positron emits from the nucleus, as illustrated in Fig. 8.2.4. Note that the positron has a +1 mass number that balances the +1 atomic number of the proton on the other side of the equation. The positron has a +1 charge that also balances the +1 charge of the proton on the other side of the equation. The mass number of a positron is zero as it has negligible mass compared to the mass of a proton or a neutron. The electrical charges are not shown in the nuclear equation.

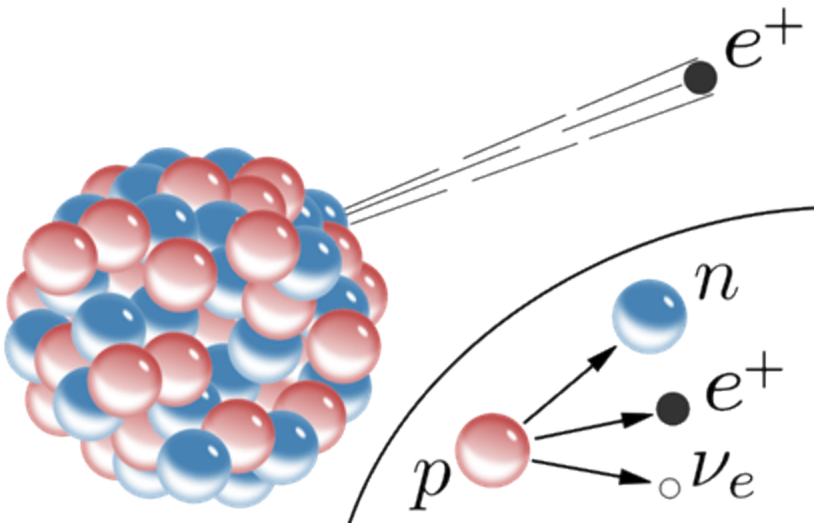


Figure 3.9.2.4: Illustration of positron-emission. A proton (p) is transformed into a neutron (n) in the nucleus, ejecting a positron (e^+) and some energy as radiation. Source: Mastermind / Public domain

Carbon-11 is an example of positron-emitter:

$$^{11}_{6}\mathrm{I} \longrightarrow {}^{11}_{5}\mathrm{B} + {}^{0}_{1}\mathrm{e}$$

Note that in the positron-emission process, the mass number remains the same, but the atomic number decreases by one in the daughter nucleus. The nuclear equation is balanced because the mass number is the same (11 = 11+0), and the atomic number is also the same (7 = 5+1) on the two sides of the equation.

Uses of some positron emittors

Positron emission is used in positron emission tomography (PET) which is a medical imaging technique. Short-lived positron-emitting isotopes ^{11}C, ^{13}N, ^{15}O, and ^{18}F used for positron emission tomography are typically produced by proton irradiation of natural or enriched targets described in a later section.

Fluorine-18 in fluorodeoxyglucose, abbreviated as [^{18}F]FDG is a positron emitter commonly used to detect cancer, and in [^{18}F]NaF is widely used for detecting bone formation. Other examples are oxygen-15 in [^{15}O]H_2O used to measure blood flow and nitrogen-13 used to tag ammonia molecules for myocardial perfusion imaging.

$$^{15}_{8}\mathrm{O} \longrightarrow {}^{15}_{7}\mathrm{N} + {}^{0}_{1}\mathrm{e}$$

$$^{13}_{7}\mathrm{N} \longrightarrow {}^{13}_{6}\mathrm{C} + {}^{0}_{1}\mathrm{e}$$

Example 3.9.2.5

Write the nuclear equation for the positron-emission of fluorine-18?

Solution

Step 1. The symbol and the atomic number of fluorine in the periodic table are F and 9, respectively. So the initial equation is:

$$^{18}_{9}\mathrm{F} \longrightarrow {}^{?}_{?}? + {}^{0}_{1}\mathrm{e}$$

Step 2. Balance the mass number on the two sides of the equation, i.e., the mass number of unknown nucleoid is 18-0 = 131:

$$^{18}_{9}\mathrm{F} \longrightarrow {}^{131}_{?}? + {}^{0}_{1}\mathrm{e}$$

Step 3. Balance the atomic number on the two sides of the equation, i.e., the atomic number of the unknown nucleoid is 9-(+1) = 8:

$$^{18}_{9}\mathrm{F} \longrightarrow {}^{131}_{8}? + {}^{0}_{1}\mathrm{e}$$

Step 4. Find the symbol of the unknown nucleoid from the periodic table of elements, i.e., the element at atomic number 8 is oxygen symbol O:

$$^{18}_{9}\mathrm{F} \longrightarrow {}^{18}_{8}? + {}^{0}_{1}\mathrm{e}$$

This is the balanced nuclear equation for the positron-emission of fluorine-18.

Gamma-emission

The gamma-rays are high-energy electromagnetic radiations that do not have mass or charge. So, pure γ-emission happens from the nucleus, but it does not result in transmutation, simply the nucleoid changes from a more unstable state, called a **metastable state**, to a relatively stable state, as illustrated in Fig. 8.2.5.

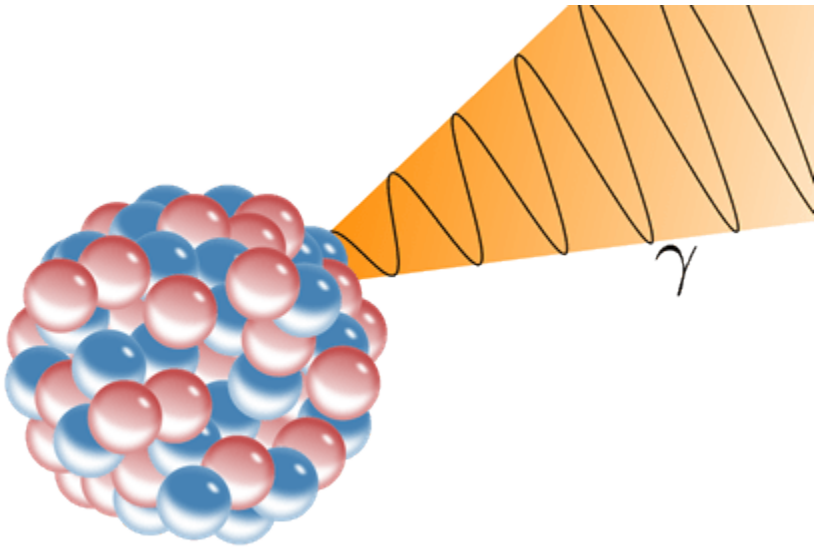


Figure 3.9.2.5: An illustration of gamma-emission, the γ-ray originates from the radioactive nucleus, changes the energy state but does not change the identity of the nucleoid. Source: Inductiveload / Public domain

A symbol *m* or * next to the mass number as a superscript to the right indicates the metastable state of the parent nucleoid. For example, technetium-99m is a γ-emitter widely used in medical imaging:

$$^{99m}_{43}\mathrm{Tc} \longrightarrow {}^{199}_{43}\mathrm{Tc} \ + \gamma$$

Similarly, boron-11m is a γ-emitter:

$$^{11m}_{5}\mathrm{B} \longrightarrow {}^{11}_{5}\mathrm{B} \ + \gamma$$

Note that the nucleoid remains the same after γ-emission, except for the change form metastable to a more stable state indicated by m.

Often, the γ-emission accompanies α-emission or β-emission. For example, polonium-210 decays by a simultaneous α-emission and γ-emissions.

$$^{210}_{84}\mathrm{Po} \longrightarrow {}^{206}_{82}\mathrm{Pb} \ + {}^{4}_{2}\mathrm{He} + \gamma$$

Similarly, iridium-192 used in implants to treat breast cancer, and cobalt-60 used as an external radiation source for cancer treatment, simultaneously emit β and γ-rays.

$$^{192}_{77}\mathrm{Ir} \longrightarrow {}^{192}_{78}\mathrm{Pt} \ + {}^{0}_{-1}\mathrm{e} + \gamma$$

$$^{60}_{27}\mathrm{Co} \longrightarrow {}^{60}_{28}\mathrm{Ni} \ + {}^{0}_{-1}\mathrm{e} + \gamma$$

Iodine-131 decays to β-particle and xenon-131m that is rapidly followed by a γ-decay of xenon-131m.

$$^{131}_{53}\mathrm{I} \longrightarrow {}^{131m}_{54}\mathrm{Xe} \ + {}^{0}_{-1}\mathrm{e}$$

$$^{131m}_{54}\mathrm{Xe} \longrightarrow {}^{131}_{54}\mathrm{Xe} + \gamma$$

Less common forms of radioactivity

Several relatively less common forms of radioactivity are known. Some examples are the following.

1. Neutron-emission is a mode of radioactive decay in which one or more neutrons are ejected from a nucleus.
2. Proton-emission is a rare form of radioactivity in which a proton emits from a nucleoid.
3. Spontaneous fission is a radioactive process in which a more massive nucleoid breaks into smaller nucleoids, often along with the emission of smaller nuclear particles.
4. In electron-capture, an external electron is captured to react with proton and produce a neutron in the nucleus.

For example, beryllium-7 decays by electron capture, as shown in the following equation.

$$^{7}_{4}\mathrm{Be} + {}^{0}_{-1}\mathrm{e} \longrightarrow {}^{7}_{3}\mathrm{Li} + \gamma$$

Note that the mass number remains the same, but the atomic number decreases by one in the electron-capture process.

Chromium-51, which is used for imaging the spleen, decays by the electron capture and γ-emission.

$$^{51}_{24}\mathrm{Cr} + {}^{0}_{-1}\mathrm{e} \longrightarrow {}^{51}_{23}\mathrm{V} + \gamma$$

Summary of decay mode of radioactive nucleoids

Fig. 8.2.6 summarizes the changes in the nucleoid composition upon different radioactive events described above.

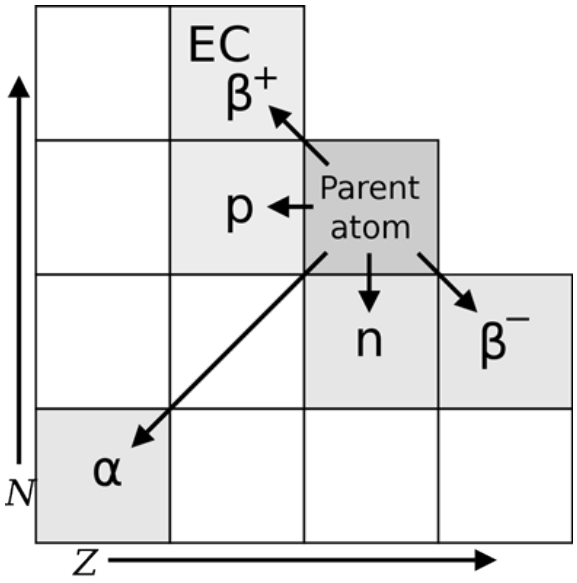


Figure 3.9.2.6: Transition diagram for decay modes of a radionuclide, with

neutron number N and atomic number Z (shown are α, , $\beta^{\pm}$ are beta-particle, and positron,$p^{+}$is proton, and n^{0} is neutron missions, EC denotes electron capture). Source: MarsRover / GFDL (http://www.gnu.org/copyleft/fdl.html)

This page titled 3.9.2: Radioactivity is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.9.3: Half-life of radioisotopes

The half-life ($t_{1/2}$) of a radioisotope is the time it takes for half of the sample to decay.

It tells the rate of decay of the radioisotope – the faster the rate of decay, the shorter the half-life.

General features of the half-life

1. The half-life is different for different nucleoids, as shown in Fig. 8.3.1, and Table 1. It varies from a fraction of a second to more than 10^{20} s, i.e., more than 3 trillion years.
2. The farther a nucleoid is away from the stable nucleoid (shown by black dots in Fig. 8.3.1), the less stable it is, and the faster it decays.
3. The half-life is independent of concentration, temperature, and pressure, i.e., the $t_{1/2}$ is a characteristic constant of a radioisotope.
4. Natural radioactive isotopes usually have a longer half-life, e.g., $t_{1/2}$ carbon-14 is 5730 years and uranium-235 is 7.0 x 10^8 years.
5. The radioisotopes used for imaging and treatment in medical sciences are usually synthesized and have a short half-life so that they may not persist in the body for an unnecessarily long time. For example, phosphorous-32, iodine-131, and technetium-99m have half-lives of 14.3 days, 8.1 days, and 6.0 hours, respectively.

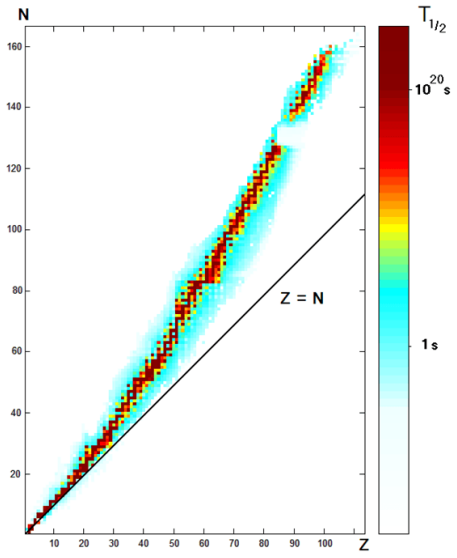


Figure 3.9.3.1: Graph of the stability of every known nucleoid. Plotted as Z (number of protons) versus N (number of neutrons). The color corresponds to the value of the half-life $t_{½}$ with a strong log scale, as it varies between 10^{-20} and 10^{20} seconds. Source: Fffred~commonswiki/ Public domain

Table 1: Half-lives ($t_{1/2}$) of some common radioisotopes

Radioisotope	Symbol	Half-life	Use
Carbon-14	$^{14}_{6}\mathrm{C}$	5730 years	Radioisotope dating
Hydrogen-3	$^{3}_{1}\mathrm{H}$	12.3 years	Radioisotope dating
Potassium-40	$^{40}_{19}\mathrm{K}$	1.3 x 10^9 years	Radioisotope dating
Rhenium-187	$^{187}_{75}\mathrm{Re}$	4.3 x 10^{10}years	Radioisotope dating

Radioisotope	Symbol	Half-life	Use
Uranium-238	$^{238}_{92}\mathrm{U}$	4.5 x 10^9years	Radioisotope dating
Uranium-235	$^{235}_{92}\mathrm{U}$	7.0 x 10^8 years	Nuclear reactor fuel
Cobalt-60	$^{60}_{27}\mathrm{Co}$	5.3 years	Medical (external radiation source)
Iodine-131	$^{131}_{53}\mathrm{I}$	8.1 days	Medical
Iron-59	$^{59}_{26}\mathrm{Fe}$	45 days	Medical
Molybdenum-99	$^{99}_{42}\mathrm{Mo}$	67 hours	Medical
Sodium-24	$^{24}_{11}\mathrm{Na}$	15 hours	Medical
Technetium-99m	$^{99m}_{43}\mathrm{Te}$	6 hours	Medical
Phosphoros-32	$^{32}_{15}\mathrm{P}$	14.3 dsys	Medical

Decay curve of radioisotopes

During each successive half-life, half of the initial amount will radioactively decay, as illustrated in Fig. 8.3.2. for the case of phosphorous-32 that decays with a half-life of 14.3 days by the following nuclear reaction

$$^{32}_{15}\mathrm{P} \longrightarrow {}^{32}_{16}\mathrm{S} + {}^{0}_{-1}\mathrm{e}$$

Suppose there is 100 mg of the phosphorous-32 in the beginning; 50 mg will be left behind after 14.3 days, i.e., after 1 half-life; and 25 mg will be left after 28.6 days, i.e., after 2 half-lives. A negligible amount of the parent isotope phosphorous-32 is left after 9 half-lives.

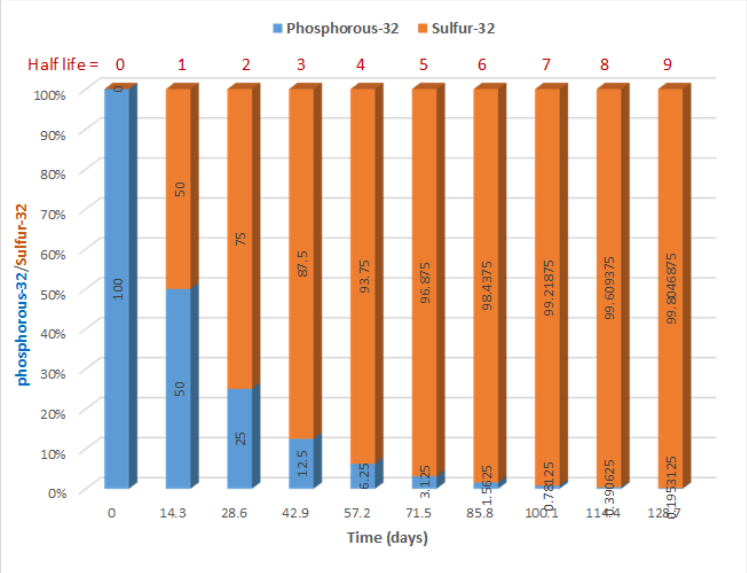


Figure 3.9.3.2: Decay of phosphorous-32 (blue color) and accumulation of sulfur-32 (orange color) after each half-life.

The amount of a radioisotope remaining after the given time can be calculated from the known initial amount and time spent, by the following formula:

$$m_f = m_i(0.5)^n$$

where m_i is the initial amount, m_f is the final amount, and n is the number of half-lives passed. The formula works even if the number of half-lives is not a whole number.

Example 3.9.3.1

If 50.0 mg of iodine-131 was injected for medical treatment, how many milligrams will be left after 40.5 days? (Half-life of iodine-131 is 8.1 days)

Solution

Given: m_i = 50.0 mg, Time = 40.5 days, Desired ? m_f

The equality: 1 half-live = 8.1 days, gives the following conversion factors.

$$\frac{1 \text{ half-life}}{8.1 \text{ days}} \quad \text{and} \quad \frac{8.1 \text{ days}}{1 \text{ half-life}}$$

For calculating the half-lives, multiple the given time with the conversion factor that cancels the time:

$$n = 40.5 \cancel{\text{ days}} \times \frac{1 \text{ half-life}}{8.1 \cancel{\text{ days}}} = 5 \text{ half-lives}$$

For calculating the amount left, plug in the values in the formula:

$$m_f = m_i(0.5)^n = 50.0\text{mg}(0.5)^5 = 1.56 \text{ mg}$$

Radioisotope dating

Natural radioactivity is used to establish the age of objects of archeological, anthropological, or historic interest. All living objects have carbon in their composition. Carbon-14 is a radioactive isotope of carbon with a half-life of 5730 years. Carbon-14 is produced by the transmutation of nitrogen-14 upon neutron bombardment by cosmic rays, as illustrated in Fig. 8.3.3. Its concentration in a carbon source for the living organism remains almost constant because its decay counterbalances its production by cosmic rays. Living organisms continuously replenish carbon, so the carbon-14 concentration remains almost constant as long as the object is alive. After the object dies, the carbon-14 decreases with time, reducing to half after one half-life. The carbon-12 isotope is not radioactive, so its concentration remains constant. Measurement of the carbon-14/carbon-12 ratio allows calculating the age of the object after its death. The age of early civilizations, like the Indus valley civilization examples shown in Fig. 8.3.4 were determined by the carbon-14 dating method.



Figure 3.9.3.4: Excavated ruins of Mohenjo-daro, Sind province of Pakistan, showing the Great Bath in the foreground. Mohenjo-daro, on the right bank of the Indus River, is a UNESCO World Heritage Site. Dated 3300 BCE to 1300 BCE. Source: Saqib Qayyum / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Because the carbon-14 decreases with time, the object older than 20,000 years of age does not have sufficient carbon-14 left to determine their age accurately. Other radioisotopes with a longer half-life, e.g., uranium-238 with a half-life of 4.5×10^9 years, are used to determine the age of ancient objects. For example, the age of rock samples from the moon, as shown in Fig. 8.3.5 was determined by uranium-238 radioisotope dating.



Figure 3.9.3.5: Lunar Olivine Basalt sample collected from the moon by the Apollo 15 mission, at station 9A on the rim of Hadley Rille. It was formed around 3.3 billion years ago. On display in the National Museum of Natural History. Source: Wknight94 talk / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

This page titled 3.9.3: Half-life of radioisotopes is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.9.4: Radiation measurements

Radioactivity measurement

Radioactivity is measured in terms of the rate of radioactive events. Nuclear radiations are ionizing radiations, i.e., they knock off electrons from atoms or molecules that come in their path, leaving behind cations. Geiger Muller counter is one of the radiations measuring instruments that counts the disintegration of radionucleotide per second by registering the current produced by the ionization action of the radiation, as illustrated in Fig. 8.4.1. It is not just one ionization event; the nuclear particle keeps ionizing the atoms in its track until its energy is exhausted, as illustrated in Fig. 8.4.2. The instrument records the flash of electric current produced by the ionization of each radioactive disintegration.

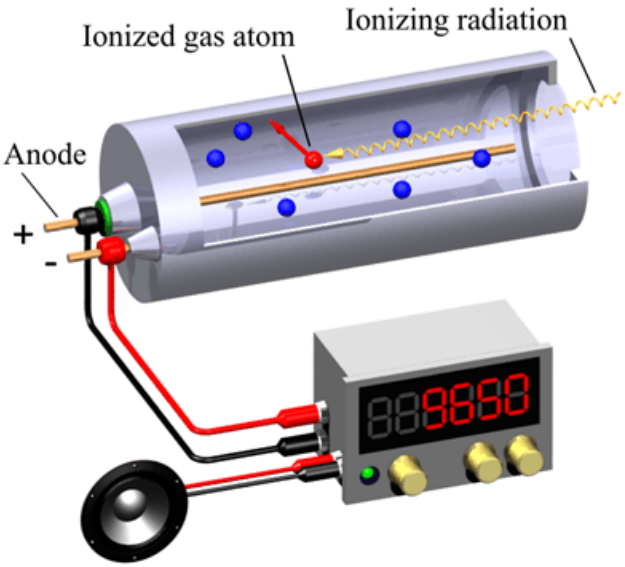


Figure 3.9.4.1: Geiger-Mueller counter. Source source Svjo-2 / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

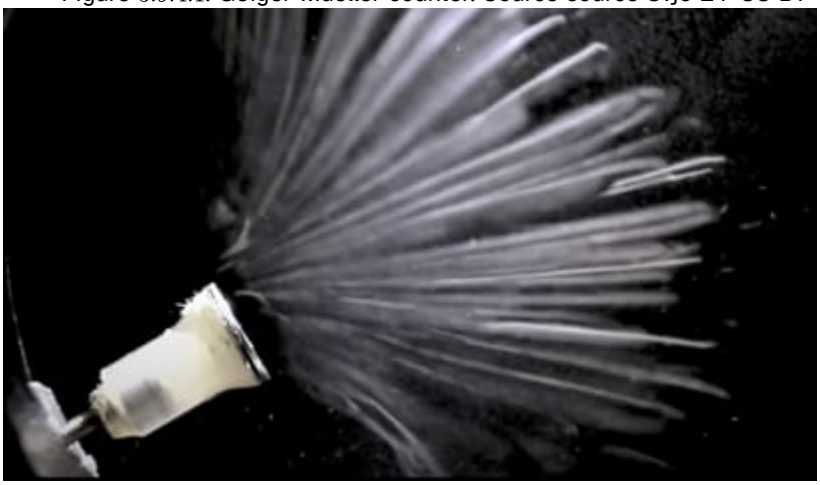


Figure 3.9.4.2: Tracks of α-particles from americium-241 made visible in a cloud chamber. Source: https://youtu.be/noP7HT-Uins

Becquerel (Bq)

The SI unit of radioactivity is Becquerel (Bq), i.e., the number of nuclei that disintegrate per second.

The common unit of radiation intensity is Curie (Ci), i.e., 3.7 x 10^{10} disintegrations per second. The relationship between Becquerel

and Curie is the following.

$$1\text{Ci} = 3.7 \times 10^{10}\ \text{Bq} = 3.7 \times 10^{10}\ \text{disintegrations}$$

Often the radioisotope for medical use has the information of millicurie per milliliter (mCi/mL) from which the volume for the desired dose can be calculated.

Example 3.9.4.1

A patient must be given a 5.0 mCi dose of iodine-131 that is available as $Na^{131}I$ solution containing 3.8 mCi/mL. What volume of the solution should be administered?

Solution

Use the reciprocal of 3.8 mCi/mL as a conversion factor:

$$5.0\ \cancel{\text{mCi}} \times \frac{1\ \text{mL}}{3.8\ \cancel{\text{mCi}}} = 1.3\ \text{mL dose}$$

Radiation exposure measurements

Absorbed dose

The ionizing radiation dose or called the absorbed dose is measured in terms of energy deposited by ionizing radiation in a unit mass of matter being irradiated, as illustrated in Fig. 8.4.3

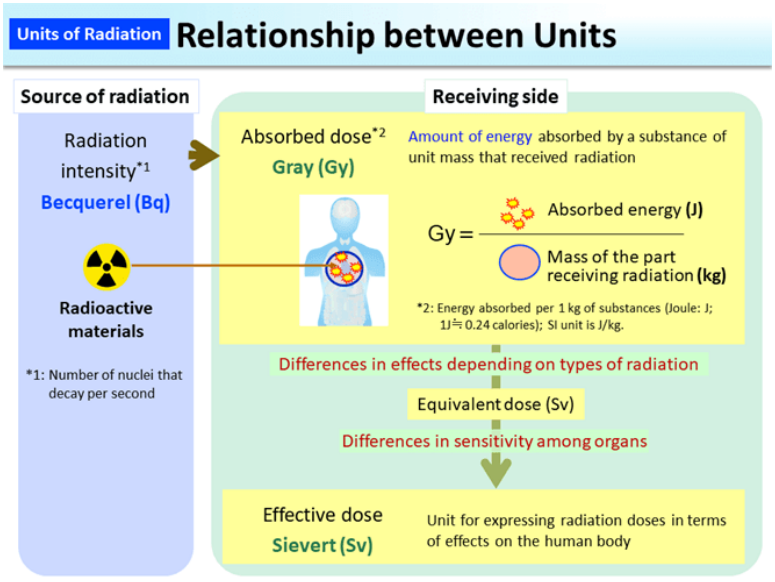


Figure 3.9.4.3: Source: Ministry of the Environment of Japan website, https://www.env.go.jp/en/chemi/rhm/b...g-02-03-03.png

Gray (Gy)

The SI unit of absorbed dose is gray (Gy) which is defined as the absorption of one joule of radiation energy per kilogram of matter (J/kg).

The common unit of absorbed dose is rad, which stands for radiation absorbed dose. The rad is one-hundredth of a gray, i.e.:

$$1\ \text{Gy} = 100\ \text{rad}$$

Equivalent dose

The same amount of energy deposited in tissues by different types of radiation carries different levels of health risks in terms of causing cancer and genetic damage, expressed as a radiation weighting factor (W_R), as illustrated in Fig. 8.4.4, and listed in Fig. 8.4.5. For example, 1 Gy of beta-particles carries a risk of 5.5% chances of eventually developing cancer, while 1 Gy of alpha-particles has 20 times more risk compared to the β-particle (ref.: https://en.Wikipedia.org/wiki/Sievert, accessed on 07/15/2020). The health risk of the ionizing radiation is measured in the units of equivalent dose. Sievert (Sv) is an SI unit of an equivalent dose of ionizing radiation that measures the health effects of low levels of ionizing radiation on the human body.

Sievert (Sv)

The equivalent doze in Sievert (Sv) is equal to the product of absorbed dose in grays (Gy) multiplied the radiation weighting factor (W_R), i.e., $\text{The equivalent dose in Sv} = \text{Absorbed dose in Gy} \times \text{W}_\text{R}$

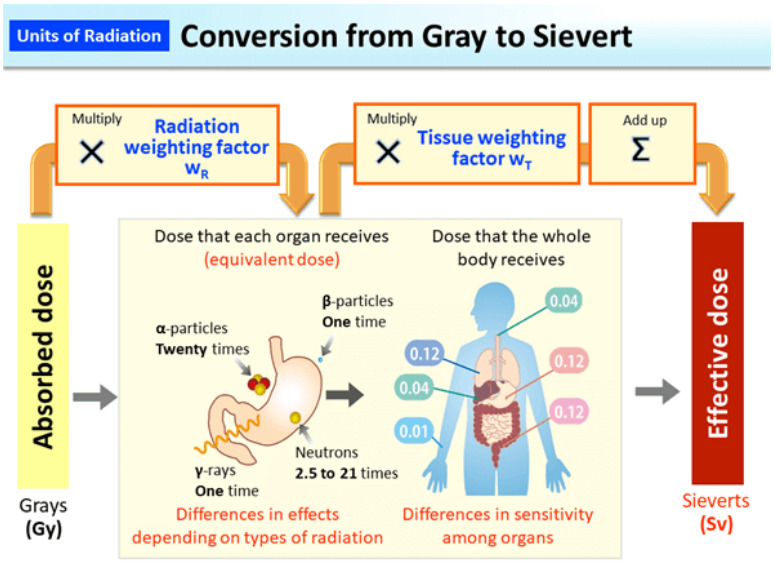


Figure 3.9.4.4: Illustration of radiation equivalent dose, effective dose, radiation weight factors (W_R), and tissue weighting factors (W_T). Source: Ministry of the Environment of Japan website, https://www.env.go.jp/en/chemi/rhm/b...g-02-03-04.png

Radiation	Energy (E)	W_R (formerly Q)
x-rays, gamma rays, beta particles, muons		1
neutrons	< 1 MeV	$2.5 + 18.2 \cdot e^{-[\ln(E)]^2/6}$
	1 MeV - 50 MeV	$5.0 + 17.0 \cdot e^{-[\ln(2 \cdot E)]^2/6}$
	> 50 MeV	$2.5 + 3.25 \cdot e^{-[\ln(0.04 \cdot E)]^2/6}$
protons, charged pions		2
alpha particles, Nuclear fission products, heavy nuclei		20

Figure 3.9.4.5: Radiation weighting factors W_R used to represent relative biological effectiveness according to ICRP report 103, Source of data: https://en.Wikipedia.org/wiki/Sievert, accessed on 07/16/2020

The common unit of equivalent dose is rem (rem stands for roentgen equivalent man), which is:

$$1\ \mathrm{Sv} = 100\mathrm{rem}$$

The personnel working in a radiation environment are required to wear film badges or electronic personal dosimeters, as shown in Fig. 8.4.6, that record the dose received. A record of each person's dose is usually maintained by the radiation facilities to comply with the allowed radiation exposure limits.



Figure 3.9.4.6: View of readout on an electronic personal dosimeter. The clip is used to attach it to the wearer's clothing. Source: Rama / CC BY-SA 2.0 FR (https://creativecommons.org/licenses...2.0/fr/deed.en)

Effective dose

The equivalent dose is equal to the effective dose in sievert (Sv) when the whole human body is exposed equally to the radiation. If part of the body is exposed, then an effective dose in sievert (Sv) is calculated by the summation of the product of equivalent dose in Sv with tissue weighting factor (W_T) for each tissue exposed to the radiation, as illustrated in Fig. 8.4.4 and calculated with example in Fig. 8.4.7. The reason for this calculation is that the effect of the same equivalent dose is different in different tissues. The tissue weighting factors (W_T) are listed in Table 1.

An effective dose takes the absorbed dose and adjusts it for radiation type and organ sensitivity, i.e.,:

$$\text{The effective dose in Sv} = \text{ Equivalent dose in Sv} \times \mathrm{W_T}$$

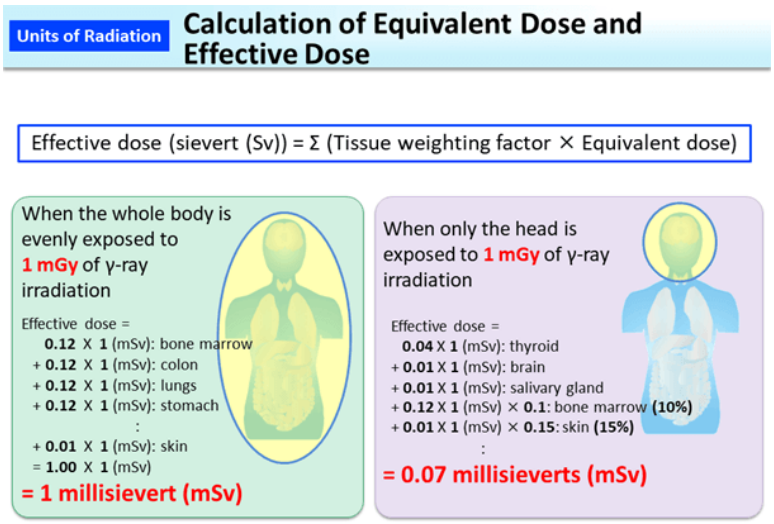


Figure 3.9.4.7: Example calculations of effective dose. Source: Ministry of the Environment of Japan website, https://www.env.go.jp/en/chemi/rhm/b...g-02-03-06.png

Table 1: Tissue weighting factors (W_T) ICRP103 (2007), source of data: Source: https://en.Wikipedia.org/wiki/Sievert, accessed on 07/16/2020

Organs	W_T
Gonads	0.08
Red bone marrow	0.12
Colon	0.12
Lung	0.12
Stomach	0.12
Breasts	0.12
Bladder	0.04
Liver	0.04
Oesophagus	0.04
Thyroid	0.04
Skin	0.01
Bone surface	0.01
Salivary glands	0.01
Brain	0.01
Remainder of body	0.12
Total	1.00

This page titled 3.9.4: Radiation measurements is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.9.5: Ionizing radiation exposures

Interaction of radiation with matter

The ionizing radiations knock off electrons from atoms and molecules. Water composes about 60% of the human body. The oxygen atom in water has eight valence electrons, four in the two bonding pairs, and four in the two lone pairs, as shown in Fig. 8.5.1. If one of them is knocked off by radiation, the result is a radical-cation. It is radical because it has one unpair electron and its octet is incomplete, and cation because an electron has been lost. The radical cation can then release a proton to become a hydroxyl radical. The radicals are very reactive species due to the incomplete octet. The radicals tend to react with any material around that causes damage to the tissues. Notably, the damage to DNA is the most dangerous, causing mutation, cancer, and hereditary problems.

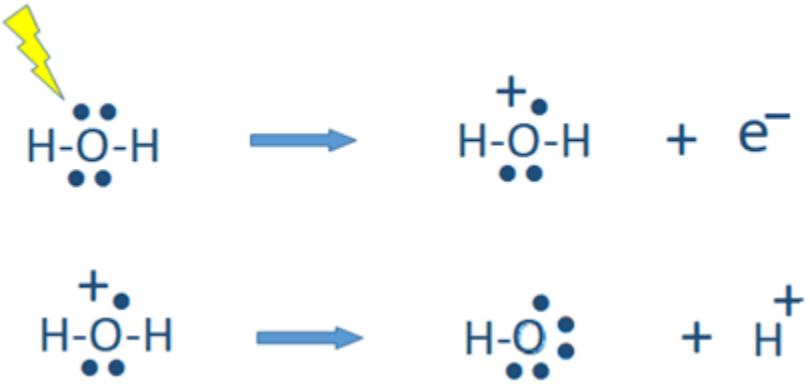


Figure 3.9.5.1: Ionization radiations knock off an electron from molecules like water, producing very reactive radical cations that lead to the production of very reactive radicals like hydroxyl radical in this case.

The cells that are proliferating are more susceptible to the harmful effects of radiation exposure, including bone marrow, skin, reproductive organs, and intestinal lining, as well as all the cells of growing children. If the bone marrow cells are damaged, the red blood cells may not be produced. Damage to the reproductive cells or the cells of a fetus may cause congenital disabilities. Fortunately, the cancer cells are multiplying and affected by the radiation exposure much more than the surrounding healthy cells, which allows for selectively killing the cancerous cells by radiation treatment. The use of ionizing radiations in cancer treatment is described in a later section.

Effects of radiation exposure on humans

Exposure of humans to less than 0.25 Sv usually does not have any noticeable effect. Whole-body exposure of 1 Sv results in a temporary decrease in the white blood cell counts. More than 1 Sv exposure may cause nausea, vomiting, fatigue, and a reduction in white blood cell count. More than 3 Sv doses to the whole body can decrease the white blood cell count to zero and cause diarrhea, hair loss, and infection. Exposure to 5 Sv can cause death in 50% of the people receiving the dose –it is called a lethal dose for one-half of the population (LD_{50}). Whole-body exposure of 6 Sv or higher is fatal to all humans within a few weeks. Fig. 8.5.2 illustrates the effects.

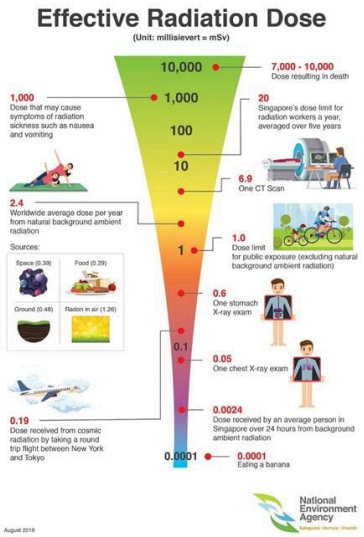


Figure 3.9.5.2: Health Effects of Ionising Radiation on People. Source: National Environment Agency, Open Data Licence, https://www.nea.gov.sg/our-services/...tion-on-people

Background radiation exposures

Humans are exposed to radiation in the cases of nuclear accidents, during nuclear radiation treatments, particularly for treating different forms of cancers, and during medical imaging for medical diagnostic purposes. Besides these human-made radiation exposures, humans are regularly exposed to natural sources of radiation, called background radiation. The background radiation can be in foods, e.g., potassium-40 is a naturally occurring radioactive isotope present in potassium-containing foods. Carbon-14, radon-222, strontium-90, and iodine-131 are other radioisotopes present in the air and foods around us. Radioactive isotopes of uranium and thorium and their decay products are the source of radiation in soil. The extraterrestrial sources of radiation are cosmic rays that are stopped in the upper atmosphere, but some may reach the ground and expose the people. The average annual radiation dose per person in the U.S. is 6.2 millisieverts (620 millirem).

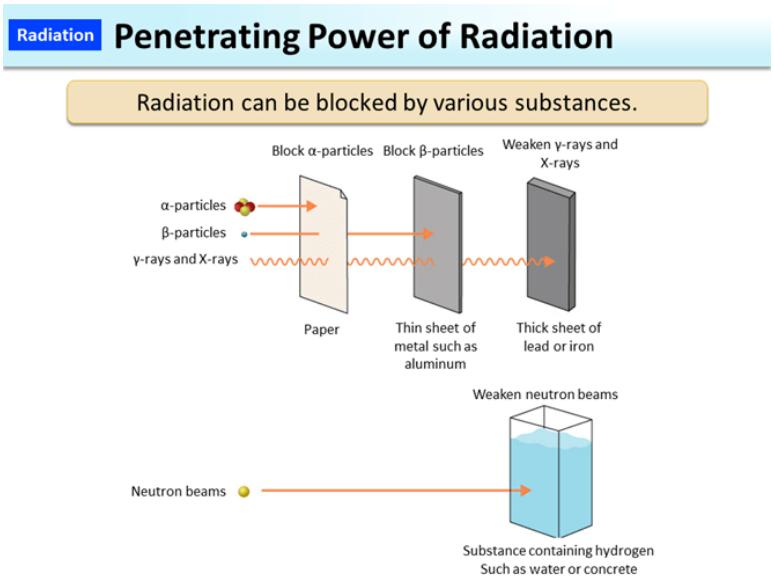


Figure 3.9.5.3: Source: Ministry of the Environment of Japan website, https://www.env.go.jp/en/chemi/rhm/b...g-01-03-08.png

Radiation protection

Different types of radiation have different penetration depths in the air and the human body. The α-particles are heavy with two protons and two neutrons; they cause much ionization but travel a few centimeters in air. A sheet of paper can stop the α-particles,

as illustrated in Fig. 8.5.3. Exposure to α-particles from an external source affects only the outer layer of the skin. However, an internal source can cause significant damage to the nearby tissues, as illustrated in Fig. 8.5.4

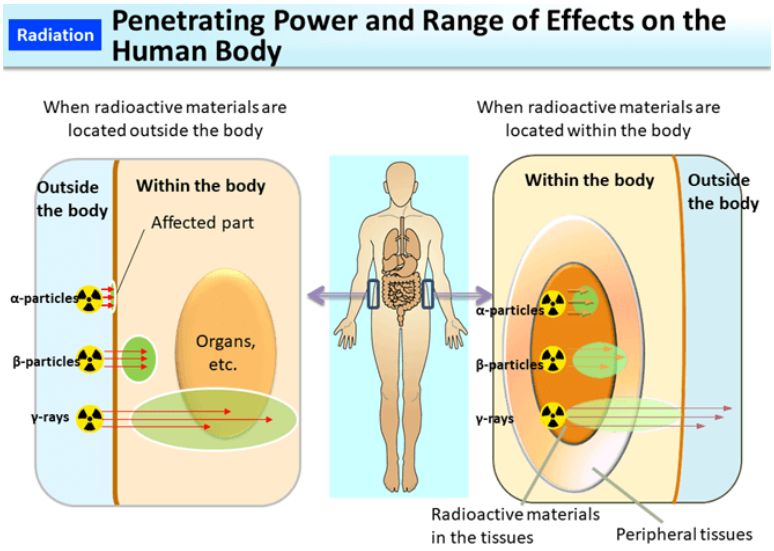


Figure 3.9.5.4: Source: Ministry of the Environment of Japan website, https://www.env.go.jp/en/chemi/rhm/b...g-01-03-10.png

The β-particles are fast-moving electrons that can travel several meters in the air. A 2-4 mm thick aluminum plate can stop them. The β-particles can penetrate 4 to 5 centimeters into the tissue. External exposure to the β-particles can burn the skin, but internal organs remain safe. Internal exposure to the β-particle is more dangerous than external exposure.

The γ-rays and X-rays can travel long distances in air, up to half a kilometer, and are not easily stopped by common material. A thick layer of lead or concrete shield is required to stop the γ-rays and X-rays. External exposure to the γ-rays and X-rays is the most dangerous because these rays can penetrate deep and damage the organs.

Neutrons are high-energy neutral particles that also have high penetrating power. They lose energy by colliding with atoms of substances. The most effective shielding is the water or concrete that has moisture in it to stop the neutrons effectively.

The workers in a radiation environment wear heavy clothes, gloves, and lab coats to provide additional protection. The radioactive materials are usually stored in shielded containers, even the syringes containing radioactive materials for injection are shielded. The general rules for protection against the radiation are:

Radiation protection measures

1. keep minimum possible time in the radiation environment -less the time means less exposure,
2. keep as much distance from the radiation source as possible -the radiation intensity drops inversely proportional to the square of the distance, and
3. keep shielding between you and the source of radiation as much as possible -the more the shielding the less the exposure.

This page titled 3.9.5: Ionizing radiation exposures is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.9.6: Medical uses of radioisotopes

The use of nuclear chemistry in medical technologies is increasing over time. The medical uses can be divided into two categories :

1. medical imaging of organs or diagnosing any malfunction and
2. therapeutic use, mainly for killing cancerous cells.

Radioisotopes in medical imaging

Criteria of radioisotope selection for medical imaging

1. those that emit γ-rays, directly or indirectly, during their decay,
2. can be delivered to the organ of interest in pure form or tagged in a compound,
3. are short-lived or can be excreted from the body after use.

Table 8.6.1 lists some radioisotopes commonly used in medical imaging. A low dose of the radioisotope is administered to a patient. The γ-rays cross over the body and are recorded like X-rays. A computer finally converts the recording into a useful image. The image is compared with an image of a healthy organ to diagnose any malfunction.

Table 8.6.1: Some of the commonly used radioisotopes in medical imaging

Radioisotope	Symbol	Mode of decay	Half-life	Use in medical imaging
Carbon-11	$^{11}_{6}\mathrm{C}$	β^+, g	20.3 m	Brain scan to trace glucose
Fluorine-19	$^{18}_{9}\mathrm{F}$	β^+, g	109m	Brain scan to trace glucose
Chromium-51	$^{51}_{24}\mathrm{Cr}$	E.C., g	27.3 d	Diagnose albinism; image the spleen and gastrointestinal track
Gallium-67	$^{67}_{31}\mathrm{Ga}$	E.C., g	78.3 h	Whole-body scan for tumors
Selenium-75	$^{75}_{34}\mathrm{Se}$	E.C., g	118 d	Pancreas scan
Krypton-81m	$^{81m}_{36}\mathrm{Kr}$	g	13.3 s	Lung ventilation scan
Xenon-133	$^{133}_{54}\mathrm{Xe}$	β	5.24 d	Lung ventilation scan
Strontium-81	$^{81}_{38}\mathrm{Sr}$	β	22.2 m	Scan for bone cancer and bone diseases
Mercury-197	$^{197}_{80}\mathrm{Hg}$	E.C., g	64.1 h	Kidney scan
Iron-59	$^{59}_{26}\mathrm{Fe}$	β, g	44.5 d	Bone marrow function and anemias
Iodine-131	$^{131}_{53}\mathrm{I}$	β, g	8.04 d	Diagnosis of thyroid malfunction
Iodine-123	$^{123}_{53}\mathrm{I}$	E.C., g	13.2 h	Diagnosis of thyroid malfunction
Technetium-99m	$^{99m}_{43}\mathrm{Te}$	g	6.01 h	Brain, liver, kidney, bone scans, and diagnosis of the damaged heart muscle
Phosphorous-32	$^{32}_{15}\mathrm{P}$	β	14.3 d	Detect eye tumors
Thallium-201	$^{201}_{81}\mathrm{Tl}$	E.C., g	3.05 d	Heart scan and exercise stress test

An example of medical imaging is the thyroid gland in the neck that produces the hormone thyroxin, which controls the overall rate of metabolism in the body. Each thyroxin molecule contains four iodine atoms. Administration of radioactive $Na^{131}I$ or $Na^{123}I$ salt accumulates the iodine in the thyroid gland in a few hours. Decay of ^{131}I and ^{123}I involves γ-emission.

$$^{131}_{53}\mathrm{I} \longrightarrow {}^{131}_{54}\mathrm{Xe} + {}^{0}_{-1}\mathrm{e} + \gamma$$

$$^{123}_{53}\mathrm{I} + {}^{0}_{-1}\mathrm{e} \longrightarrow {}^{123}_{52}\mathrm{Xe} + \gamma$$

The γ-emission from the iodine localized in the thyroid gland is recorded, as shown in Fig. 8.6.1. An overactive thyroid (hyperthyroidism) cumulates more and underactive thyroid (hypothyroidism) cumulates less iodine than a healthy thyroid.

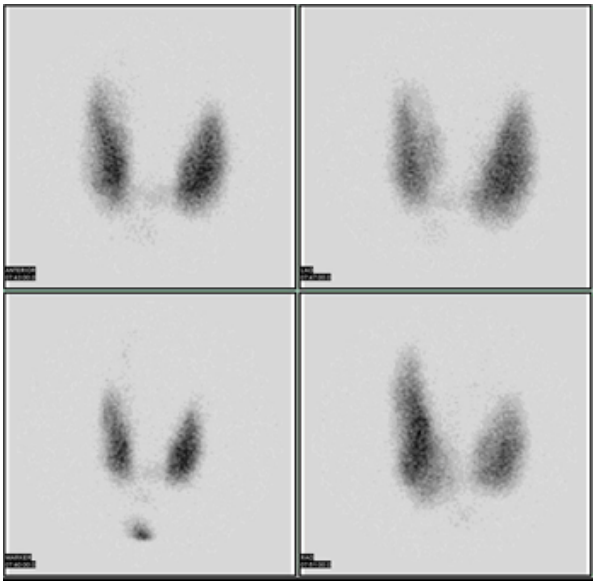


Figure 3.9.6.1: Thyroid scan with iodine-123 for evaluation of hyperthyroidism. Source: Myohan at en.Wikipedia / CC BY (https://creativecommons.org/licenses/by/3.0)

Another example is positron emission tomography (PET). Positron emitters like carbon-11 and fluorine-18 incorporated in a suitable compound like glucose allow following the metabolic path of the compound. For example, 18-fluorodeoxyglucose)18-FDG) is a glucose molecule in which one of the oxygen is replaced with ^{18}F. Intravenous injection of the 18-FDG ultimately results in the cumulation of 18-FDG in the brain and other body organs where glucose is used in the metabolic process. The ^{18}F emits a positron, which, being an anti-particle of the electron, reacts with the electron and produces two g γ-rays.

$$^{18}_{9}\mathrm{F} \longrightarrow {}^{18}_{8}\mathrm{O} + {}^{0}_{1}\mathrm{e} + \gamma$$

$${}^{0}_{1}\mathrm{e} + {}^{0}_{1}\mathrm{e} \longrightarrow 2\gamma$$

The γ-rays are used to obtain an image of the organ. The image reveals problem areas in the form of an abnormal concentration of glucose in the part of the organ. For example, Fig. 8.6.2 compares the PET image of a healthy brain versus a brain with Alzheimer's disease.

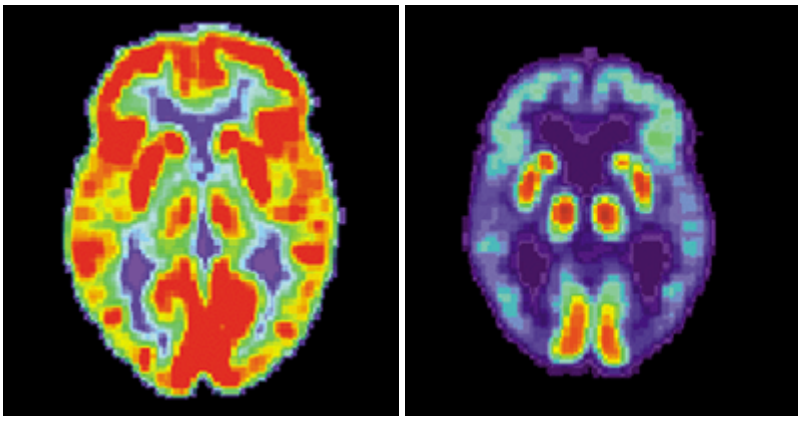


Figure 3.9.6.2: PET scan of brain: healthy (left) and Alzheimer's disease (right). Source: US National Institute on Aging, Alzheimer's Disease Education and Referral Center / Public domain

Since glucose metabolism happens in all organs, whole-body PET scans can be used to diagnose lung, colorectal, head and neck, and esophageal cancers as well as other diseases that involve abnormal glucose metabolism. For example, tumors have high metabolic rates; the PET scans using 18-FDG are used to detect them, as shown in Fig. 8.6.3.

File:PET-MIPS-anim.gif Figure 3.9.6.3: Whole-body PET scan using ^{18}F-FDG. The normal brain and kidneys are labeled, and radioactive urine from breakdown of the FDG is seen in the bladder. In addition, a large metastatic tumor mass from colon cancer is seen in the liver. Source: Jens Maus (http://jens-maus.de/) / Public domain

Medical imaging without using radioisotopes

Major examples of medical imaging using external radiation sources include the following.

1. X-rays, which are external radiation source, is commonly used for medical imaging of organs and bones.
2. Computed tomography (CT) scan uses computer processing of many X-ray measurements, from different angles, to produce a cross-sectional view (virtual slices) of the organ.
3. Magnetic resonance image (MRI) is another powerful medical imaging technique that is based on the fact that hydrogen atoms splint into two two energy states when placed in a strong magnetic field. When illuminated with infrared (IR) radiation of the energy matching with the energy gap between the two groups, the hydrogen atoms are excited from the lower to higher energy state. The decay of the excited state emits the IR radiations that are recorded to obtain the image of soft tissues that contain many hydrogen atoms in the form of water molecules. Fig. 8.6.4 shows some examples.

Figure 3.9.6.4: Spectrum of medical imaging. Source: Martin Tornai, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>, via Wikimedia Commons

Radiation therapy

The purpose of radiation therapy is to selective kill the diseased cells or tissues by exposing them to radiation. Higher radiation doses are required for therapy than for imaging. The radiation source can be external or internal.

External radiation therapy

In the external irradiation, the radiation from the source, such as coba which are often used, other radiation sources are being developed, such as proton beam from the cyclotron, which have been used to kill inoperable tumors near or in the eye, skull base, or spine. Proton therapy uses a beam of proton to deliver radiation directly to the tumor.



Figure 3.9.6.5: Radiation therapy for Hodgkin's Lymphoma in a Versa HD.. Souce: Jakembradford, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

Internal radiation therapy

Internal radiation therapy is used when a short-lived radioisotope can be made to selectively concentrate in the organ or tissue of interest. For example, iodine-131 is β and γ-emitter, is administered to a patient, is picked up by the thyroid gland, and is used to treat hyperthyroidism. Another example is actinium-225, which is an α-emitter with a half-life of 10 days. Actinium-225 installs in a monoclonal antibody that is attached to a prostate-specific antigen and delivered selectively to treat prostate tumors.

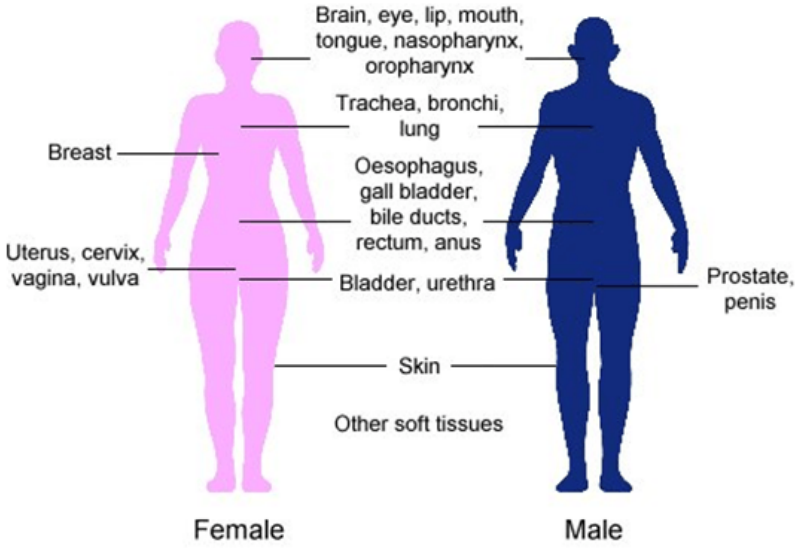


Figure 3.9.6.6: Body sites in which brachytherapy can be used to treat cancer. Source: Rock mc1 / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0)

Brachytherapy

Brachytherapy or seed implantation is a form of internal radiation treatment that delivers a high dose for a short period compared to the external radiation treatment. Fig. 8.6.6 shows the sites in the body where brachytherapy can be used to treat cancer. For example, 40 or more titanium capsules, about the size of a rice grain (Fig. 8.6.7), are implanted to treat prostate cancer. The seed contains γ-emitter like iodine-125 (half-life 60 days), palladium-103 (half-life 17 days), or cesium-131 (half-life 10 days). The seed may be left in the body because, due to the short half-life, they are no more significantly radioactive after the treatment.



Figure 3.9.6.7: Photo of radioactive seeds. These implants are a form of radiation therapy for prostate cancer. Source: Nuclear Regulatory Commission from US / Public domain

Another option is temporary brachytherapy, e.g., iridium-192 needles that deliver higher radiation doses are inserted to treat prostate cancer and removed after 5 to 10 min. The iridium-192 wires are also used as a follow-up treatment after breast cancer surgery to kill any residual or recurring cancer cells. The iridium-192 wires are inserted through a catheter implanted in the space from where the tumor was removed. The wires are removed after delivering the required radiation dose. The process is repeated

twice a day for five days. The catheter is removed, and no radioactive material is left in the body after the treatment.

This page titled 3.9.6: Medical uses of radioisotopes is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.9.7: Making radioisotopes for medical uses

Natural radioisotopes usually have a long half-life and are not best suited for medical applications. The medical application usually requires short-lived radioisotopes. The radioisotopes are usually produced in nuclear reactors where particles, like α-particles, β-particles, and neutrons, are abundant. Particle accelerators, such as the one shown in Fig. 8.7.1 also accelerate and direct the nuclear particles at the targets. The high-energy nuclear particles may be absorbed by and transmute the target nuclei to radioisotopes in a nuclear reaction.

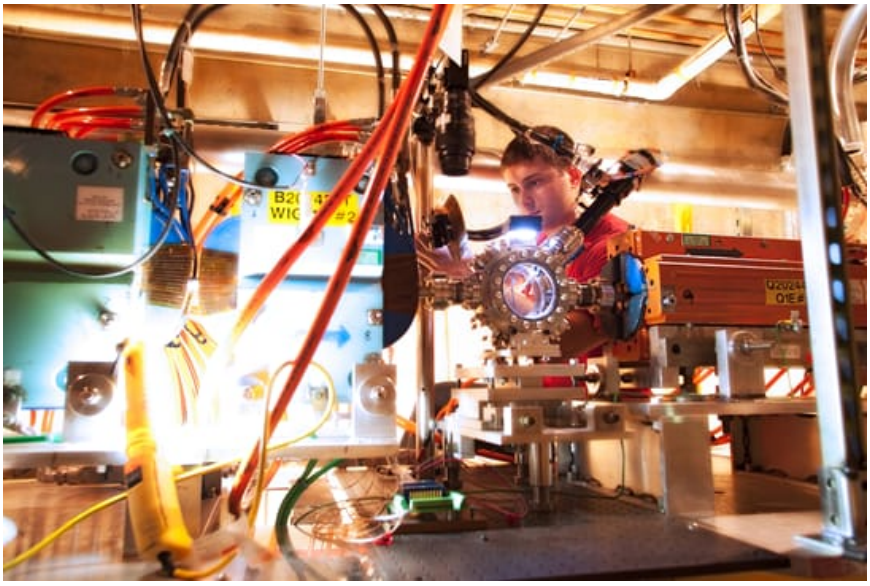


Figure 3.9.7.1: SLAC's particle accelerator may be two miles long, but researchers at FACET are working to develop more compact versions that could be widely used in medicine and industry -- particle accelerators are used for cancer research, processing computer chips, and even producing the shrink wrap used to keep your Thanksgiving turkey fresh. Source: ENERGY.GOV / Public domain

Radioisotopes in medical applications are usually produced by the particle bombardment method. One reaction that happens naturally by neutron bombardment from cosmic rays on nitrogen-14 is the following.

$$^{14}_{7}\mathrm{N} + ^{1}_{0}\mathrm{n} \longrightarrow ^{14}_{6}\mathrm{C} + ^{1}_{1}\mathrm{p}$$

An example of an artificial nuclear reaction initiated by α-particle bombardment on nitrogen, observed by Rutherford that lead to the discovery of proton, is illustrated in Fig. 8.7.2.

$$^{4}_{2}\mathrm{He} + ^{14}_{7}\mathrm{N} \longrightarrow ^{17}_{8}\mathrm{O} + ^{1}_{1}\mathrm{p}$$

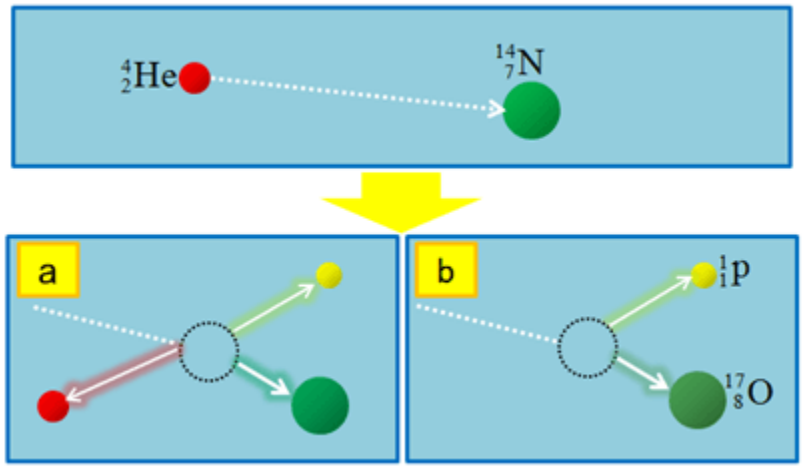


Figure 3.9.7.2: A nuclear reaction observed by

Rutherford using a cloud chamber. When an alpha ray strikes nitrogen, reaction "a" (proton knock-off) does not occur; reaction "b" (α-p) occurs.). Source: アリオト / Public domain

Another example is the nuclear reaction initiated by α-particles on beryllium, observed by James Chadwick, which lead to the discovery of the neutron.

$$ {}^{4}_{2}\mathrm{He} + {}^{9}_{4}\mathrm{Be} \longrightarrow {}^{12}_{6}\mathrm{C} + {}^{1}_{0}\mathrm{n} $$

An example of radioisotope production for medical uses is the following. Gold-198, used as a tracer in the liver, is produced by neutron bombardment on gold-197.

$$ {}^{197}_{79}\mathrm{Au} + {}^{1}_{0}\mathrm{n} \longrightarrow {}^{198}_{79}\mathrm{Au} $$

Similarly, gallium-67 used in medical diagnostics is produced by proton bombardment on zinc-66.

$$ {}^{66}_{30}\mathrm{Zn} + {}^{1}_{1}\mathrm{p} \longrightarrow {}^{67}_{31}\mathrm{Ga} $$

Molybdenum-99 is a radioactive isotope produced in a nuclear reactor by neutron bombardment of molybdenum-98.

$$ {}^{98}_{42}\mathrm{Mo} + {}^{1}_{0}\mathrm{n} \longrightarrow {}^{99}_{42}\mathrm{Mo} + \gamma $$

Molybdenum-99 is also produced as a fission product of uranium-235. Molybdenum-99 decays to technetium-99m that has several uses in nuclear medical imaging and treatment.

$$ {}^{99}_{42}\mathrm{Mo} \longrightarrow {}^{99m}_{43}\mathrm{Tc} + {}^{0}_{-1}\mathrm{e} $$

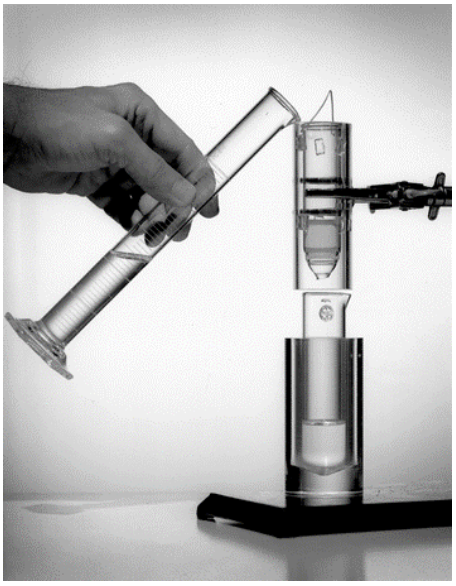


Figure 3.9.7.1: The first technetium-99m generator developed at Brookhaven National Laboratory, circa 1958, shown without shielding. A Tc-99m pertechnetate solution is being eluted from Mo-99 molybdate bound to a chromatographic substrate. Source: Brookhaven National Laboratory. / Public domain

Technetium-99m is short-lived (half-life 6 h), and needs to be produced in the hospital to minimize its decay during the transport. Its parent molybdenum-99 has a half-life of 66h and can be transported without significant decay during the transport. Molybdenum-99/technetium-99m generators are supplied to the hospitals in a shielded container. Fig. 8.7.3 illustrates the first Molybdenum-99/technetium-99m generator developed at Brookhaven National Laboratory. Molybdate (MoO_4^{2}) ion is adsorbed onto alumina adsorbent in a column. When molybdenum-99 decays to technetium-99m, the ion change to pertechnetate (TcO_4^{-}), which is less tightly bound to the alumina. Pouring a saline solution through the column elutes the technetium-99m as TcO_4^{-}ion, which is then used for medical purposes in the hospitals.

Destruction of an inoperable tumor has also been tested by α-emission from boron-10 upon neutron bombardment.

$$^{1}_{0}\mathrm{n} + {}^{10}_{5}\mathrm{B} \longrightarrow {}^{7}_{4}\mathrm{Li} + {}^{4}_{2}\mathrm{He}$$

This page titled 3.9.7: Making radioisotopes for medical uses is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

3.9.8: Nuclear fusion and fission

Conversion of matter into energy in nuclear reactions

When free nucleons come together to make a nucleoid, they release energy equal to the nuclear binding energy. The energy is produced at the expense of the mass of the nucleons following the famous Einstein equation:

$$E = mc^2$$

, where *E* is the energy, *m* is the mass, and c is the speed of light. In other words, the mass of nucleons in nucleoids is slightly less than the mass of the free nucleons, as some of their mass is released as nuclear binding energy.

The nuclear binding energy per nucleon is different for different nucleoids, as shown in Fig. 8.8.1. The composition of nucleoids changes during the nuclear reaction, and the difference in the nuclear binding energy is released as energy during the process, which is the source of energy in the sun, stars, nuclear power plants, and nuclear weapons.

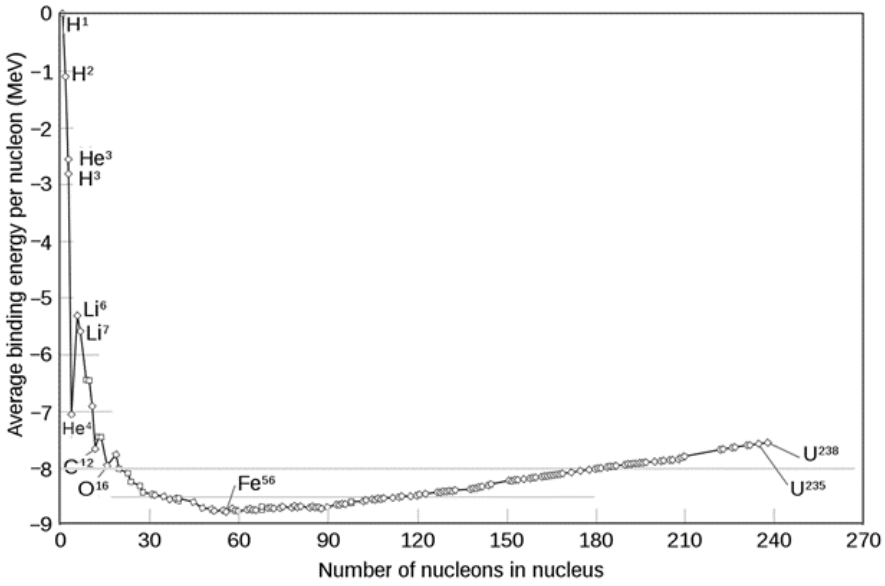


Figure 3.9.8.1: Graph of the binding energy released by an atom as a function of its nucleons, the energy is in megaelectron volts (MeV) where 1 Mev = 1.60218 10^{-13} J. Source: Modified from Fastfission/ Public domain

The most stable nucleoid with the highest average binding energy per nucleon is iron-56.

Nuclear fusion

The lighter nucleoids tend to combine and make heavier nucleoids that are more stable and lease a tremendous amount of energy –a process called nuclear fusions.

Nuclear fission

The heavier nucleoids, particularly those having a mass number higher than 92, tend to split into two smaller nucleoids that are more stable and lease a large amount of energy –a process called nuclear fission.

Both the fusion and fission processes release a tremendous amount of energy, called nuclear energy or atomic energy, which is the difference between the nuclear binding energy of the parent and the daughter nucleoid.

Nuclear fusion -a source of energy in the sun and stars

The universe is composed of about 98% hydrogen and helium. Sun is composed of about 74% hydrogen, 25% helium, and 1% all other elements. The light nucleoid like hydrogen and helium fuse to make heavier nucleoids and subatomic particles like neutrons in a process called nuclear fusion. Nuclear fusion releases energy that is the difference in the nuclear binding energy of their nucleons in the daughter and parent nucleoids. Fusion energy is the source of energy for the sun and stars. Some of the nuclear fusion reactions, illustrated in Fig. 8.1.1 are the following.

$$ {}_1^1\mathrm{H} + {}_1^1\mathrm{H} \longrightarrow {}_2^2\mathrm{H} + {}_1^0\mathrm{e} $$

$$ {}_1^1\mathrm{H} + {}_1^2\mathrm{H} \longrightarrow {}_2^3\mathrm{H} $$

$$ {}_2^3\mathrm{H} + {}_2^3\mathrm{H} \longrightarrow {}_2^4\mathrm{H} + 2\,{}_1^1\mathrm{H} $$

$$ {}_2^3\mathrm{H} + {}_1^1\mathrm{H} \longrightarrow {}_2^4\mathrm{H} + {}_1^0\mathrm{e} $$

Although hydrogen is abundant on earth and a potential source of energy, the controlled fusion reaction is not yet economically feasible. The main difficulty is that the fusion of nuclei requires extremely high temperature, greater than a hundred thousand Celsius, and extremely high pressure, greater than a hundred thousand atmospheres, to overcome the repulsive forces of the like charges of the nuclei before they can fuse. For this reason, nuclear fusion reactions are also called thermonuclear reactions. Such conditions exist in sun and stars where these reactions routinely happen, but there are many technical problems yet to overcome before the energy can be harnessed economically from the nuclear fusion reactions on earth. A nuclear explosion can create the conditions needed for the thermonuclear reactions and are used to carry out uncontrolled thermonuclear reactions to boost the explosive power of nuclear weapons. Such a nuclear weapon is called a thermonuclear or hydrogen bomb.

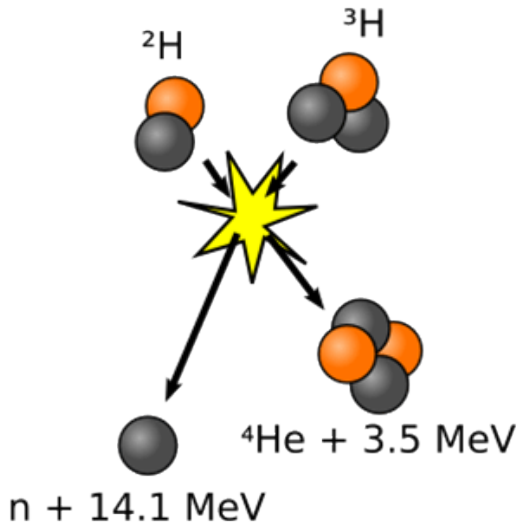


Figure 3.9.8.2: Fusion of deuterium with tritium creating helium-4, freeing a neutron, and releasing 17.59 MeV as kinetic energy of the products while a corresponding amount of mass disappears, in agreement with kinetic $E = \Delta mc^2$, where Δm is the decrease in the total rest mass of particles. Source: Wykis (talk · contribs), Public domain, via Wikimedia Commons

One of the potential candidate reactions for nuclear fusion reactors is the fusion of deuterium with tritium, illustrated in Fig. 8.8.2, which requires a little less harsh conditions. Most of the research is focused on the device called Tokamak, which uses strong magnetic fields to contain and heat the materials for the reaction, but so far it is not economically feasible to use it as a source of energy for commercial uses.

Nuclear fission -the source of energy in nuclear power plants and in nuclear weapons

Nuclear fission is a radioactive decay process in which a heavier nucleoid splits into two or lighter nucleoids and releases a

tremendous amount of energy that is the difference in the binding energy of nucleons in the daughter and the parent nucleoids. The nucleoids that do spontaneous fission are marked green in Fig. 8.2.1.

Nuclear fuel

Induced fission is fission that occurs upon bombardment of a nucleoid with a nuclear particle, like a neutron. For example, the bombardment of uranium-235 or plutonium-239 by a neutron causes the fission and releases about three neutrons along with the smaller nucleoids, as illustrated in Fig. 8.8.3. The neutrons released by the fission cause fission of another nucleoid starting a **nuclear chain reaction**. The nuclei like uranium-235 or plutonium-239 that produce neutrons in the product that are capable of continuing the fission in a chain reaction are called **fissile**. The fissile nucleoids are nuclear fuels used in nuclear power plants to produce nuclear energy.

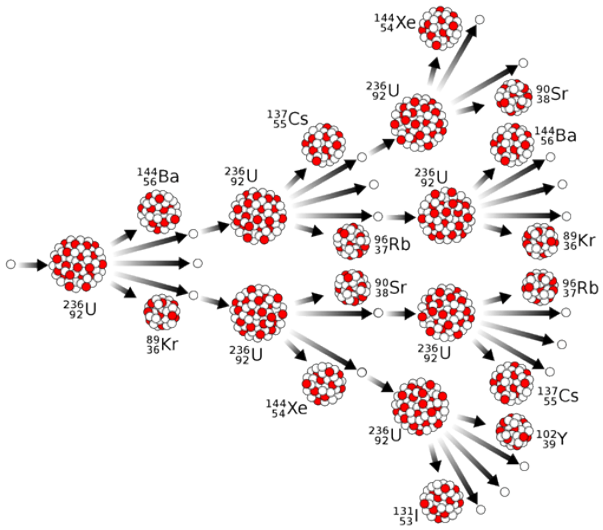


Figure 3.9.8.3: Illustration of a nuclear fission chain reaction. Source: MikeRun / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0)

Fate of neutron released in a fission event

The neutrons released in a fission event may be lost, absorbed by another nucleus like uranium-238 that does not fission, or may be absorbed by another fissile nucleoid and repeat the fission, as illustrated in Fig. 8.8.4. Three situations may arise, If, on the average;

1. less than one neutron released causes new fission, the process dyes,
2. one neutron released causes new fission, the fission continues in a controlled way as in nuclear power plants,
3. more than one neutron released causes new fission, the fission increases exponentially resulting in a nuclear explosion, as illustrated in Fig. 8.8.3.

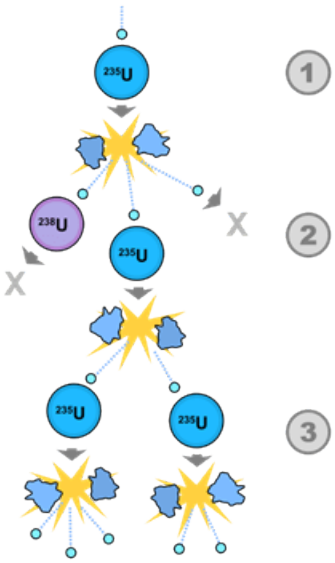


Figure 3.9.8.4: A schematic nuclear fission chain reaction. 1. A uranium-235 atom absorbs a neutron and fissions into two new atoms (fission fragments), releasing three new neutrons and some binding energy. 2. One of those neutrons is absorbed by an atom of uranium-238 and does not continue the reaction. Another neutron is simply lost and does not collide with anything, also not continuing the reaction. However, the one neutron does collide with an atom of uranium-235, which then fissions and releases two neutrons and some binding energy. 3. Both of those neutrons collide with uranium-235 atoms, each of which fissions and releases between one and three neutrons, which can then continue the reaction. Source: Fastfission/ Public domain

Note that each fission event does not necessarily split the parent nucleoid into exact same daughter nucleoids, it can split into different combinations of daughter nucleoids resulting in a mixture of products that are usually radioactive, as illustrated in Fig. 8.8.4.

Atomic bomb

A certain minimum mass of the fissile material, called **critical mass**, is necessary for a fission event to grow exponentially and cause a nuclear explosion. The critical mass is less for a more enriched fissile material. Therefore, usually enriched uranium-235 or plutonium 239 is used as an explosive in an atomic bomb. The mass more than the critical mass, called supercritical mass, is split into sub-critical portions which are kept separate in the weapon to avoid a nuclear explosion. When needed, the conventional chemical explosion is used to force the sub-critical portions to combine and make a supercritical mass for a nuclear explosion to happen, as illustrated in Fig. 8.8.5.

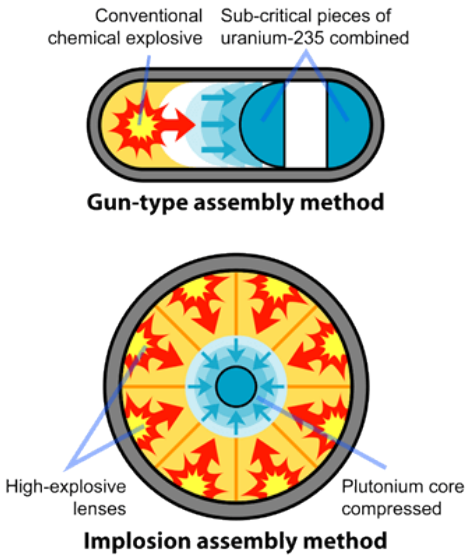


Figure 3.9.8.5: Schematic representation of the two methods with which to

assemble a fission bomb. Source: Fastfission/ Public domain

Nuclear power plant

Nuclear power plants, like the one illustrated in Fig. 8.8.6, harness the energy released during fission for electricity production. The fissile material is assembled in the form of fuel rods in the core of a nuclear reactor. It is important to sustain the nuclear chain reaction but it is critical to keep it under control so that the chain reaction may not grow exponentially and cause a nuclear explosion. Rods of material, like palladium or boron that are efficient neutron absorbers, are included in the design of the reactor core. The neutron absorber rods can be lowered or raised to control the neutron flux available to sustain the fission chain reaction.

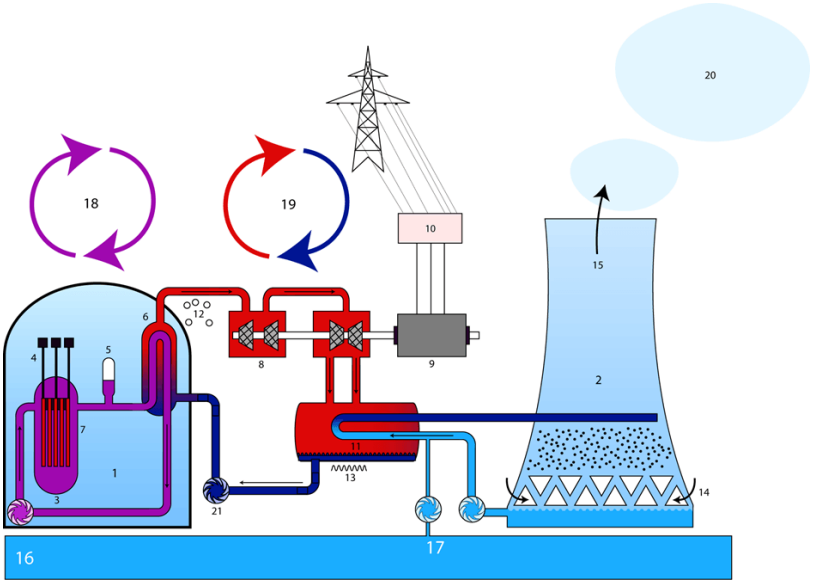


Figure 3.9.8.1: Design of a nuclear power plant with a pressurized water reactor (PWR).

1. reactor block, 2. cooling tower, 3. reactor, 4. control rod, 5. support for pressure, 6. steam generator, 7. fuel element, 8. turbine, 9. generator, 10. transformer, 11. condenser, 12. gaseous, 13. liquid, 14.air, 15.air (humid), 16. river, 17. cooling-water circulation, 18. primary circuit, 19. secondary circuit, 20. water vapor, 21. pump. Source: Steffen Kuntoff / CC BY-SA 2.0 DE (https://creativecommons.org/licenses...2.0/de/deed.en)

The reactor core is surrounded by a coolant and moderator, which may be regular water (H_2O) or heavy water (D_2O), or some other material. It is called coolant because it takes away the heat produced by the fission reaction and moderator because it also slows down the fast neutrons released in fission. The slow neutrons are more effective for the fission reaction. The coolant transfers its heat to a steam generator. All of these components are housed in a containment building to keep the radioactivity contained, even if a nuclear accident happens. The steam is used to drive a turbine for electricity production, and the condensed water is cooled using water from the cooling tower and returned for reuse, as illustrated in Fig. 8.8.6.

Nuclear power plants produce about 20% of the electricity in the USA. Fig. 8.8.7 shows a nuclear power plant in Arkansas. In some countries, the contribution of nuclear power to the total electricity is much higher, e.g., about 70% of electricity is produced by nuclear power plants in France. Nuclear power is considered an energy source of the future, or at least for a transition from fossil fuel to the next major energy source when the fossil fuel reserves will exhaust.



Figure 3.9.8.1: Russellville nuclear power plant in Arkansas, showing a cooling tower in the front with non-radioactive steam evaporating from it. Source: Edibobb, CC BY 3.0 <https://creativecommons.org/licenses/by/3.0>, via Wikimedia Commons

The fission products in the nuclear power plants become **radioactive waste** that needs to be stored for at least 10 half-lives to reach an acceptable radioactivity level. Based on the 28.8 years half-life of strontium-90 which is the long-lived and dangerous product in nuclear waste, a storage time of 300 years is needed. Handling radioactive waste is an issue of concern associated with nuclear power production. Plutonium-239 is another radioactive isotope produced from neutron absorption by uranium-238. Uranium-238 is a major constituent (99.28%) in the natural uranium along with fissile uranium-235 (0.71%). Plutonium-239 has a long half-life (24,000 Years) but it can be extracted in a fuel reprocessing plant and used as a nuclear fuel.

This page titled 3.9.8: Nuclear fusion and fission is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

4: GEOLOGY

Chapter 4.1: Science of Geology

4.1.1: What is Geology?

4.1.2: The Science of Geology

4.1.3: The Branches of Geology

4.1.4: Scientific Processes in Geology

4.1.5: The Nature of Science

4.1.6: Scientific Tools

4.1.7: Geologic Tools

4.1.8: Maps

4.1.9: Location and Direction

4.1.10: Scientific Models

4.1.11: Geologic Modelling

Chapter 4.2: Rock Forming Minerals

4.2.1: The Building Blocks of Matter

4.2.2: Atomic Bonds

4.2.3: Physical Characteristics of Minerals

4.2.4: Classifying Minerals

4.2.5: Abundance of Elements in Earth's Crust

4.2.6: Silicate Minerals and the Silicate Tetrahedron

Chapter 4.3: Rocks and the Rock Cycle

4.3.1: Characteristics of Igneous Rocks

4.3.2: How Are Igneous Rocks Classified?

4.3.3: Identifying Igneous Rocks

4.3.4: Uses of Igneous Rocks

4.3.5: Characteristics of Sedimentary Rocks

4.3.6: Characteristics of Metamorphic Rocks

Chapter 4.4: Weathering and Soil Formation

4.4.1: What Is Weathering?

4.4.2: Mechanical Weathering

4.4.3: Chemical Weathering

4.4.4: Influences on Weathering

4.4.5: Soil Texture and Composition

4.4.6: Soil Horizons and Profiles

4.4.7: Types of Soils

4.4.8: Causes of Soil Erosion

4.4.9: Avoiding Soil Loss

Chapter 4.5: Plate Tectonics

4.5.1: Continental Drift

4.5.2: Wegener and the Continental Drift Hypothesis

4.5.3: Theory of Plate Tectonics

4.5.4: Developing the Theory

4.5.5: Supercontinents

4.5.6: Volcanic Arcs

4.5.7: Hot Spots

Chapter 4.6: Earth's Interior

4.6.1: Characteristics of the Crust

4.6.2: Characteristics of the Mantle

4.6.3: Characteristics of the Inner and Outer Core

4.6.4: The Lithosphere and Asthenosphere

4.6.5: Studying the Earth's Interior

4.6.6: Understanding the Earth's Interior

Chapter 4.7: Crustal Deformation

4.7.1: Stress In Earth's Crust

4.7.2: Strain

4.7.3: Folds

4.7.4: Faults

4.7.5: Mountains

Chapter 4.8: Earthquakes

4.8.1: The Nature of Earthquakes

4.8.2: Causes of Earthquakes

4.8.3: Common Locations of Earthquakes

4.8.4: The Ring of Fire

4.8.5: Body Waves

4.8.6: Surface Waves

4.8.7: Measuring Earthquakes

4.8.8: Magnitude versus Intensity

4.8.9: Earthquake Hazards

4.8.10: Earthquake Prediction

4.8.11: Seismicity and Earthquake Prediction

Chapter 4.9: Volcanoes

4.9.1: Types of Volcanoes

4.9.2: Types of Eruptions

4.9.3: Supervolcanoes

4.9.4: Volcanoes at Plate Boundaries

4.9.5: Volcanoes Hotspots

4.9.6: Pyroclastic Deposits
4.9.7: Dikes and Sills
4.9.8: Volcanic Landforms
4.9.9: Igneous Features
4.9.10: Volcanic Gases, Pyroclastic Flow, and Tephra
4.9.11: Lahars, Landslides, and Lavaflow
4.9.12: Signs of an Eruption
4.9.13: Tools to Monitor Volcanoes

Chapter 4.10: Mass Movement

4.10.1: Slides
4.10.2: Falls
4.10.3: Slump and Creep
4.10.4: Natural Mass Wasting
4.10.5: Human Impact on Mass Wasting

Chapter 4.11: Hydrology

4.11.1: Phases of the Hydrologic Cycle
4.11.2: Types of Streams and Rivers
4.11.3: Drainage Basins
4.11.4: Geologic Processes and Flowing Water
4.11.5: Groundwater
4.11.6: Porosity and Permeability
4.11.7: Aquifers and Well-Potential
4.11.8: Groundwater Withdrawal
4.11.9: Quality of Groundwater
4.11.10: Karst Topography
4.11.11: Geysers
4.11.12: Springs
4.11.13: Geologic Features
4.11.14: Oceans and Shorelines
4.11.15: Ocean Movement
4.11.16: Shorelines
4.11.17: Coastal Hazards
4.11.18: Human Modifications of Coastal Processes

Chapter 4.12: Geological Resources

4.12.1: Ore Genesis
4.12.2: Petroleum
4.12.3: Natural Gas
4.12.4: Coal
4.12.5: Mining
4.12.6: Effects of Mining

4.12.7: Earth’s Energy
4.12.8: Energy Consumption on Earth
4.12.9: Use and Conservation of Resources

This page titled 4: Geology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle.

CHAPTER 4.1: SCIENCE OF GEOLOGY

Learning Objectives

Define geology as a science and identify methods used in its study.

- Define the science of geology.
- Identify and use scientific processes and the scientific methods as used by geologists and with other related scientific disciplines.
- List tools and concepts commonly used by geologists.

4.1.1: What is Geology?

4.1.2: The Science of Geology

4.1.3: The Branches of Geology

4.1.4: Scientific Processes in Geology

4.1.5: The Nature of Science

4.1.6: Scientific Tools

4.1.7: Geologic Tools

4.1.8: Maps

4.1.9: Location and Direction

4.1.10: Scientific Models

4.1.11: Geologic Modelling

This page titled Chapter 4.1: Science of Geology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.1: What is Geology?

Define the science of geology.

This section will introduce you to the science of geology. You will not only learn what geology is but you will also learn the various fields and concentrations of geology.

What You'll Learn to Do

- Describe "geology" and identify some of its main principles.
- Identify the main branches of geology.

Learning Activities

The learning activities for this section include the following:

- Reading: The Science of Geology
- Reading: The Branches of Geology
- Self Check: What is Geology?

Contributors and Attributions

CC licensed content, Original

- **Authored by**: Lumen Learning and SBCTC. **License**: *CC BY: Attribution*

This page titled 4.1.1: What is Geology? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.2: The Science of Geology

Why Do People Study Geology?

A lot of people are attracted to geology and other earth sciences because they love to be outdoors. These people wonder how the magnificent rock formations that they see, like those in Yosemite in California (see figure 1), were formed. They want to study the processes that create and modify landforms.



Figure 1. Half dome in Yosemite, National Park, California, USA

But earth science is not just about what we can see with the naked eye—the molten lava, icy mountain peaks, steep canyons, and towering waterfalls. Some people want to go deeper, to learn about what drives the surface processes and other features of the planet; for example, why does Earth have a magnetic field? These people are interested in learning about the layers of material that lie beneath the surface, the mantle and the core. Since more than 70 percent of Earth is covered with oceans, it's not surprising that many people wonder what lies within and at the bottom of the seas.

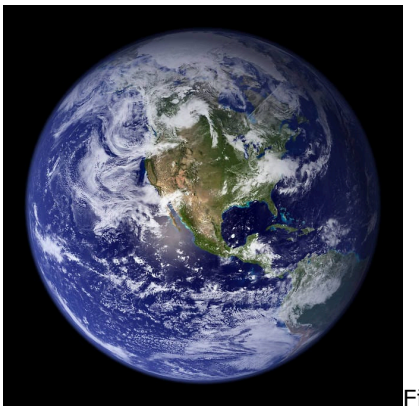


Figure 2. The "Blue Marble"

Some people look up and wonder what lies beyond our skies. These people are interested in applying what we know about Earth to our more distant surroundings. They want to understand our near neighbors, the planets and satellites of our Solar System, and objects that lie far beyond.

What is Geology?

In its broadest sense, geology is the study of Earth—its interior and its exterior surface, the rocks and other materials that are around us, the processes that have resulted in the formation of those materials, the water that flows over the surface and lies underground, the changes that have taken place over the vastness of geological time, and the changes that we can anticipate will take place in the near future. Geology is a science: we use deductive reasoning and scientific methods to understand geological problems.

Geology is arguably the most integrated of all of the sciences because it involves the understanding and application of all of the other sciences: physics, chemistry, biology, mathematics, astronomy, and others. But unlike most of the other sciences, geology has an extra dimension, that of time—billions of years of it. Geologists study the evidence that they see around them, but in most cases, they are observing the results of processes that happened thousands, millions, and even billions of years in the past. Those were processes that took place at incredibly slow rates—millimeters per year to centimeters per year—but because of the amount of time available, they produced massive results.


Figure 3. Rearguard Mountain and Robson Glacier in the Rocky Mountains of British Columbia

Geology is displayed on a grand scale in mountainous regions, perhaps nowhere better than the Rocky Mountains in Canada (Figure 3). The peak on the right is Rearguard Mountain, which is a few kilometers northeast of Mount Robson, the tallest peak in the Canadian Rockies (3,954 m). The large glacier in the middle of the photo is the Robson Glacier. The river flowing from Robson Glacier drains into Berg Lake in the bottom right.

The sedimentary rock that these mountains are made of formed in ocean water over 500 million years ago. A few hundred million years later, these beds were pushed east for tens to hundreds of kilometers by tectonic plate convergence and also pushed up to thousands of meters above sea level. Over the past two million years this area—like most of the rest of Canada—has been repeatedly glaciated, and the erosional effects of those glaciations are obvious. The Robson Glacier is now only a small remnant of its size during the Little Ice Age of the fifteenth to eighteenth centuries, as shown by the distinctive line on the slope on the left. Like almost all other glaciers in the world, it is now receding even more rapidly because of human-caused climate change.

Geology is also about understanding the evolution of life on Earth; about discovering resources such as metals and energy; about recognizing and minimizing the environmental implications of our use of those resources; and about learning how to mitigate the hazards related to earthquakes, volcanic eruptions, and slope failures. All of these aspects of geology, and many more, are covered in this course.

Contributors and Attributions

CC licensed content, Original

- Revision and adaptation. **Provided by**: Lumen Learning. **License**: *CC BY-NC: Attribution-NonCommercial*

CC licensed content, Shared previously

- 1.2: Earth Science and Its Branches. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-Middle-School/section/1.2/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Chaptert2: Studying Earth Science. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-Concepts-For-High-School/section/2.0/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Half Dome in Yosemite National Park, California, USA. **Authored by**: Heiko von Rauu00dfendorff. **Located at**: https://commons.wikimedia.org/wiki/File:03_yosemite_half_dome.jpg. **License**: *CC BY-SA: Attribution-ShareAlike*

- 1.1 What Is Geology?. **Authored by**: Steven Earle. **Located at**: https://opentextbc.ca/geology/chapter/1-1-what-is-geology/. **Project**: Physical Geology. **License**: *CC BY: Attribution*

Public domain content

- Blue Marble 2002. **Provided by**: NASA. **Located at**: http://www.nasa.gov/content/goddard/earth-from-space-15-amazing-things-in-15-years. **License**: *Public Domain: No Known Copyright*

This page titled 4.1.2: The Science of Geology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.3: The Branches of Geology

As we mentioned, there are many varieties of geology. There is so much to know about our home planet that most geologists become specialists in one area. These specialties are known as **branches of geology**, and have specific titles. For example, a mineralogist studies minerals while a seismologist monitors earthquakes to help protect people and property from harm (figure 1).

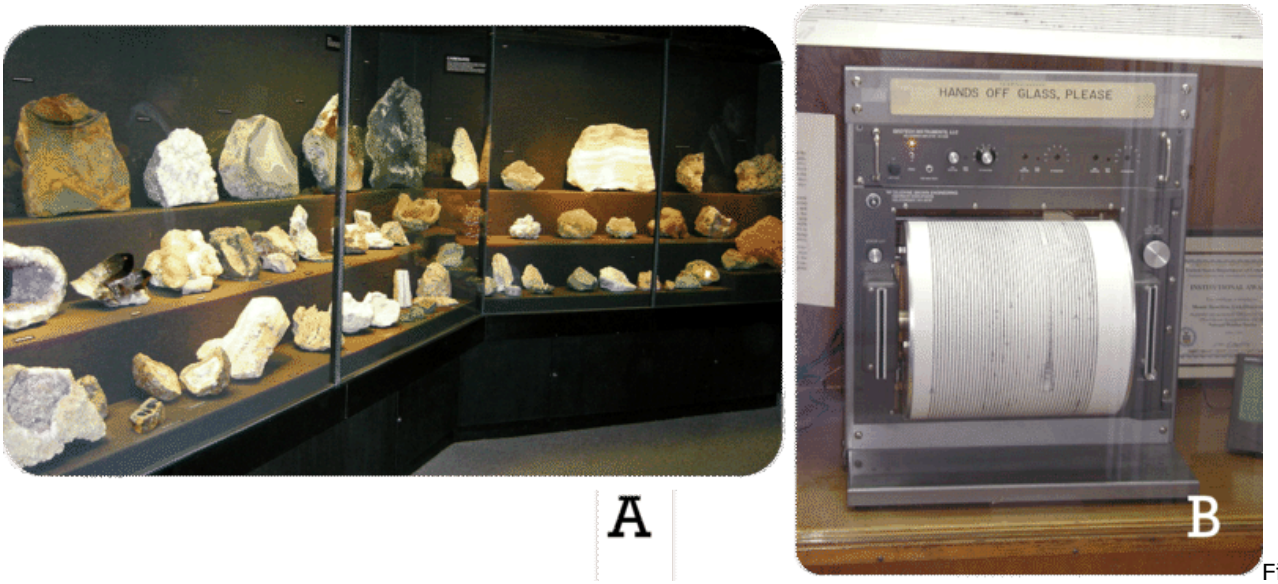


Figure 1. (A) Mineralogists focus on all kinds of minerals. (B) Seismographs are used to measure earthquakes and pinpoint their origins.



Figure 2. These folded rock layers have bent over time. Studying rock layers helps scientists to explain these layers and the geologic history of the area.

Volcanologists brave molten lava to study volcanoes. Scientists who compare the geology of other planets to Earth are planetary geologists. Some geologists study the Moon. Others look for petroleum. Still others specialize in studying soil. Some geologists can tell how old rocks are and determine how different rock layers formed (figure 2). There is probably an expert in almost anything you can think of related to Earth!

Geologists might study rivers and lakes, the underground water found between soil and rock particles, or even water that is frozen in glaciers. Earth scientists also need geographers who explore the features of Earth's surface and work with cartographers, who make maps. Studying the layers of rock beneath the surface helps us to understand the history of planet Earth.

Some Branches of Geology

As you've seen, different branches of geology study one particular part of earth. Since all of the branches are connected, specialists work together to answer complicated questions. Let's look at some other important branches of geology.

Geochemistry

Geochemistry is the study of the chemical processes which form and shape the Earth. It includes the study of the cycles of matter and energy which transport the Earth's chemical components and the interaction of these cycles with the hydrosphere and the atmosphere.

It is a subfield of inorganic chemistry, which is concerned with the properties of all the elements in the periodic table and their compounds. Inorganic chemistry investigates the characteristics of substances that are not organic, such as nonliving matter and minerals found in the Earth's crust.

Oceanography

Oceanography is the study of the composition and motion of the water column and the processes which are responsible for that motion. The principal oceanographic processes influencing continental shelf waters include waves and tides as well as wind-driven and other oceanic currents. Understanding the oceanography of shelf waters and the influence this has on seabed dynamics, contributes to a wide range of activities such as the following:

- assessment of offshore petroleum production infrastructure
- seabed mapping and characterisation for environmental management
- marine biodiversity surrogacy research
- assessment of renewable energy potential

Paleontology

Paleontologists are interested in fossils and how ancient organisms lived. Paleontology is the study of fossils and what they reveal about the history of our planet. In marine environments, microfossils collected within in layers of sediment cores provide a rich source of information about the environmental history of an area.

Sedimentology

Sedimentology is the study of sediment grains in marine and other deposits, with a focus on physical properties and the processes which form a deposit. Deposition is a geological process where geological material is added to a landform. Key physical properties of interest include:

- the size and shape of sediment grains
- the degree of sorting of a deposit
- the composition of grains within a deposit
- sedimentary structures.

These properties together provide a record of the mechanisms active during sediment transportation and deposition which allows the interpretation of the environmental conditions that produced a sediment deposit, either in modern settings or in the geological record.

Additional Branches

- **Benthic Ecology.** Benthic ecology is the study of living things on the seafloor and how they interact with their environment.
- **Biostratigraphy.** Biostratigraphy is the branch of stratigraphy that uses fossils to establish relative ages of rock and correlate successions of sedimentary rocks within and between depositional basins.
- **Geochronology.** Geochronology is a discipline of geoscience which measures the age of earth materials and provides the temporal framework in which other geoscience data can be interpreted in the context of Earth history.

- **Geophysics.** Information relating to various techniques including: airborne electromagnetics, gravity, magnetics, magnetotellurics, radiometrics, rock properties and seismic.
- **Marine Geochemistry.** Marine geochemistry is the science used to help develop an understanding of the composition of coastal and marine water and sediments.
- **Marine Geophysics.** Marine geophysics is a scientific discipline which uses the quantitative observation of physical properties to understand the seafloor and sub-seafloor geology.
- **Marine Surveying.** The survey environment varies from oceanographic studies in the water column to investigating sediment and geochemical processes on the seafloor and imaging the sub-seafloor rocks. Surveys are carried over Australia's entire marine jurisdiction, from coastal estuaries and bays, across the continental shelf and slope, to the deep abyssal plains.
- **Spectral Geology.** Spectral geology is the measurement and analysis of portions of the electromagnetic spectrum to identify spectrally distinct and physically significant features of different rock types and surface materials, their mineralogy and their alteration signatures.

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- 1.2: Earth Science and Its Branches. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-Middle-School/section/1.2/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Disciplines and Techniques. **Provided by**: Geoscience Australia. **Located at**: http://www.ga.gov.au/scientific-topics/disciplines. **License**: *CC BY: Attribution*

This page titled 4.1.3: The Branches of Geology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.4: Scientific Processes in Geology

Identify and use scientific processes and the scientific methods as used by geologists and with other related scientific disciplines.

This section discusses the importance of the scientific process as it relates to geology and other scientific disciplines.

What You'll Learn to Do

- Order the steps of the scientific process.
- Form a hypothesis and use it to design a scientific experiment.

Learning Activities

The learning activities for this section include the following:

- Reading: The Nature of Science
- Self Check: Scientific Processes in Geology

Contributors and Attributions

CC licensed content, Original

- **Authored by**: Lumen Learning and SBCTC. **License**: *CC BY: Attribution*

This page titled 4.1.4: Scientific Processes in Geology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.5: The Nature of Science

Sometime in your life you've asked a question about the world around you. Probably you've asked a lot of questions over the years. The best way to answer questions about the natural world is by using science. Scientists ask questions every day, and then use a set of steps to answer those questions. The steps are known as the scientific method. By following the scientific method, scientists come up with the best information about the natural world. As a scientist, you need to do experiments to find out about the world. You also need to wonder, observe, talk, and think. Everything we learn helps us to ask new and better questions.

Scientific Method

The scientific method is a set of steps that help us to answer questions. When we use logical steps and control the number of things that can be changed, we get better answers. As we test our ideas, we may come up with more questions. The basic sequence of steps followed in the scientific method is illustrated in figure 1.

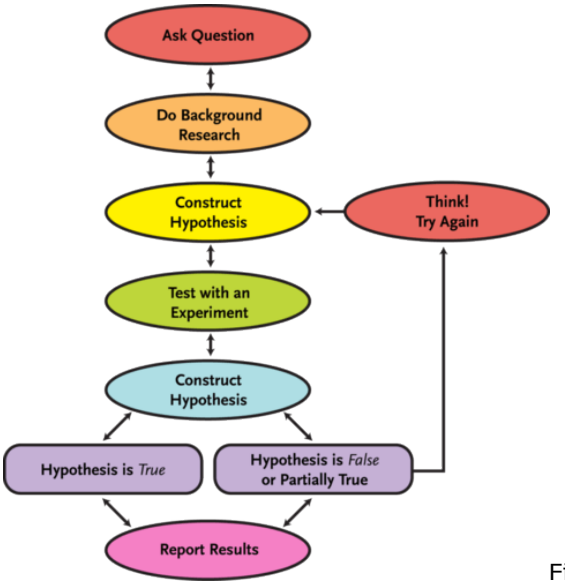


Figure 1. The Scientific Method

Questions



Figure 2. Soil is often lost from ground that has been plowed.

Asking a question is one really good way to begin to learn about the natural world. You might have seen something that makes you curious. You might want to know what to change to produce a better result.

Let's say a farmer is having an erosion problem. She wants to keep more soil on her farm. The farmer learns that a farming method called "no-till farming" allows farmers to plant seeds without plowing the land. She wonders if planting seeds without plowing will reduce the erosion problem and help keep more soil on her farmland. Her question is this: "Will using the no-till method of farming help me to lose less soil on my farm?" (figure 2).

Research



Figure 3. Rather than breaking up soil like in this picture, the farmer could try no-till farming methods.

Before she begins, the farmer needs to learn more about this farming method. She can look up information in books and magazines in the library. She may also search the Internet. A good way for her to learn is to talk to people who have tried this way of farming. She can use all of this information to figure out how she is going to test her question about no-till farming. Farming machines are shown in the figure 3.

Hypothesis

After doing the research, the farmer will try to answer the question. She might think, "If I don't plow my fields, I will lose less soil than if I do plow the fields. Plowing disrupts the soil and breaks up roots that help hold soil in place." This answer to her question is a **hypothesis**. A hypothesis is a reasonable explanation. A hypothesis can be tested. It may be the right answer, it may be a wrong answer, but it must be testable. Once she has a hypothesis, the next step is to do experiments to test the hypothesis. A hypothesis can be proved or disproved by testing. If a hypothesis is repeatedly tested and shown to be true, then scientists call it a **theory**.

Experiment

When we design experiments, we choose just one thing to change. The thing we change is called the **independent variable**. In the example, the farmer chooses two fields and then changes only one thing between them. She changes how she plows her fields. One field will be tilled and one will not. Everything else will be the same on both fields: the type of crop she grows, the amount of water and fertilizer that she uses, and the slope of the fields she plants on. The fields should be facing the same direction to get about the same amount of sunlight. These are the experimental **controls**. If the farmer only changes how she plows her fields, she can see the impact of the one change. After the experiment is complete, scientists then measure the result. The farmer measures how much soil is lost from each field. This is the **dependent variable**. How much soil is lost from each field "depends" on the plowing method.

Data and Experimental Error


Figure 4. A pair of farmers take careful measurements in the field.

During an experiment, a scientist collects data. The data might be measurements, like the farmer is taking in figure 4. The scientist should record the data in a notebook or onto a computer. The data is kept in charts that are clearly labeled. Labeling helps the scientist to know what each number represents. A scientist may also write descriptions of what happened during the experiment. At the end of the experiment the scientist studies the data. The scientist may create a graph or drawing to show the data. If the scientist can picture the data the results may be easier to understand. Then it is easier to draw logical conclusions.

Even if the scientist is really careful it is possible to make a mistake. One kind of mistake is with the equipment. For example, an electronic balance may always measure one gram high. To fix this, the balance should be adjusted. If it can't be adjusted, each measurement should be corrected. A mistake can come if a measurement is hard to make. For example, the scientist may stop a stopwatch too soon or too late. To fix this, the scientist should run the experiment many times and make many measurements. The average of the measurements will be the accurate answer. Sometimes the result from one experiment is very different from the other results. If one data point is really different, it may be thrown out. It is likely a mistake was made in that experiment.

Conclusions

The scientist must next form a conclusion. The scientist must study all of the data. What statement best explains the data? Did the experiment prove the hypothesis? Sometimes an experiment shows that a hypothesis is correct. Other times the data disproves the hypothesis. Sometimes it's not possible to tell. If there is no conclusion, the scientist may test the hypothesis again. This time he will use some different experiments. No matter what the experiment shows the scientist has learned something. Even a disproved hypothesis can lead to new questions.

The farmer grows crops on the two fields for a season. She finds that 2.2 times as much soil was lost on the plowed field as compared to the unplowed field. She concludes that her hypothesis was correct. The farmer also notices some other differences in the two plots. The plants in the no-till plots are taller. The soil moisture seems higher. She decides to repeat the experiment. This time she will measure soil moisture, plant growth, and the total amount of water the plants consume. From now on she will use no-till methods of farming. She will also research other factors that may reduce soil erosion.

Theory

When scientists have the data and conclusions, they write a paper. They publish their paper in a scientific journal. A journal is a magazine for the scientists who are interested in a certain field. Before the paper is printed, other scientists look at it to try to find mistakes. They see if the conclusions follow from the data. This is called peer review. If the paper is sound it is printed in the journal.

Other papers are published on the same topic in the journal. The evidence for or against a hypothesis is discussed by many scientists. Sometimes a hypothesis is repeatedly shown to be true and never shown to be false. The hypothesis then becomes a

theory. Sometimes people say they have a ”theory” when what they have is a hypothesis.

In science, a theory has been repeatedly shown to be true. A theory is supported by many observations. However, a theory may be disproved if conflicting data is discovered. Many important theories have been shown to be true by many observations and experiments and are extremely unlikely to be disproved. These include the theory of plate tectonics and the theory of evolution.

Contributors and Attributions

CC licensed content, Shared previously

- 1.1: The Nature of Science. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-Middle-School/section/1.1/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.1.5: The Nature of Science is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.6: Scientific Tools

List tools and concepts commonly used by geologists.

This section discusses some of the common tools and concepts geologists employ.

What You'll Learn to Do

- Discuss several tools and their use in geology.
- Understand how to read various maps.
- Find a location using longitude and latitude.
- Identify commonly used geological models.

Learning Activities

The learning activities for this section include the following:

- Reading: Geologic Tools
- Reading: Maps
- Video: Exploring Topographic Maps
- Reading: Location and Direction
- Reading: Scientific Models
- Reading: Geologic Modelling
- Self Check: Scientific Tools

This page titled 4.1.6: Scientific Tools is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.7: Geologic Tools

Geologists use a lot of tools to aid their studies. Some of the most common tools used are compasses, rock hammers, hand lenses, and field books.

Compasses

There are a number of different (specialised) magnetic compasses used by geologists to measure orientation of geological structures, as they map in the field, to analyse (and document) the geometry of bedding planes, joints, and/or metamorphic foliations and lineations.[1] In this aspect the most common device used to date is the analogue compass.

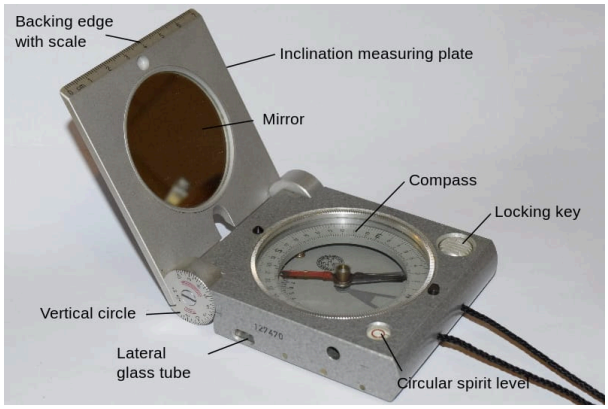
Classic geological compasses

Classic geological compasses that are of practical use combine two functions, direction finding and navigation (especially in remote areas), and the ability to measure strike and dip of bedding surfaces and/or metamorphic foliation planes. Structural geologists (i.e. those concerned with geometry and the pattern of relative movement) also have a need to measure the plunge and plunge direction of lineations.

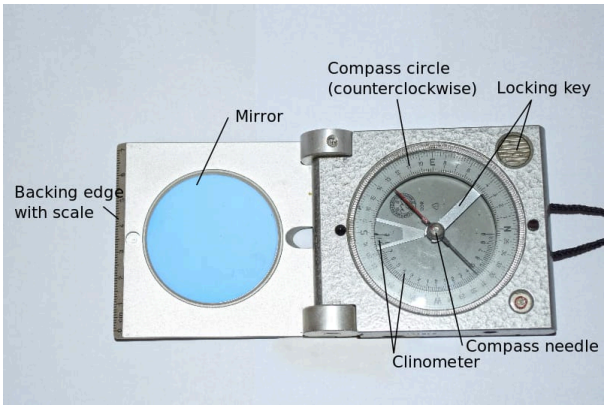
Compasses in common use include the Brunton compass and the Silva compass.

Modern Geological Compasses

The concept of modern geological compass was developed by Eberhard Clar of the University of Vienna during his work as structural geologist. He published it in 1954.[2] An advantage of his concept is that strike and dip is measured in one step, using the vertical circle for dip angle and the compass for the strike direction. The first implementation was done by the VEB Freiberger Präzisionsmechanik in Freiberg, Germany. The details of the design were made in a close cooperation with the Freiberg University of Mining and Technology.[3]



Setup of a modern geological compass after Prof. Clar (Freiberger), total view

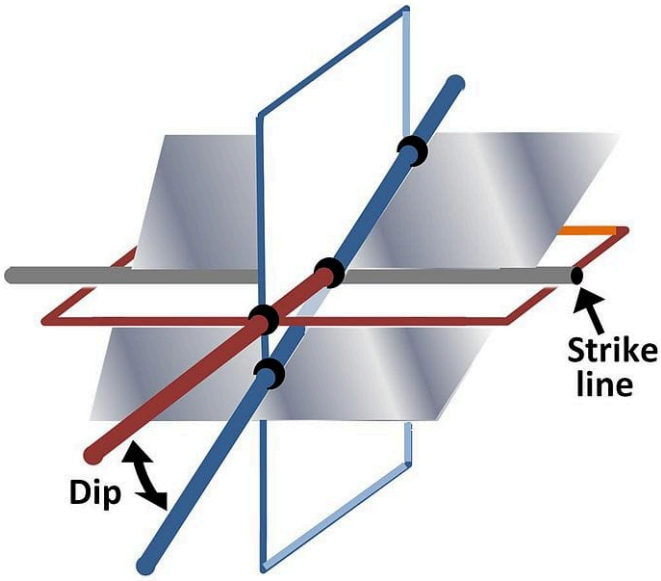


top view



bottom side

Usage



Strike line and dip of a plane describing attitude relative to a horizontal plane and a vertical plane perpendicular to the strike line

At first sight it appears confusing to the novice user, for the numbers on the compass dial ascend in an anticlockwise direction. This is because the compass is used to determine dip and dip-direction of surfaces (foliations), and plunge and plunge-direction of lines (lineations). To use the compass one aligns the lid of the compass with the orientation of the surface to be measured (to obtain dip and dip direction), or the edge of the lid of the compass with the orientation of the line (to obtain plunge and plunge direction). The compass must be twisted so that the base of the compass becomes horizontal, as accomplished using the spirit level incorporated in it. The needle of the compass is then freed by using the side button, and allowed to spin until the damping action slows its movement, and then stabilises. The side button is released and the needle is then firmly held in place, allowing the user thereafter to conveniently read the orientation measured. One first reads the scale that shows the angle subtended by the lid of the compass, and then depending on the colour shown (red or black) the end of the compass needle with the corresponding colour. Data are then recorded as (for example) 25°->333° (dip and dip-direction) or (plunge and plunge-direction).

This compass has the most use by structural geologists, measuring foliation and lineation in metamorphic rocks, or faults and joints in mining areas.

Digital Compasses

With the advent of the smartphone, geological compass programs based on the 3-axis teslameter and the 3-axis accelerometer have also begun to appear. These compass programs use vector algebra to compute plane and lineation orientations from the accelerometer and magnetometer data, and permit rapid collection of many measurements. However, some problems are potentially present. Smartphones produce a strong magnetic field of their own which must be compensated by software; as well, because the Earth's magnetic field fluctuates rapidly, measurements made by smartphone geological compasses can potentially be susceptible to considerable noise. Users of a smartphone compass should carefully calibrate their devices and run several tests against traditional magnetic compasses in order to understand the limitations of their chosen program.

Rock Hammers


A geologist's hammer used to break up rocks, as well as a scale in the photograph

A **geologist's hammer**, **rock hammer**, **rock pick**, or **geological pick** is a hammer used for splitting and breaking rocks. In field geology, they are used to obtain a fresh surface of a rock to determine its composition, nature, mineralogy, history, and field estimate of rock strength. In fossil and mineral collecting, they are employed to break rocks with the aim of revealing fossils inside. Geologist's hammers are also sometimes used for scale in a photograph.

Shape

Geologist's hammers, as with most hammers, have two heads, one on either side. Most commonly, the tool consists of a flat head on one end, with either a chisel or a pick head at the other end.[4]

- A chisel head (pictured), which is shaped like a chisel, is useful for clearing covering vegetation from exposures and is sometimes (though inadvisedly) used to pry open fissures. Some rocks can be easily split, like slate or shale, to reveal any fossils.
- A pick head, which terminates in a sharp point to deliver maximum pressure, is often preferred for harder rocks. A geologist's hammer bearing a pick end is often referred to as a rock pick or geological pick instead of a geologist's hammer.
- A flat head is used to deliver a blow to a rock with the intention of splitting it. Specimens or samples can be trimmed to remove sharp corners or reduce them in size.

Construction


A geologist's hammer with tubular shaft and chisel head

The effective power of a geologist's hammer is mainly considered to be a reflection of its head weight and handle length. Head weight may range from 8 oz (225 g) or less on a small hammer—such as would generally be used for casual use or by children—to 24 oz (680 g) and greater. A hammer of 16 oz (450 g) is often quoted as sufficient for all rock types, although metamorphic or igneous rocks often require heavier hammers for a more powerful blow.

The best geologist's hammers are forged from one piece of hardened steel, which renders them sturdy and long-lasting. Alternatives such as tubular- and wooden-shafted hammers are more commonly used, in part due to their lower cost. Such alternative handles sacrifice strength and make the hammer unsuitable for high-strain activities such as prying.

The form and weighting of the shaft defines the balance, which itself defines the ease, efficiency, and comfort of use of the geologist's hammer.

Hand Lenses


Loupe used by a geologist

The hand lens is a vital geological field tool used to identify small mineral crystals and structures in rocks. It is a simple, small magnification device used to see small details more closely. Unlike a magnifying glass, a loupe does not have an attached handle, and its focusing lens(es) are contained in an opaque cylinder or cone or fold into an enclosing housing that protects the lenses when not in use.

Three basic types of hand lenses exist:

- Simple lenses, which result in the highest degree of optical aberration and are generally lower magnification.
- Multiple lenses, generally higher magnification because of the reduced optical aberration.
- Prismatic, Multiple lenses with prisms used to change the perspective.

Jewelers typically use a monocular, handheld loupe in order to magnify gemstones and other jewelry that they wish to inspect.[5] A 10x magnification is good to use for inspecting jewelry and hallmarks[6] and is the Gemological Institute of America's standard for grading diamond clarity. Stones will sometimes be inspected at higher magnifications than 10x, although the depth of field, which is the area in focus, becomes too small to be instructive.[7] The accepted standard for grading diamonds is therefore that inclusions and blemishes visible at 10x impact the clarity grade.[8]

Field Books

Field books are used to take fieldnotes; they can be anything from a composition type notebook to a spiral, but most use an actual "field book" like those available for purchase here. **Fieldnotes** refer to qualitative notes recorded by scientists during or after their observation of a specific phenomenon they are studying. They are intended to be read as evidence that gives meaning and aids in the understanding of the phenomenon. Fieldnotes allow the researcher to access the subject and record what they observe in an unobtrusive manner.

One major disadvantage of taking fieldnotes is that they are recorded by an observer and are thus subject to (a) memory[9] and (b) possibly, the conscious or unconscious bias of the observer. It is best to record fieldnotes immediately after leaving the site to avoid forgetting important details.

Fieldnotes are particularly valued in geology and other descriptive sciences such as ethnography, biology, and archaeology.

Structure

There are two components of fieldnotes: descriptive information and reflective information.[10]

- Descriptive information is factual data that is being recorded. Factual data includes time and date, the state of the physical setting, social environment, descriptions of the subjects being studied and their roles in the setting, and the impact that the observer may have had on the environment.[11]
- Reflective information is the observer's reflections about the observation being conducted. These reflections are ideas, questions, concerns, and other related thoughts.[12]

Fieldnotes can also include sketches, diagrams, and other drawings. Visually capturing a phenomenon requires the observer to pay more attention to every detail as to not overlook anything.[13]

---

1. *The Mapping of Geological Structures* (Geological Society of London Handbook Series) [Paperback] K. R. McClay; *Statistics of Earth Science Data: Their Distribution in Time, Space and Orientation* [Paperback] Graham J. Borradaile (Author) ↩
2. Clar, E.: *A dual-circle geologist's and miner's compass for the measurement of areal and linear geological elements*. Separate print from the negotiations of the Federal Institute of Geology Vienna, 1954, vol. 4 ↩
3. "Geologist's Compass: Operating Manual," *Freiberger Präzisionsmechanik*. http://www.fpm.de/downloads/GeologistCompass_eng.pdf. ↩
4. MartinS (2006). "'Geological Hammers' info page"↩
5. "Jewelry - How to Use a Loupe - Using Jewelry Magnifiers". Jewelry.About.com ↩
6. *Ibid*. ↩
7. http://www.gemsociety.org/article/10x-loupe-the-gemologists-best-friend/ ↩
8. "The 4C's Of Diamonds: Diamond Clarity". Leibish & Co. ↩
9. Canfield, Michael (2011). *Field Notes on Science & Nature*. Harvard University Pres. p. 21. ↩
10. Labaree, Robert V. "Research Guides: Organizing Your Social Sciences Research Paper: Writing Field Notes". *libguides.usc.edu*. Retrieved 2016-04-12. ↩
11. *Ibid*. ↩
12. *Ibid*. ↩
13. Canfield,Michael (2011). *Field Notes on Science & Nature*. p. 162. ↩

Contributors and Attributions

CC licensed content, Original

- Revision, Adaptation, and Original Content. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY-SA: Attribution-ShareAlike*

CC licensed content, Shared previously

- Geological compass. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Geological_compass. **License**: *CC BY-SA: Attribution-ShareAlike*
- Geologist's hammer. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Geologist%27s_hammer. **License**: *CC BY-SA: Attribution-ShareAlike*
- Loupe. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Loupe. **License**: *CC BY-SA: Attribution-ShareAlike*
- Fieldnotes. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Fieldnotes. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.1.7: Geologic Tools is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.8: Maps

Maps are essential tools in geology. Maps are as important in geology as written texts are in the study of literature. By studying maps, a geologist can see the shape and geology of the earth's surface and deduce the geological structures that lie hidden beneath the surface. Geologists are trained in map reading and map making. Many geologists have experience mapping some part of the earth's surface.

It takes some training to read maps skillfully. You are not expected to become a geological expert in reading maps. However, you will be expected to develop your map reading skills as you use maps to help you learn geology.

Topographic Maps

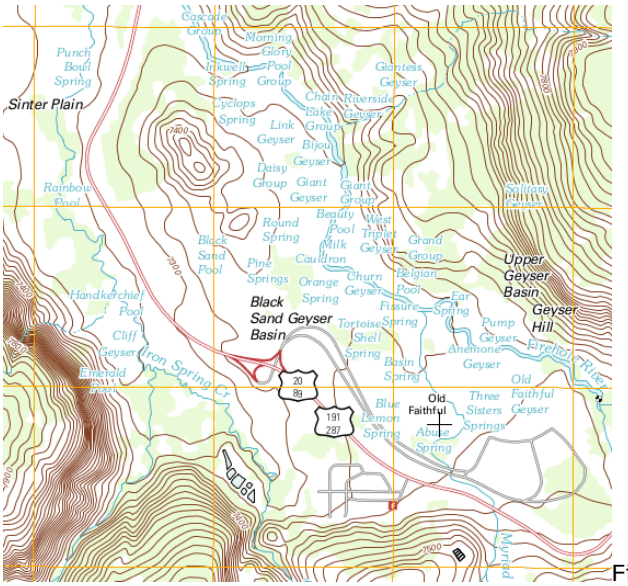


Figure 1. Map of Yellowstone.

A **topographic map** (like the one in figure 1) is one type of map used by geologists. Topographic maps show the three-dimensional shape of the land and features on the surface of the earth. Topographic maps are also used by hikers, planners who make decisions on zoning and construction permits, government agencies involved in land use planning and hazard assessments, and civil engineers. The topographic maps drawn and published by the U. S. Geological Survey portray the grids that are used on deeds to identify the location of real estate, so homeowners and property owners sometimes find it useful to refer to topographic maps of their area.

Most topographic maps make use of contour lines to depict elevations above sea level. The contour lines reveal the shape of the land in the vertical direction, allowing the 3-dimensional shape of the land to be portrayed on a 2-dimensional sheet of paper or computer screen. When you know how to read contour lines, you can look at them on a topographic map and visualize the mountains, plains, ridges, or valleys that they portrays.

Topographic maps are important in geology because they portray the surface of the earth in detail. This view of the surface shows patterns that provide information about the geology beneath the surface.

The landforms of the earth result from surface processes such as erosion or sedimentation combined with internal geological processes such as magma rising to create a volcano or a ridge of bedrock pushed up along a fault. By studying the shape of the earth's surface through topographic maps, geologists can understand the nature of surface processes in a given area, including zones subjected to landsliding, places undergoing erosion and places where sediment is accumulating. They can also find clues to the underlying geologic structure and geologic history of the area.

In addition to a topographic map, a complete understanding of the underlying geologic structure and history of an area requires completion of a geologic map and cross-sections. A topographic map provides the frame of reference upon which most geologic maps are constructed.

Reading a Topographic Map

Reading a topographic map requires familiarity with how it portrays the three-dimensional shape of the land, so that in looking at a topographic map you can visualize the shape of the land. To read a topographic map, you need to understand the rules of contour lines.

Rules for Contour Lines

- *A contour line connects all the points on a map area that are at a specific elevation. For example, every point on a 600-foot contour line represents a point on earth that is 600 feet above sea level. You can visualize a contour line as the shoreline that would exist if the ocean were to cover the land to that elevation.*
- *The contour interval is the vertical distance, also known as the elevation difference, between adjacent contour lines. On a map with a 40-foot contour interval, the **vertical distance** between two contour lines that are next to each other, is 40 feet, regardless of the horizontal distance between the two lines on the map.*
- Contour lines do not intersect each other, because a point on the surface of the earth cannot be at two different elevations. *(However, in the rare case of a vertical cliff showing up on a topographic map, contour lines along the cliff may appear to join together into a single line.)*
- Circles that are closed contours generally signify hills.
- Depressions that have no outlet are signified by closed contours with short lines that stick out of them and point toward the center. (The short lines sticking out of the contour lines are called hachures, hatch marks, or tick marks.)
- Contour lines on standard US Geological Survey topographic maps are brown — *except on the surfaces of glaciers, where the contour lines are blue.*
- The elevation of a point on the map that is not on a contour line must be estimated as greather than the elevation of the nearest contour line below it, and less than elevation of the nearest contour line above it. *For example, a point lying midway between the contours 5440 ft and 5480 ft would be at approximately 5460 ft elevation.*
- Contour lines curve upstream when they cross a valley. *This leads to the **"Rule of Vs"**: Where they cross streams, contour lines make Vs that point upstream.*
- Where contours are close together, the topography is steep; where contour lines are far apart, the slope is gentle or flat.
- The relief on a landscape is the elevation difference between two given points. *The maximum relief on a topographic map is the elevation difference between the highest and lowest points on the map.*

Map Quadrangles, Latitudes, and Longitudes

Standard United States Geological Survey topographic maps cover a quadrangle. A map quadrangle spans a fraction of a degree of longitude east-to-west and the same fraction of a degree of latitude north-to-south. Because lines of longitude degrees (also called meridians) in the Northern Hemisphere come closer and closer together the nearer they get to the North Pole, whereas lines of latitude degrees remain the same distance apart as they circle the earth, quadrangle maps span less distance east-to-west than they do north-to-south.

Latitude is how far north or south of the equator a point is on earth, measured in degrees, from 0° at the equator to 90° at the poles. When specifying a latitude, always state whether it is in the Northern Hemisphere (N) or Southern Hemisphere (S).

Longitude is how far east or west, up to a maximum of 180°, a point on earth is from the Prime Meridian. The Prime Meridian, 0° longitude, is a north-south line that runs through Greenwhich, England. When specifying a longitude, state whether it is in the Western Hemisphere (W) or Eastern Hemisphere (E).

Meridians, lines of longitude, run from the South Pole to the North Pole, converging (coming together) at the poles. Because the meridians converge at the poles, a degree of longitude gets smaller and smaller near each pole. In contrast, a degree of latitude remains approximately 69 miles across, no matter how near or far from the poles or equator it is.

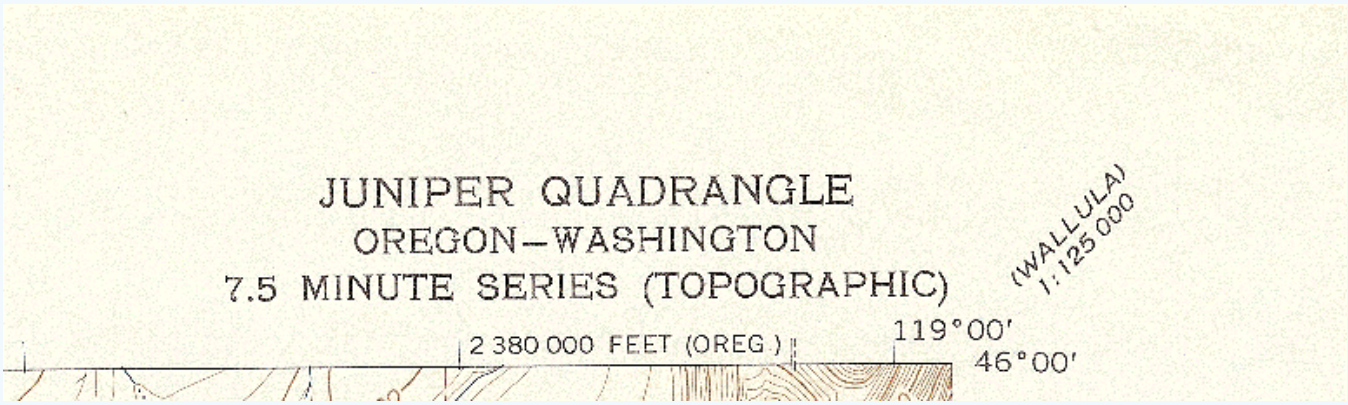
Degrees of latitude and longitude are divided into arc minutes and arc seconds. In this context, they are usually just called minutes and seconds, but it must be kept in mind that these minutes and seconds are units of angles, not units of time. These units, which divide angles into smaller parts, work as follows:

1. There are 60 arc minutes in 1 degree.
2. The symbol for minutes is a single apostrophe: ‘.
3. In symbols, 60′ = 1° means there are 60 minutes in 1 degree.

4. There are 60 arc seconds in 1 arc minute.
5. To convert arc minutes into a decimal fraction of a degree, multiply the number of arc minutes by 1°/60′. For example, to convert 15′ into a decimal fraction of a degree, 15′ x 1°/60′ = 0.25°. In simpler terms, just divide the number of arc minutes by 60 to convert to decimal degrees.
6. The symbol for arc seconds is a double apostrophe or quotation mark: ".
7. In symbols, 60″ = 1′ means there are 60 seconds in 1 minute.

Two common quadrangle sizes are 7.5 minutes (1/8 of a degree), and 15 minutes (1/4 of a degree).

Name, Size, and Latitude-Longitude of a Topographic Map Quadrangle



The image above shows the northeastern corner portion of the topographic map of the Juniper quadrangle, which spans the border of the states of Oregon and Washington. The name of the quadrangle comes from the name of a place on the map. Find the following information using this corner of the map:

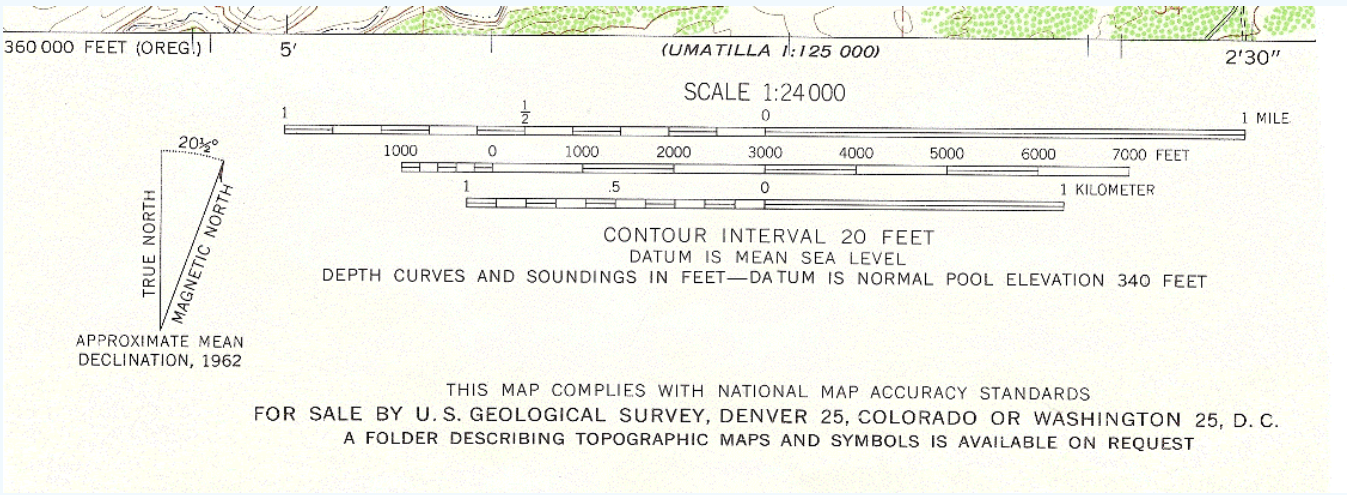
1. The name of the quadrangle
2. The state(s) in which the quadrangle is located
3. The size of the quadrangle
4. The name and fractional scale of the quadrangle map that is located adjacent to the northeast of the corner corner
5. The longitude of the eastern boundary of the map
6. The latitude of the northern boundary of the map

[reveal-answer q="588761"]Show Answer[/reveal-answer]
[hidden-answer a="588761"]

1. Juniper
2. Oregon and Washington
3. 7.5 minutes
4. Wallula, 1:125,000
5. 119° 00′
6. 46° 00′

[/hidden-answer]

Map Scale, Contour Interval, and Magnetic Declination



Important information is shown at the bottom of a USGS quadrangle map, including the map scale, the contour interval, and the magnetic declination. The image above is from the bottom of the Juniper 7.5 minute quadrangle. It tells you, among other things:

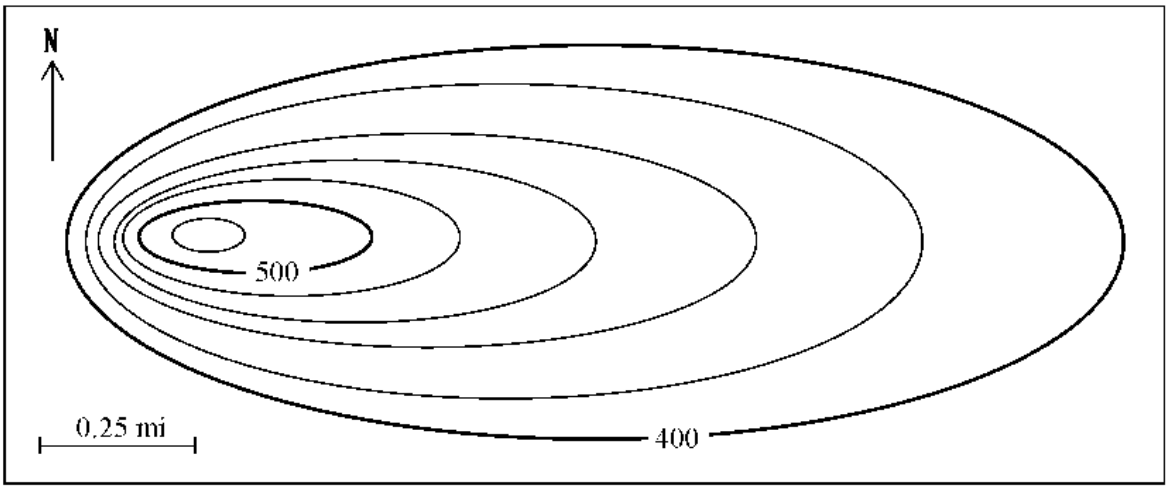
1. The map scale. The map scale is listed in terms of the fractional scale as 1:24,000. This means that 1 inch on the map corresponds to 24,000 inches in the real world represented by the map, or 1 cm equals 24,000 cm; in other words, distances on the map have been shrunk by a factor of 24,000 from their real-world size. Beneath the fractional scale, the map scale is also depicted a different way, in bar scales using three different units. One of the bar scales is in miles, one is in units of thousands of feet, and one is in kilometers.
2. The contour interval, the difference in elevation between adjacent contour lines on the map, is listed below the map scale as 20 feet.
3. There is also a reminder that elevations shown on the map are elevations above mean (average) sea level on earth.
4. (You might note that this map does something unusual for a topographic map. It shows depths in the Columbia River in feet below the surface of the river when the river is backed up in its reservoir behind a dam to a normal pool surface elevation of 340 feet above sea level.)
5. To the left of the bar scales, the magnetic declination is shown as an arrow diverging from a line oriented toward true north. True north is the direction toward the geographic North Pole. The geographic North Pole is where the northern end of earth's axis of rotation is located. The magnetic North Pole is in northeastern Canada. In 1962, the magnetic North Pole, as measured from the Juniper quadrangle, was located 20.5° east of true north. If you took a magnetic compass to the Juniper quadrangle in 1962, its north arrow would point 20.5° east of true north, so you would have had to set your magnetic compass to compensate for the declination. The magnetic North Pole wanders a few miles every year and the magnetic declination of 20.5° east of true north was determined in the year 1962; it may be slightly different now.

Constructing a Topographic Profile

One of the important tools you can use to extract the vertical information from a topographic map, and see more clearly the shape of the earth's surface that it represents is a topographic profile.

Construction of a topographic profile allows you to visualize the vertical component of a landscape. A topographic profile is similar to the view you have of a landscape while standing on earth, looking at hills and valleys from the side rather than from above.

Given a topographic map such as the one below, here's how to construct a topographic profile.

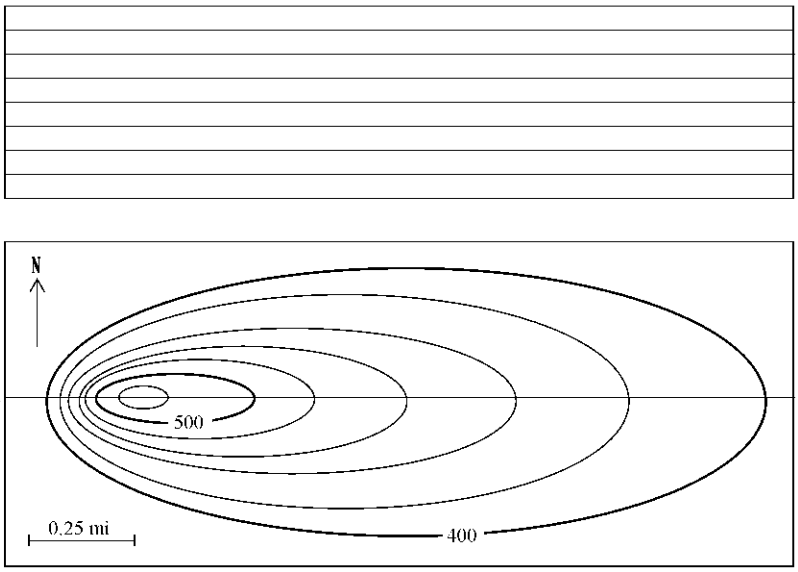


Step 1

Determine the line of profile, the line across that part of the map that you want to see in profile or cross-section view. Depending on which part of the map you want to see in profile, you can draw your line of profile in any direction you choose, across any part of the map you choose. For the map used in this example, we choose to draw the profile from A to A' as shown in the diagram below, to see the entire length of the hill in profile.

Step 2

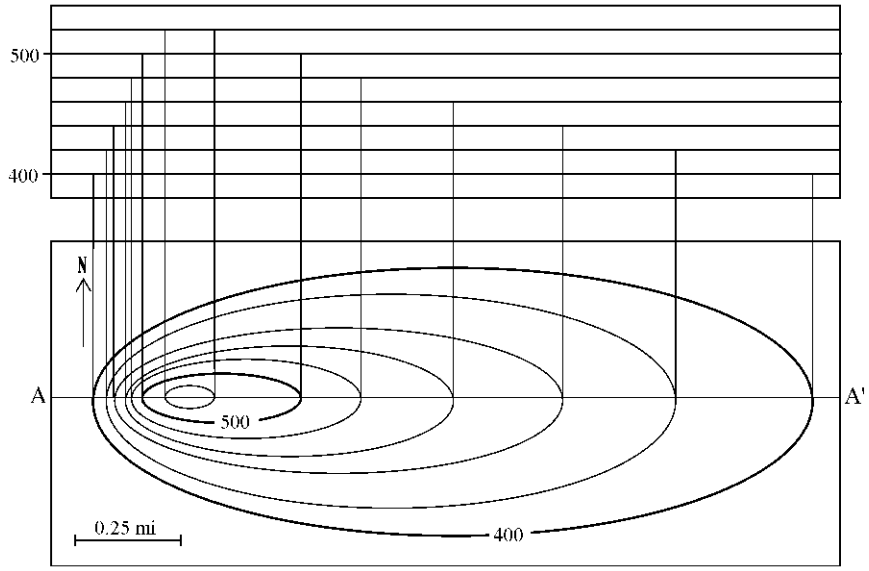
Draw a grid that will contain the profile. The width of the grid should be the same as the length of the line of profile. To draw the profile, the grid must be crossed by evenly-spaced horizontal lines that represent the contour elevations. The grid must extend high enough to span the elevation range of the contour lines spanned by the line of profile. You can see that the grid, shown below, includes the range of elevations that the line of profile crosses on the map. In addition, the grid must have an extra horizontal line at the bottom and top to accommodate the parts of the profile that go above the highest contour elevation and below the lowest contour elevation. That is why the grid in the example below goes below 400 feet and above 500 feet in elevation.



Step 3

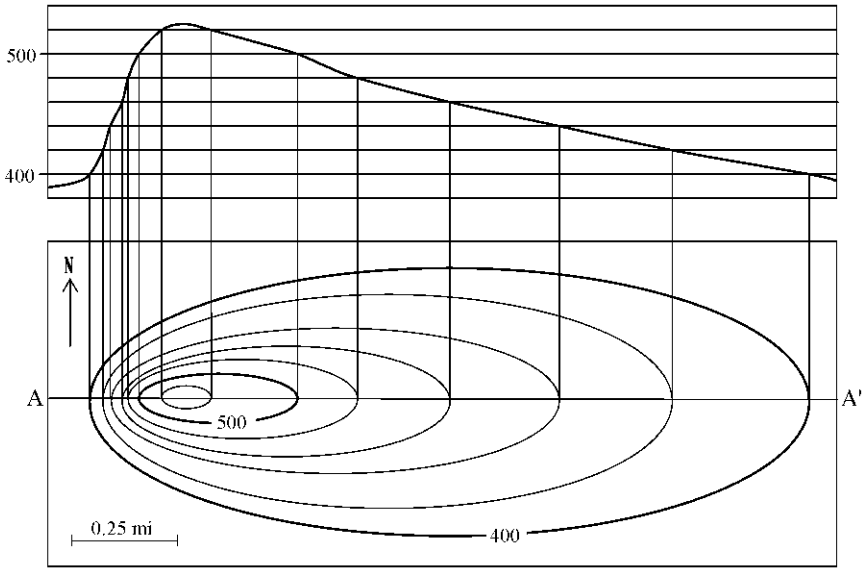
Transfer the contour elevations from the topographic map to the profile grid. The point where each contour line crosses the line of

profile on the topographic map determines the horizontal coordinate of each corresponding point on the grid of the topographic profile. The elevation of each contour line corresponds to the vertical coordinate of each corresponding point on the profile grid, as shown on the diagram below.



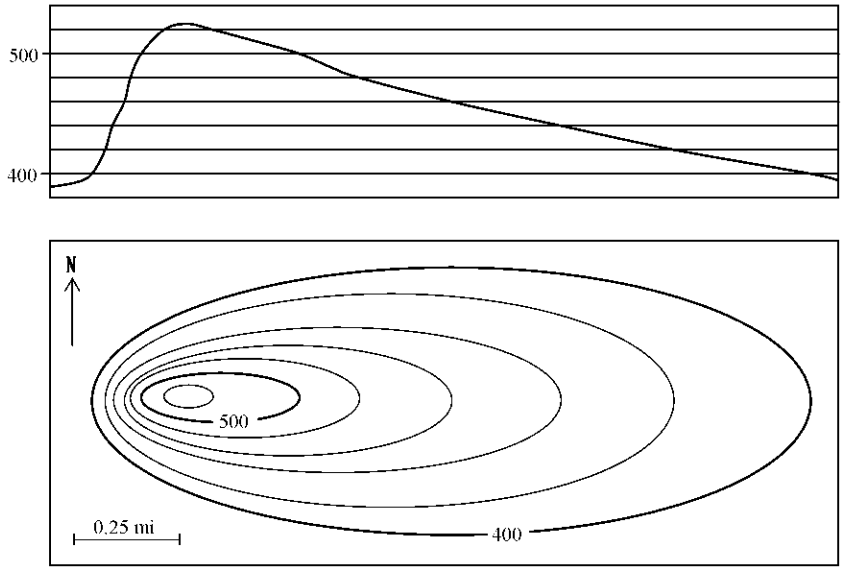
Step 4

Now that you have marked the elevation points on the profile grid, draw a smooth line connecting the data points as shown below. Note that the ends of this profile go below the 400 foot contour elevation but they do not extend to the 380 foot elevation because on the map the line of profile did not reach the 380 foot contour line. Also note that the top of the profile reaches a peak above 520 feet but less than 540 feet because the line of profile does not cross the 540 foot contour line.



Step 5

The completed topographic profile and the map it was drawn from are shown below. Topographic profiles are usually constructed without drawing any lines on the map. Instead, the edge of a piece of paper is laid along the line of profile and the contour line data is transferred to the edge of the piece of paper. From the edge of the piece of paper, the data are transfered to the profile grid, which is on a separate piece of paper.



Notice on the topographic profile constructed above that the peak of the hill is above 520 ft, but below 540 ft. Similarly, the ends of the profile are below 400 ft but above 380. This is consistent with the elevations of those parts of the line of profile on the map.

Note that the vertical scale on the profile is very different from the horizontal scale on the map. In this example, the map covers 0.25 miles horizontally in less distance than the profile covers 100 feet vertically. As a result, the topographic profile is greatly exaggerated vertically. In an actual view of the hill, looking at it from the side, it would not look nearly as steep as it does in the topographic profile that we have constructed.

If the vertical scale on a topographic profile is different from the map scale, as it is in this case, then the profile will exhibit a *vertical exaggeration*. The vertical exaggeration of a topographic profile can be calculated. It is the fractional scale of the topographic profile's vertical axis, divided by the fractional scale of the map. For example, if the vertical scale on the profile is 1:200 and the map scale is 1:24,000, the vertical exaggeration is \displaystyle\frac{\left(\frac{1}{200}\right)}{\left(\frac{1}{24,000}\right)}. To divide by a fraction, you can invert and multiply, so this becomes \displaystyle\left(\frac{1}{200}\right)\times\left(\frac{24,000}{1}\right)=\frac{24,000}{200}=120. A topographic profile with a VE of 120 would be a very exaggerated topographic profile. It would be as if a rubber model of the landscape has been pulled in the vertical direction, until it is 120 times taller than it really is.

If the vertical scale of a topographic profile is different from the map scale, the vertical exaggeration should be listed next to the profile, such as VE = 10 or VE 10x if the vertical exaggeration is 10.

Compare the profile to the topographic map. You will see that the hill is steeper on the west (left) side than on the east (right) side. This is consistent with the contour lines being spaced more closely on the west side of the hill and farther apart on the east side of the hill. This accords with the rules of contour lines, which state that slopes are steeper where contour lines are more closely spaced, and slopes are less steep where contour lines are more widely spaced.

If you drew a profile from north to south across the peak of the hill, do you think the profile would be symmetric or asymmetric?

Checklist for a Complete Topographic Profile

A properly drawn topographic profile will have the following attributes:

- The topographic profile is drawn on a rectilinear graph with evenly spaced grid lines. (Vertical grid lines are not required.)
- Elevation lines are labeled along the left-hand vertical axis.
- The profile is a smooth curve where its gradient changes, rather than straight-line segments connecting the dots and only bending at the dots.
- If the vertical scale on the profile is different from the map scale, the resulting amount of vertical exaggeration is listed.
- The ends and any high points or low points of the topographic profile should be above or below elevation lines, not on them,

except in cases where an end, high point, or low point of a line of profile happens to fall right on a contour line.

Explore More: Topographic Maps

Use this resource to answer the questions that follow. You may stop watching at the 4:04 mark.

A link to an interactive elements can be found at the bottom of this page.

1. What do topographic maps do and how do they do it?
2. What are the meanings of the terms topographic map, contour line, contour interval and index contour?
3. If you were to walk along a contour line, what would happen to your elevation?
4. If you walk perpendicular to contour lines what are you doing?
5. What do close contour lines indicate?

Bathymetric Maps

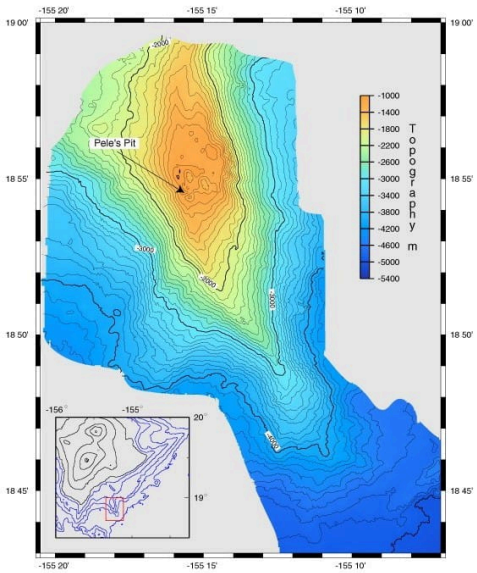


Figure 3. Loihi volcano growing on the flank of Kilauea volcano in Hawaii. Black lines in the inset show the land surface above sea level and blue lines show the topography below sea level. Click on the image to view a larger version.

A **bathymetric map** is like a topographic map with the contour lines representing depth below sea level, rather than height above. Numbers are low near sea level and become higher with depth.

Kilauea is the youngest volcano found above sea level in Hawaii. On the flank of Kilauea is an even younger volcano called Loihi. The bathymetric map pictured in figure 3 shows the form of Loihi.

Geologic Maps

A **geologic map** shows the geological features of a region (see figure 4 for an example). Rock units are color-coded and identified in a key. Faults and folds are also shown on geologic maps. The geology is superimposed on a topographic map to give a more complete view of the geology of the region.

A geologic map shows mappable rock units, mappable sediment units that cover up the rocks, and geologic structures such as faults and folds. A mappable unit of rock or sediment is one that a geologist can consistently recognize, trace across a landscape, and describe so that other people are able to recognize it and verify its presence and identity. Mappable units are shown as different colors or patterns on a base map of the geographic area.

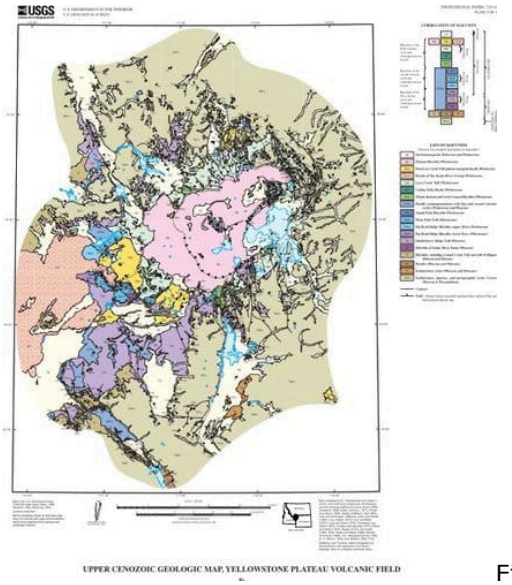


Figure 4. A geologic map of the region around Old Faithful, Yellowstone National Park.

Geologic maps are important for two reasons. First, as geologists make geologic maps and related explanations and cross-sections, they develop a theoretical understanding of the geology and geologic history of a given area.

Second, geologic maps are essential tools for practical applications such as zoning, civil engineering, and hazard assessment. Geologic maps are also vital in finding and developing geological resources, such as gravel to help build the road you drive on, oil to power the car you travel in, or aluminum to build the more fuel-efficient engine in your next vehicle. Another resource that is developed on the basis of geologic maps is groundwater, which many cities, farms, and factories rely on for the water they use.

Essential Components of Geologic Maps

A complete geologic map has at least two features:

- the map itself
- the map legend or key that explains all the symbols on the map.

Professional geologic maps usually have two other components as well:

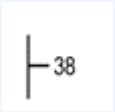
- an accompanying explanation of the rock or sediment units
- geologic cross-sections of the map area.

The legend or key to a geologic map is usually printed on the same page as the map and follows a customary format. The symbol for each rock or sediment unit is shown in a box next to its name and brief description. These symbols are stacked in age sequence from oldest at the bottom to youngest at the top. The geologic era, or period, or epoch–the geologic age–is listed for each rock unit in the key. By stacking the units in age sequence from youngest at the top to oldest at the bottom, and identifying which interval of geologic time each unit belongs to, the map reader can quickly see the age of each rock or sediment unit. The map key also contains a listing and explanation of the symbols shown on the map, such as the symbols for different types of faults and folds. See the Table of Geologic Map Symbols for pictures and an overview of the map symbols, including strikes and dips, faults, folds, and an overview.

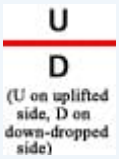
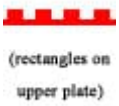
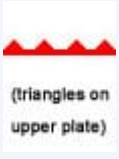
Table of Geologic Map Symbols

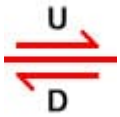
Strike and Dip Symbols

Strike and dip are a way of representing the three-dimensional orientation of a planar surface on a two-dimensional map. The strike is the compass direction of a horizontal line on the plane. All the horizontal lines on a plane are parallel, so they all have the same characteristic compass direction. The dip is the angle at which the plane slopes downhill from the horizontal, at its maximum slope, which is at right angles (90°) from strike.

Map Symbol	Definition	Explanation of symbol
	strike and dip of beds other than horizontal or vertical	• strike (longer line) is horizontal line on bedding plane • strike parallels nearby contacts between stratified rocks • dip shows which way beds run downhill • dip angle, number at end of dip symbol, is how much beds tilt down from horizontal
	horizontal beds	• because the bed is horizontal it strikes in all directions • because the bed is horizontal, the dip is 0%
	strike and dip of vertical beds	• strike (longer line) is horizontal line on bedding plane • because the bed dips vertically (has a 90% dip), it dips equally in either direction at right angles to strike, so the dip line is shown extending in both directions

Geologic Fault Symbols

Type of Fault	Map Symbol	Definition	Type of Regional Stress	Geologic Associations
normal		hanging wall down, footwall up	tension	• zones of crustal extension • divergent plate boundaries • edges of horsts and grabens • Basin and Range region
detachment		low-angle normal fault, footwall—gneiss, hanging wall—shallow-crust rocks	tension	• boundaries of metamorphic core complexes
thrust		hanging wall up, footwall down	compression	• zones of crustal compression • convergent plate boundaries
reverse		high-angle (45° or more dip) thrust fault	compression	• zones of crustal compression • convergent plate boundaries

Geologic Fault Symbols				
strike-slip		rocks on either side move horizontally in opposite directions	shear	• continental margins undergoing oblique (not straight on) plate convergence • transform plate boundaries
oblique-slip		combines horizontal and vertical motion	combination	• orogenic mountain belts • continental margins undergoing oblique (not straight on) plate convergence

Geologic Fold Symbols			
Type of Fold	**Map Symbol**	**Definition**	**Appearance of Beds in Map View**
anticline		up fold	• roughly parallel stripes • dip away from center (away from axis) • oldest at center (along axis) • youngest farthest from center
plunging anticline		up fold with tilted axis	• roughly a U-shaped pattern • plunges in direction U points • oldest at center (along axis) • youngest farthest from center
syncline		down fold	• roughly parallel stripes • dip toward center (toward axis) • oldest farthest from center • youngest at center (along axis)
plunging syncline		down fold with tilted axis	• roughly a U-shaped pattern • plunges in direction U opens • oldest farthest from center • youngest at center (along axis)
monocline		strata tilted in one direction	• all dip in same direction
structural dome		upward bulge in layered rocks	• roughly a bull's eye pattern • dip away from center • oldest in center • youngest farthest from center

Geologic Fold Symbols			
structural basin		downward bulge in layered rocks	• roughly a bull's eye pattern • dip toward center • youngest in center • oldest farthest from center

The explanations of rock units are often given in a separate pamphlet that accompanies the map. The explanations include descriptions with enough detail for any geologist to be able to recognize the units and learn how their ages were determined.

If included, cross-sections are usually printed on the same page as the geologic map. They are important accompaniments to geologic maps, especially if the map focuses on the geology of the bedrock underneath the soil and loose sediments.

Geologic Cross-Sections

A geologic cross-section is a sideways view of a slice of the earth. It shows how the different types of rock are layered or otherwise configured, and it portrays geologic structures beneath the earth's surface, such as faults and folds. Geologic cross-sections are constructed on the basis of the geology mapped at the surface combined with an understanding of rocks in terms of physical behavior and three-dimensional structures.

Summary

- Earth scientists regularly use topographic, bathymetric, and geologic maps.
- Topographic maps reveal the shape of a landscape. Elevations indicate height above sea level.
- Bathymetric maps are like topographic maps of features found below the water. Elevations indicate depth below sea level.
- Geologic maps show rock units and geologic features like faults and folds.

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- 2.5: Maps. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-Concepts-For-High-School/section/2.5/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Basics -- Topographic & Geologic Maps. **Authored by**: Ralph L. Dawes, Ph.D. and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: https://commons.wvc.edu/rdawes/G101OCL/Basics/mapkey.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

Public domain content

- Loihi Bathemetric. **Authored by**: NOAA. **Located at**: https://commons.wikimedia.org/wiki/File:LoihiBathemetric.jpg. **License**: *Public Domain: No Known Copyright*

This page titled 4.1.8: Maps is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.9: Location and Direction

If you found this feature while out in the field, could you find it again?


Figure 1. Old Faithful

If you're going to make observations of geological features, you're going to need to know the location where you are so you can mark it on a map. If you find a rock formation filled with gold, you'll want to be able to find the location again!

You may need to tell someone when your truck gets stuck when you're in the field so you'll need a direction to give them.

The photo in figure 1 is of Old Faithful Geyser in Yellowstone National Park. Let's explore just a few of the ways we can pinpoint the location of this famous geological icon.

Location

How would you find Old Faithful? One way is by using latitude and longitude. Any **location** on Earth's surface — or on a map — can be described using these coordinates. Latitude and longitude are expressed as degrees that are divided into 60 minutes. Each minute is divided into 60 seconds.

Latitude

A look on a reliable website shows us that Old Faithful Geyser is located at N44^{o}27' 43''. What does this mean?

Latitude tells the distance north or south of the Equator. Latitude lines start at the Equator and circle around the planet. The North Pole is 90°N, with 90 degree lines in the Northern Hemisphere. Old Faithful is at 44 degrees, 27 minutes and 43 seconds north of the Equator. That's just about exactly half way between the Equator and the North Pole!

Longitude

The latitude mentioned above does not locate Old Faithful exactly, since a circle could be drawn that latitude north of the Equator. To locate Old Faithful we need another point – longitude. At Old Faithful the longitude is W110^{o}49'57''.

Longitude lines are circles that go around the Earth from north to south, like the sections of an orange. Longitude is measured perpendicular to the Equator. The Prime Meridian is 0° longitude and passes through Greenwich, England. The International Date Line is the 180° meridian. Old Faithful is in the Western Hemisphere, between the Prime Meridian in the east and the International Date Line in the west.

Elevation

An accurate location must take into account the third dimension. **Elevation** is the height above or below sea level. **Sea level** is the average height of the ocean's surface or the midpoint between high and low tide. Sea level is the same all around Earth.

Old Faithful is higher above sea level than most locations at 7,349 ft (2240 m). Of course, the highest point on Earth, Mount Everest, is much higher at 29,029 ft (8848 m).

Global Positioning System

Satellites continually orbit Earth and can be used to indicate location. A **global positioning system** receiver detects radio signals from at least four nearby GPS satellites. The receiver measures the time it takes for radio signals to travel from a satellite and then calculates its distance from the satellite using the speed of radio signals. By calculating distances from each of the four satellites the receiver can triangulate to determine its location. You can use a GPS meter to tell you how to get to Old Faithful.

Direction

Direction is important if you want to go between two places. **Directions** are expressed as north (N), east (E), south (S), and west (W), with gradations in between. The most common way to describe direction in relation to the Earth's surface is with a **compass**, a device with a floating needle that is actually a small magnet. The compass needle aligns itself with the Earth's magnetic north pole. Since the magnetic north pole is 11.5 degrees offset from its geographic north pole on the axis of rotation, you must correct for this discrepancy.

Without using a compass, we can say that to get to Old Faithful, you enter Yellowstone National Park at the South Entrance, drive north-northeast to West Thumb, and then drive west-northwest to Old Faithful.

Summary

- Latitude is the distance north or south of the Equator and is expressed as a number between 0 and 90 degrees north or south.
- Longitude is the distance east or west of the Prime Meridian and is expressed as a number between 0 and 180 degrees east or west.
- Elevation is the height above sea level.
- Direction is expressed as north, south, east, or west, or some gradation between them.

Explore More

Use this resource to answer the questions that follow.

A link to an interactive elements can be found at the bottom of this page.

1. What are lines of latitude?
2. How far apart are the lines of latitude, in degrees, in miles?
3. What are the latitudes of the Equator, the Tropic of Cancer, and the Tropic of Capricorn? What are the characteristics of the regions found between the Tropic of Cancer and Tropic of Capricorn?
4. Where are the Arctic and Antarctic circle? What are the characteristics of the regions that are found poleward of these circles?
5. What are lines of longitude?
6. Where do the meridians meet?
7. What is the Prime Meridian? Where is it located?
8. How are longitude and latitude measured?

Contributors and Attributions

CC licensed content, Original

- Revision and adaptation. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- 2.4: Location and Direction. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-Concepts-For-High-School/section/2.4/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Old Faithful Rainbow. **Authored by**: Flicka. **Located at**: https://commons.wikimedia.org/wiki/File:Old_Faithful_Rainbow.jpg. **License**: *CC BY-SA: Attribution-ShareAlike*

All rights reserved content

- Latitude and Longitude. **Authored by**: Andy Jensen. **Located at**: https://youtu.be/swKBi6hHHMA. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

This page titled 4.1.9: Location and Direction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.10: Scientific Models

Scientists use models to help them understand and explain ideas. Models explain objects or systems in a more simple way. Models often only show only a part of a system. The real situation is more complicated. Models help scientists to make predictions about complex systems. Some models are something that you can see or touch. Other types of models use an idea or numbers. Each type is useful in certain ways.

Scientists create models with computers. Computers can handle enormous amounts of data. This can more accurately represent the real situation. For example, Earth's climate depends on an enormous number of factors. Climate models can predict how climate will change as certain gases are added to the atmosphere. To test how good a model is, scientists might start a test run at a time in the past. If the model can predict the present it is probably a good model. It is more likely to be accurate when predicting the future.

Physical Models

A **physical model** is a representation of something using objects. It can be three-dimensional, like a globe. It can also be a two-dimensional drawing or diagram. Models are usually smaller and simpler than the real object. They most likely leave out some parts, but contain the important parts. In a good model the parts are made or drawn to scale. Physical models allow us to see, feel and move their parts. This allows us to better understand the real system.

An example of a physical model is a drawing of the layers of Earth (figure 1). A drawing helps us to understand the structure of the planet. Yet there are many differences between a drawing and the real thing. The size of a model is much smaller, for example. A drawing also doesn't give good idea of how substances move. Arrows showing the direction the material moves can help. A physical model is very useful but it can't explain the real Earth perfectly.

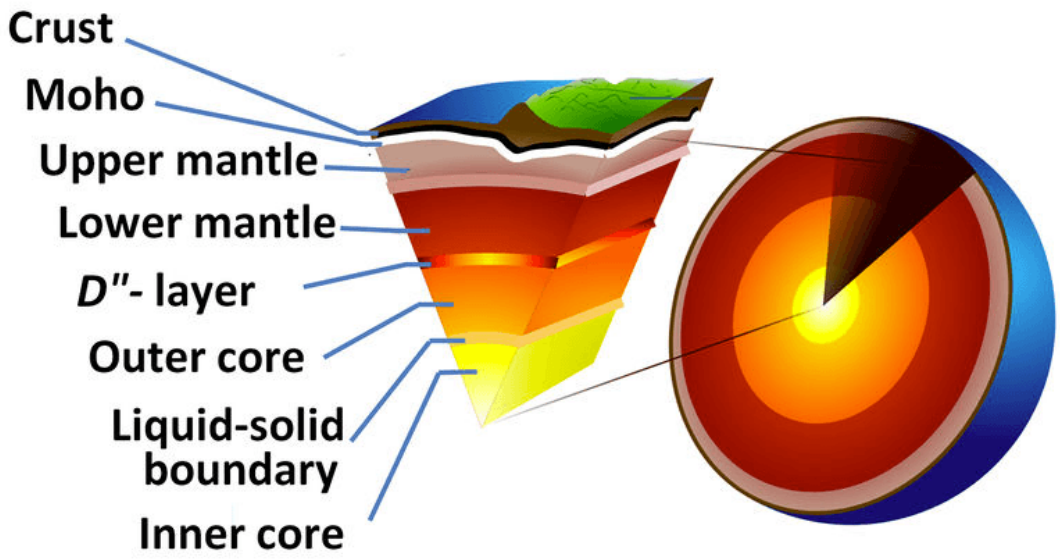


Figure 1. Earth's Center.

Ideas as Models


Figure 2. A collision showing a meteor striking Earth.

Some models are based on an idea that helps scientists explain something. A good idea explains all the known facts. An example is how Earth got its Moon. A Mars-sized planet hit Earth and rocky material broke off of both bodies (figure 2). This material orbited Earth and then came together to form the Moon. This is a model of something that happened billions of years ago. It brings together many facts known from our studies of the Moon's surface. It accounts for the chemical makeup of rocks from the Moon, Earth, and meteorites. The physical properties of Earth and Moon figure in as well. Not all known data fits this model, but much does. There is also more information that we simply don't yet know.

Models that Use Numbers

Models may use formulas or equations to describe something. Sometimes math may be the only way to describe it. For example, equations help scientists to explain what happened in the early days of the universe. The universe formed so long ago that math is the only way to describe it. A climate model includes lots of numbers, including temperature readings, ice density, snowfall levels, and humidity. These numbers are put into equations to make a model. The results are used to predict future climate. For example, if there are more clouds, does global temperature go up or down? Models are not perfect because they are simple versions of the real situation. Even so, these models are very useful to scientists. These days, models of complex things are made on computers.

Contributors and Attributions

CC licensed content, Shared previously

- 1.1: The Nature of Science. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-Middle-School/section/1.1/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.1.10: Scientific Models is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.1.11: Geologic Modelling

Figure 1. Geological mapping software displaying a screenshot of a structure map generated for an 8500ft deep gas & Oil reservoir in the Erath field, Vermilion Parish, Erath, Louisiana. The left-to-right gap, near the top of the contour map indicates a Fault line. This fault line is between the blue/green contour lines and the purple/red/yellow contour lines. The thin red circular contour line in the middle of the map indicates the top of the oil reservoir. Because gas floats above oil, the thin red contour line marks the gas/oil contact zone.

Geologic modelling, or **Geomodelling**, is the applied science of creating computerized representations of portions of the Earth's crust based on geophysical and geological observations made on and below the Earth surface. A Geomodel is the numerical equivalent of a three-dimensional geological map complemented by a description of physical quantities in the domain of interest. Geomodelling is related to the concept of Shared Earth Model; which is a multidisciplinary, interoperable and updatable knowledge base about the subsurface.

Geomodelling is commonly used for managing natural resources, identifying natural hazards, and quantifying geological processes, with main applications to oil and gas fields, groundwater aquifers and ore deposits. For example, in the oil and gas industry, realistic geologic models are required as input to reservoir simulator programs, which predict the behavior of the rocks under various hydrocarbon recovery scenarios. A reservoir can only be developed and produced once; therefore, making a mistake by selecting a site with poor conditions for development is tragic and wasteful. Using geological models and reservoir simulation allows reservoir engineers to identify which recovery options offer the safest and most economic, efficient, and effective development plan for a particular reservoir.

Geologic modelling is a relatively recent subdiscipline of geology which integrates structural geology, sedimentology, stratigraphy, paleoclimatology, and diagenesis;

In 2-dimensions (2D), a geologic formation or unit is represented by a polygon, which can be bounded by faults, unconformities or by its lateral extent, or crop. In geological models a geological unit is bounded by 3-dimensional (3D) triangulated or gridded surfaces. The equivalent to the mapped polygon is the fully enclosed geological unit, using a triangulated mesh. For the purpose of property or fluid modelling these volumes can be separated further into an array of cells, often referred to as voxels (volumetric elements). These 3D grids are the equivalent to 2D grids used to express properties of single surfaces.

Geomodelling generally involves the following steps:

1. Preliminary analysis of geological context of the domain of study.
2. Interpretation of available data and observations as point sets or polygonal lines (e.g. "fault sticks" corresponding to faults on a vertical seismic section).
3. Construction of a structural model describing the main rock boundaries (horizons, unconformities, intrusions, faults)
4. Definition of a three-dimensional mesh honoring the structural model to support volumetric representation of heterogeneity (see Geostatistics) and solving the Partial Differential Equations which govern physical processes in the subsurface (e.g. seismic wave propagation, fluid transport in porous media).

Geologic modelling components

Structural framework

Incorporating the spatial positions of the major formation boundaries, including the effects of faulting, folding, and erosion (unconformities). The major stratigraphic divisions are further subdivided into layers of cells with differing geometries with relation to the bounding surfaces (parallel to top, parallel to base, proportional). Maximum cell dimensions are dictated by the minimum sizes of the features to be resolved (everyday example: On a digital map of a city, the location of a city park might be adequately resolved by one big green pixel, but to define the locations of the basketball court, the baseball field, and the pool, much smaller pixels – higher resolution – need to be used).

Rock type

Each cell in the model is assigned a rock type. In a coastal clastic environment, these might be beach sand, high water energy

marine upper shoreface sand, intermediate water energy marine lower shoreface sand, and deeper low energy marine silt and shale. The distribution of these rock types within the model is controlled by several methods, including map boundary polygons, rock type probability maps, or statistically emplaced based on sufficiently closely spaced well data.

Reservoir quality

Reservoir quality parameters almost always include porosity and permeability, but may include measures of clay content, cementation factors, and other factors that affect the storage and deliverability of fluids contained in the pores of those rocks. Geostatistical techniques are most often used to populate the cells with porosity and permeability values that are appropriate for the rock type of each cell.

Fluid saturation

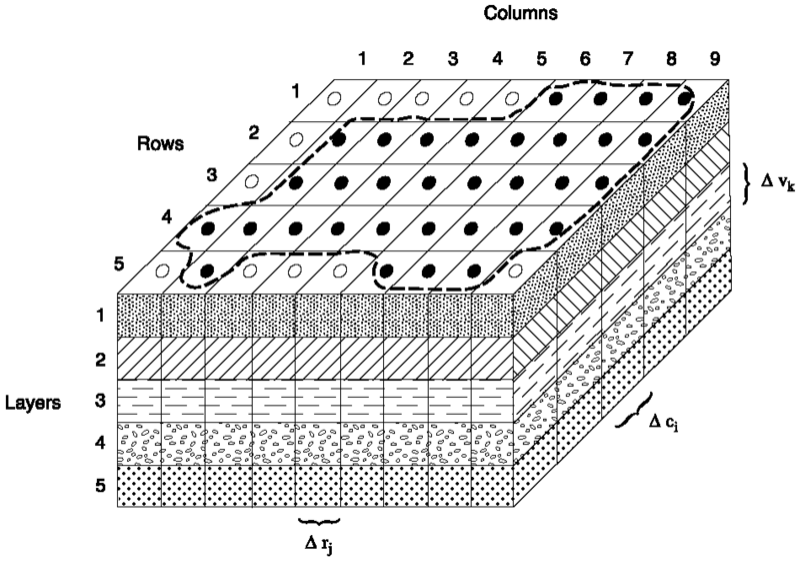


Figure 2. A 3D finite difference grid used in MODFLOW for simulating groundwater flow in an aquifer.

Most rock is completely saturated with groundwater. Sometimes, under the right conditions, some of the pore space in the rock is occupied by other liquids or gases. In the energy industry, oil and natural gas are the fluids most commonly being modelled. The preferred methods for calculating hydrocarbon saturations in a geologic model incorporate an estimate of pore throat size, the densities of the fluids, and the height of the cell above the water contact, since these factors exert the strongest influence on capillary action, which ultimately controls fluid saturations.

Geostatistics

An important part of geologic modelling is related to geostatistics. In order to represent the observed data, often not on regular grids, we have to use certain interpolation techniques. The most widely used technique is kriging which uses the spatial correlation among data and intends to construct the interpolation via semi-variograms. To reproduce more realistic spatial variability and help assess spatial uncertainty between data, geostatistical simulation based on variograms, training images, or parametric geological objects is often used.

Mineral Deposits

Geologists involved in mining and mineral exploration use geologic modelling to determine the geometry and placement of mineral deposits in the subsurface of the earth. Geologic models help define the volume and concentration of minerals, to which economic constraints are applied to determine the economic value of the mineralization. Mineral deposits that are deemed to be economic may be developed into a mine.

Technology

Geomodelling and CAD share a lot of common technologies. Software is usually implemented using object-oriented programming technologies in C++, Java or C# on one or multiple computer platforms. The graphical user interface generally consists of one or several 3D and 2D graphics windows to visualize spatial data, interpretations and modelling output. Such visualization is generally achieved by exploiting graphics hardware. User interaction is mostly performed through mouse and keyboard, although 3D pointing devices and immersive environments may be used in some specific cases. GIS (Geographic Information System) is also a widely used tool to manipulate geological data.

Geometric objects are represented with parametric curves and surfaces or discrete models such as polygonal meshes.

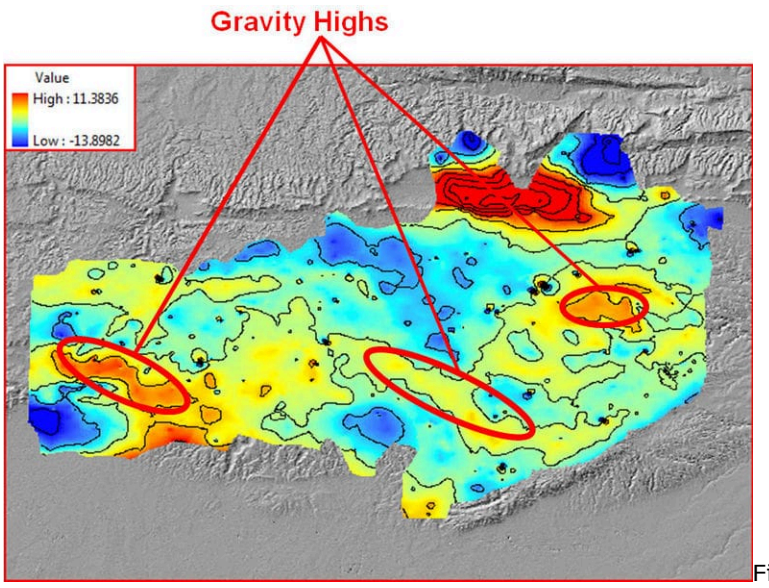


Figure 3. Gravity Highs

Research in Geomodelling

Problems pertainting to Geomodelling cover

- Defining an appropriate Ontology to describe geological objects at various scales of interest
- Integrating diverse types of observations into 3D geomodels: geological mapping data, borehole data and interpretations, seismic images and interpretations, potential field data, well test data, etc.
- Better accounting for geological processes during model building
- Characterizing uncertainty about the geomodels to help assess risk. Therefore, Geomodelling has a close connection to Geostatistics and Inverse problem theory
- Applying of the recent developed Multiple Point Geostatistical Simulations (MPS) for integrating different data sources
- Automated geometry optimization and topology conservation

History

In the 1970s, geomodelling mainly consisted of automatic 2D cartographic techniques such as contouring, implemented as FORTRAN routines communicating directly with plotting hardware. The advent of workstations with 3D graphics capabilities during the 1980s gave birth to a new generation of geomodelling software with graphical user interface which became mature during the 1990s.

Since its inception, geomodelling has been mainly motivated and supported by oil and gas industry.

Geologic modelling software

Software developers have built several packages for geologic modelling purposes. Such software can display, edit, digitise and

automatically calculate the parameters required by engineers, geologists and surveyors. Current software is mainly developed and commercialized by oil and gas or mining industry software vendors:

Geologic modelling and visualisation

- SGS Genesis
- IRAP RMS Suite
- Geomodeller3D
- Geosoft provides GM-SYS and VOXI 3D modelling software
- GSI3D
- Petrel
- Rockworks
- Move

Groundwater modelling

- FEFLOW
- FEHM
- MODFLOW
  - GMS
  - Visual MODFLOW
- ZOOMQ3D

Moreover, industry Consortia or companies are specifically working at improving standardization and interoperability of earth science databases and geomodelling software:

- Standardization: GeoSciML by the Commission for the Management and Application of Geoscience Information, of the International Union of Geological Sciences.
- Standardization: RESQML(tm) by Energistics
- Interoperability: OpenSpirit, by TIBCO(r)

Contributors and Attributions

CC licensed content, Original

- Revision and adaptation. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY-SA: Attribution-ShareAlike*

CC licensed content, Shared previously

- Geologic modelling. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Geologic_modelling. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.1.11: Geologic Modelling is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.2: ROCK FORMING MINERALS

Learning Objectives

Identify and compare common rock forming minerals, mineral structures and mineral properties.

- Describe the basic chemistry involved in mineral formation and structures.
- Identify and classify common rock forming minerals.
- Identify the most common elements in the Earth's crust and their order of abundance.

4.2.1: The Building Blocks of Matter

4.2.2: Atomic Bonds

4.2.3: Physical Characteristics of Minerals

4.2.4: Classifying Minerals

4.2.5: Abundance of Elements in Earth's Crust

4.2.6: Silicate Minerals and the Silicate Tetrahedron

This page titled Chapter 4.2: Rock Forming Minerals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.2.1: The Building Blocks of Matter

Why study the chemistry of minerals? Minerals are made of **atoms**, which have an impact on the behavior and characteristics of the mineral. Thus, to understand, explain, and predict the behavior of minerals, and rocks—which are made of minerals—we must understand some basic facts about atoms and how they behave. This requires a basic understanding of some chemistry. We will begin by constructing atoms in our thinking in terms of the three sub-atomic particles of which atoms are made.

Atoms

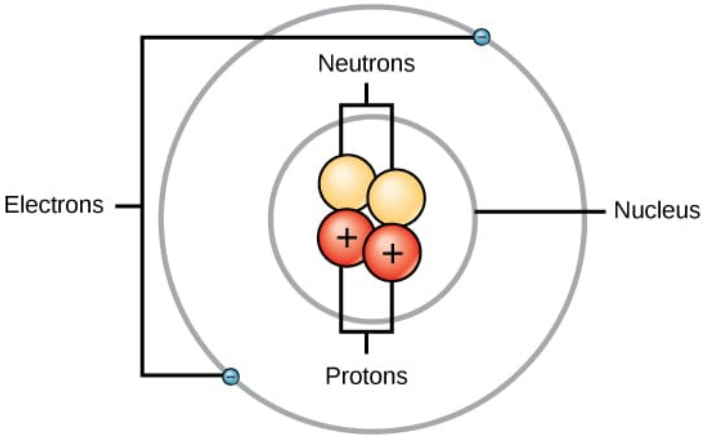


Figure 1. Elements, such as helium, depicted here, are made up of atoms. Atoms are made up of protons and neutrons located within the nucleus, with electrons in orbitals surrounding the nucleus.

Atoms consist of protons, neutrons, and electrons. **Protons** have a positive (+) electrical charge. **Electrons** have a negative (–) charge that is exactly equal and opposite to the electrical charge of a proton. **Neutrons** are electrically neutral.

Most of the mass of an atom is packed into its tiny nucleus. An atomic nucleus is made of protons and neutrons, which have approximately the same mass (about 1.67×10^{-24} grams). Electrons, on the other hand, are arranged in specific orbitals around the nucleus of an atom; they are also much smaller in mass than protons and neutrons, weighing only 9.11×10^{-28} grams, or about 1/1800 the weight of protons and neutrons.

Even though the mass of an electron is a tiny mass compared to the mass of a proton or a neutron, the electrons occupy most of the volume of an atom (see Figure 1).

Ions

A neutral atom has the same number of electrons as it does protons. An atom that has lost or gained any electrons is no longer an electrically neutral atom. That type of atom, which is not electrically neutral and has an electrical charge associated with it, is called an **ion**. Atoms that have gained electrons are negatively (–) charged ions, or **anions**. Atoms that have lost electrons are positively (+) charged ions, or **cations**.

It is also possible to have ions that are actually small groups of atoms bonded together. These are known as polyatomic ions. One example of a polyatomic ion is the carbonate ion, $(CO_3)^{2-}$, which has two extra electrons, giving it the net electrical charge of 2–.

Chemical Formulas

We've just seen that a carbonate ion can also be called $(CO_3)^{2-}$. But what exactly does this mean?

First let's look at the letters: CO. Atoms have chemical symbols; each element has been assigned one or two letters to represent it. Thus, C stands in for carbon, and O stands in for oxygen (all of these chemical symbols can be seen in the periodic table in figure 2 below).

As we read above, the $^{2-}$ means that the ion has two extra electrons. But what about the $_3$? This means there are three oxygen atoms in the ion. The number of atoms in a particular formula is always notated in subscript. Charge is always written in superscript at the end of the formula (a superscript at the beginning of the formula means something else—we'll get to this when we discuss isotopes below). The parentheses around CO_3 indicate that the charge belongs to the whole polyatomic unit rather than just the O_3.

Thus the carbonate ion is one carbon atom (C), three oxygen atoms (O_3), and two extra electrons ($^{2-}$), which charge the whole polyatomic ion.

The Periodic Table

Naturally occurring atoms found in the earth range from hydrogen, with just one proton in its nucleus, to uranium, with 92 protons in its nucleus. These are the naturally occurring chemical elements, which includes such commonly known elements as carbon, oxygen, iron, and so on. The periodic table lists all the chemical elements in a way that tells us how many protons each of them has, how its electrons are arranged, and what the general chemical behavior of each element is.

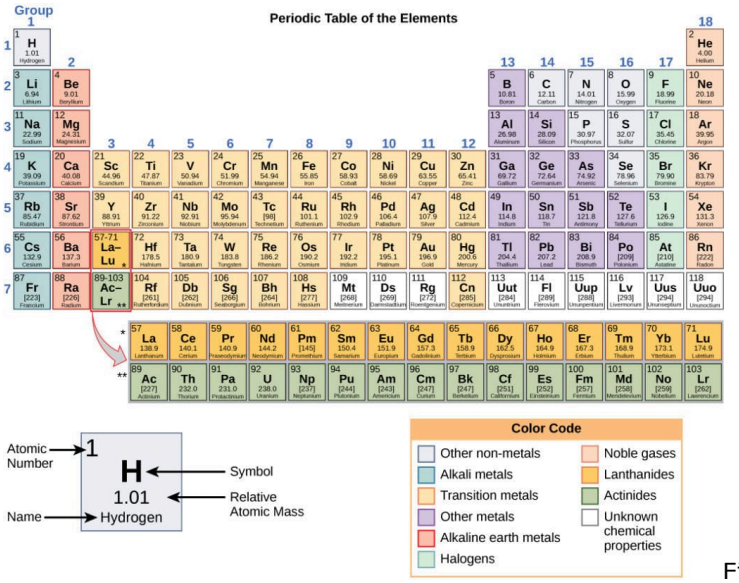


Figure 2. Follow this link to a large version of the Periodic Table. The link will open in a new window so you can easily refer to it as you read through this page.

As shown in Figure 2, the periodic table consists of eighteen groups and seven periods. Two additional rows of elements, known as the lanthanides and actinides, are placed beneath the main table. These elements are placed separately to make the table more compact. All the elements in a group have a similar chemical behavior. This is because all the elements in a group have a similar arrangement of electrons in their atoms, and it is the electron arrangement that determines the chemical behavior of an element.

For each element, the name, atomic symbol, atomic number, and atomic mass are provided. The atomic number is a whole number that represents the number of protons: each chemical element is distinguished by the number of protons in its nucleus. For example, every atom of the element oxygen has eight protons in its nucleus. That is why the atomic number of oxygen is 8. If an atom has greater or fewer than eight protons in its nucleus, it is not oxygen, it is some other chemical element. In the periodic table, the atomic number of each element is listed above the chemical symbol of the element.

The atomic mass, which is the average mass of different isotopes, is estimated to two decimal places. For example, hydrogen has the atomic symbol H, the atomic number 1, and an atomic mass of 1.01. The atomic mass is always larger that the atomic number. For most small elements, the atomic mass is approximately double the atomic number, as the number of protons and neutrons is

about equal.

The elements are divided into three categories: metals, nonmetals and metalloids. These form a diagonal line from period two, group thirteen to period seven, group sixteen. All elements to the left of the metalloids are metals, and all elements to the right are nonmetals.

The periodic table was created to help chemists better understand elements and how they function. It is a map to elemental behavior.

Isotopes

Every atom of a specific element must have the same number of protons in its nucleus. This number is its atomic number. However, there is a range of possible numbers of neutrons it can have in its nucleus. The fact that atoms of a chemical element may have different numbers of neutrons results in each chemical element having several isotopes. Isotopes are atoms of a given chemical element that have different numbers of neutrons in their nuclei.

For example, while all atoms of the element oxygen have eight protons in their nuclei, those oxygen atoms may have eight, nine, or ten neutrons. The different numbers of neutrons in the nucleus distinguishes the three isotopes of oxygen. Oxygen-16 is the isotope of oxygen with 8 neutrons in its nucleus. The number 16 is called the atomic mass number. The atomic mass number is the total number of protons and neutrons in the nucleus of an isotope. From this definition, and knowing that all oxygen atoms have 8 protons in the nucleus, you can deduce that oxygen-17 is the oxygen isotope with 9 neutrons and oxygen-18 is the oxygen isotope with 10 neutrons. Abbreviated into symbols, the three isotopes of oxygen are written as ^{16}O, ^{17}O and ^{18}O.

Isotopes are not very important for understanding minerals, but are important in understanding how to apply chemistry and nuclear physics to geology, such as how to use measurements of radioactive isotopes to measure the ages of rocks and minerals and how to use oxygen isotopes from layers of glacial ice to determine what the temperature of the earth was during an ice age.

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Basics - Minerals. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/minerals.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*
- Periodic Table of Elements. **Provided by**: OpenStax. **Located at**: http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@9.74:6/Biology. **License**: *CC BY: Attribution*

All rights reserved content

- The Periodic Table: Crash Course Chemistry #4. **Authored by**: CrashCourse. **Located at**: https://youtu.be/0RRVV4Diomg. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

This page titled 4.2.1: The Building Blocks of Matter is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.2.2: Atomic Bonds

Chemical Reactions

Minerals form as a result of chemical reactions. Chemical reactions are driven mainly by the arrangement and rearrangement of electrons in atoms. In a mineral, the atoms are held together by chemical bonds, which derive from the electrons.

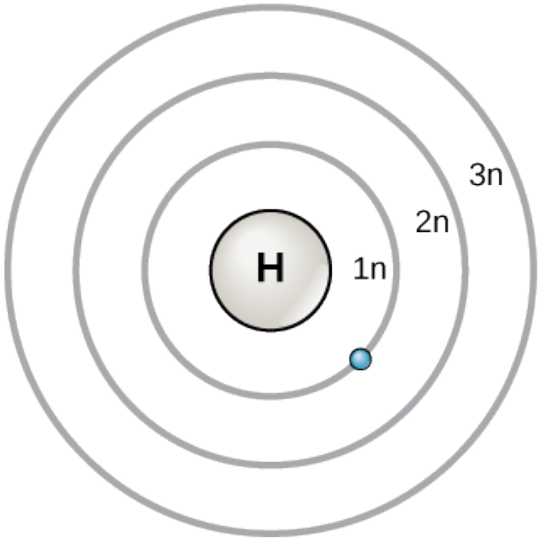


Figure 1. The Bohr model of the atom.

Electrons can be thought of as occupying energy levels, or shells, in an atom. The lowest-energy shell is closest to the nucleus. Each shell can accommodate only a limited number of electrons. The first shell can hold two electrons, the second and third shells can hold eight electrons, and the next two shells can hold eighteen electrons. Unless energy is added to an atom to excite it from its low-energy state, the electrons in the atom will occupy the lowest-energy shells available to them.

The Bohr model (see figure 1) was developed by Niels Bohrs in 1913. In this model, electrons exist within principal shells. An electron normally exists in the lowest energy shell available, which is the one closest to the nucleus. Energy from a photon of light can bump it up to a higher energy shell, but this situation is unstable, and the electron quickly decays back to the ground state.

Chemical Bonds

If atoms interact with other atoms, they can gain or lose electrons to the other atoms, or share electrons with other atoms. In an individual atom, the most stable arrangement is a full outer shell of electrons. Therefore, chemical reactions will occur, and chemical bonds will form that hold atoms together to each other, when atoms encounter other atoms and change their electron configurations toward more stable, lower-energy arrangements, which generally involves achieving full outer electron shells in the atoms.

This stable configuration—a full outer shell of electrons—is exemplified by the inert gases. In the periodic table the inert gases are the elements of group 18 or VIIIA, the last column on the right. Inert gases do not have to undergo any chemical reactions or form any chemical bonds with other atoms in order to have a full outer shell of electrons. The inert gases already have full outer shells of electrons. That is why they are chemically inert. Their electrons are as stable as can be arranged. For this reason, inert gases are extremely unlikely to undergo any chemical reactions and it is almost impossible for them to bond with any other atoms. Because they do not bond with any other atoms to form a liquid, a solid, a molecule, or a mineral, the inert gases consist of atoms that stay separate from each other, in the gas state.

Individual atoms of all the other chemical elements, when they are neutral atoms, do not have full outer shells of electrons like the inert gases do. Therefore, they do not have the most stable arrangement of electrons that they possibly can. That is why most chemical elements have a strong tendency to either gain or lose electrons, or to enter into other arrangements of their valence

electrons, the electrons in their outer shell. Chemical reactions and chemical bonds are generally a result of electrons being rearranged within and among atoms to give the atoms full outer electron shells.

For an atom to lose or gain one electron takes less energy than to lose or gain two, which in turn takes less energy than to lose or gain a third electron. For an individual atom to gain or lose four electrons will only occur in extremely high-energy environments such as in a star. In common chemical reactions on earth, and in the formation of chemical bonds, no element will completely gain or lose four electrons. This limits the charges of atomic cations to +1, +2 or +3 and the charges of atomic anions to –1, –2, or –3.

Reading this far, you have learned about one group of elements in the periodic table, group 18, the inert gases. Another group of chemical elements in the periodic table is the alkali elements. The alkali elements compose group 1 or IA, the left hand column, including the elements sodium (Na) and potassium (K).

Hydrogen is not usually considered as an alkali element because, even though it is shown in group 1 in the periodic table. Hydrogen is so light and small, with just a single proton in its nucleus, that it has some unique behaviors and is considered in a class by itself.

The alkali elements have a single electron in their outer electron shell. If an alkali element loses a single electron, it becomes an ion with a +1 charge and a full outer shell. If an opportunity arises, alkali elements will easily turned into +1 cations.

Ionic bonds

Now look at group 17 or VIIA in the periodic table, which includes the chemical elements fluorine (F), chlorine (Cl) and so on. These are the halogen elements. If a halogen element gains a single electron, it becomes an ion with a–1 charge and a full outer electron shell. If an opportunity arises, halogen elements have a strong tendency to take in an extra electron and become -1 anions because by doing so they achieve a full outer shell of electrons, which is the most stable arrangement of electrons possible.

If sodium and chlorine atoms get together in the right conditions, such as in an evaporating solution of salt water, each sodium atoms will give up an electron to a chlorine atom. This turns the sodium atoms into sodium ions, Na^+, and the chlorine atoms into chloride ions, Cl^-. Opposite electrical charges attract, so the sodium ions and chloride ions will tend to stick together with each other, joined by what are called ionic bonds.

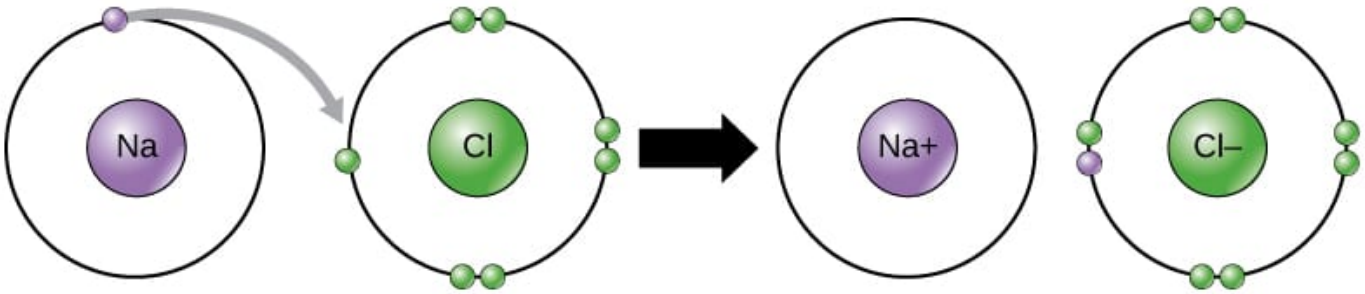


Figure 2. In the formation of an ionic compound, metals lose electrons and nonmetals gain electrons to achieve an octet. Ionic bonds are formed between ions with opposite charges. For instance, positively charged sodium ions and negatively charged chloride ions bond together to make crystals of sodium chloride, or table salt, creating a crystalline molecule with zero net charge.

Not only will the sodium and chloride ions have a very strong tendency to join together with each other via ionic bonds, in most situations they will naturally arrange into a configuration where there is no wasted space and no wasted energy. This leads them to form the crystal lattice of the mineral halite. Halite is a mineral with the chemical formula NaCl, sodium chloride, in which the bonds between the atoms are all ionic bonds.

Look at the diagram of halite showing the sodium and chloride ions arranged into the crystal lattice. All the ionic bonds are at the same angle and the same distance, so they are all of equal strength. This is the lowest-energy arrangement of the ions, the most stable arrangement. If any of the ions were spaced located at different angles or at different distances, there would be extra energy available. This extra energy would drive the ions toward equal angles and distances from each other, until the extra energy is used up and the ions are arranged into their lowest energy state. That is why minerals form, as a natural way for atoms to arrange themselves into the lowest energy state currently available to them.

Covalent bonds

Some elements, such as carbon (C) and silicon (Si) have a half-full valence shell. (The valence shell is another name for the outer shell, where the most reactive electrons are.) If an element such as carbon were to gain 4 electrons or lose 4 electrons, it would have a full valence shell. However, it is very difficult for an atom to gain or lose four electrons—the energy barrier becomes too strong. Therefore, carbon and silicon, along with a few other elements, tend to form a different type of bond in which they share their outer electrons with other atoms, which in turn share their outer electrons with the carbon (or silicon) atom. The atoms all end up with a full outer shell of electrons, even though some or all of those electrons are being shared with neighboring atoms. This electron sharing keeps the atoms bonded together. This type of chemical bond is called a covalent bond.

It is not uncommon for covalent bonds to be relatively strong. An extreme example can be in diamond. Diamond is a mineral consisting of nothing but carbon atoms, so its chemical formula is simply C. Each carbon atom in the diamond crystal lattice is covalently bonded to—sharing its valence electrons with—four neighboring carbon atoms. A diamond crystal is held together by nothing but extremely strong covalent bonds in all directions, which makes diamond a very hard mineral, the hardest known.

Metallic bonds

Gold forms a naturally occurring mineral of more or less pure gold, Au, held together by another type of bond, the metallic bond. Metallic elements such as gold and copper, when they bond with other metallic elements, are sharing some of their electrons not just with adjacent atoms, but throughout the whole substance. That is why metallic substances such as copper, gold, and aluminum make such good electrical conductors, because it is so easy to get the “loose” electrons to respond through the whole extent of the metal.

Hydrogen bonds

Another type of chemical bond that occurs in some minerals is the hydrogen bond. Hydrogen bonds are caused by the positive and negative ends of polar molecules attracting each other strongly enough to hold each other in fixed positions. For example, water molecules can join together through hydrogen bonds to form the mineral known as ice. In a water molecule, H_2O, each of the hydrogen atoms forms a covalent bond with the oxygen atom.

Contributors and Attributions

CC licensed content, Shared previously

- Basics - Minerals. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/minerals.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*
- Atoms, Isotopes, Ions, and Molecules: The Building Blocks. **Provided by**: OpenStax. **Located at**: http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@10.8. **Project**: Biology. **License**: *CC BY: Attribution*

This page titled 4.2.2: Atomic Bonds is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.2.3: Physical Characteristics of Minerals

What are Minerals?

All rocks except obsidian and coal are made of minerals. (Obsidian is a volcanic rock made of glass and coal is made of organic carbon.) Most rocks contain several minerals in a mixture characteristic of the particular rock type. When identifying a rock you must first identify the individual minerals that make up that rock.

Minerals are naturally occurring, inorganic solids with a definite chemical composition and a crystal lattice structure. Although thousands of minerals in the earth have been identified, just ten minerals make up most of the volume of the earth's crust—plagioclase, quartz, orthoclase, amphibole, pyroxene, olivine, calcite, biotite, garnet, and clay.

Together, the **chemical formula** (the types and proportions of the chemical elements) and the **crystal lattice** (the geometry of how the atoms are arranged and bonded together) determine the physical properties of minerals.

The chemical formula and crystal lattice of a mineral can only be determined in a laboratory, but by examining a mineral and determining several of its physical properties, you can identify the mineral. First, you need to become familiar with the physical properties of minerals and how to recognize them.

Minerals can be identified by their physical characteristics. The physical properties of minerals are related to their chemical composition and bonding. Some characteristics, such as a mineral's hardness, are more useful for mineral identification. Color is readily observable and certainly obvious, but it is usually less reliable than other physical properties.

How are Minerals Identified?



Figure 1. This mineral has shiny, gold, cubic crystals with striations, so it is pyrite.

Mineralogists are scientists who study minerals. One of the things mineralogists must do is identify and categorize minerals. While a mineralogist might use a high-powered microscope to identify some minerals, most are recognizable using physical properties.

Check out the mineral in figure 1. What is the mineral's color? What is its shape? Are the individual crystals shiny or dull? Are there lines (striations) running across the minerals?

Color, Streak, and Luster

Diamonds are popular gemstones because the way they reflect light makes them very sparkly. Turquoise is prized for its striking greenish-blue color. Notice that specific terms are being used to describe the appearance of minerals.

Color



Figure 2. This mineral is shiny, very soft, heavy, and gold in color, and is actually gold.

Color is often useful, but should not be relied upon. Different minerals may be the same color. Real gold, as seen in figure 2, is very similar in color to the pyrite in figure 1.

Additionally, Some minerals come in many different colors. Quartz, for example, may be clear, white, gray, brown, yellow, pink, red, or orange. So color can help, but do not rely on color as the determining property. Figure 3 shows one sample of quartz that is colorless and another quartz that is purple. A tiny amount of iron makes the quartz purple. Many minerals are colored by chemical impurities.



Figure 3. Purple quartz, known as amethyst, and clear quartz are the same mineral despite the different colors.

Luster

Luster describes the reflection of light off a mineral's surface. Mineralogists have special terms to describe luster. One simple way to classify luster is based on whether the mineral is metallic or non-metallic. Minerals that are opaque and shiny, such as pyrite, have a metallic luster. Minerals such as quartz have a non-metallic luster.

Luster is how the surface of a mineral reflects light. It is not the same thing as color, so it crucial to distinguish luster from color. For example, a mineral described as "shiny yellow" is being described in terms of luster ("shiny") and color ("yellow"), which are two different physical properties. Standard names for luster include metallic, glassy, pearly, silky, greasy, and dull. It is often useful to first determine if a mineral has a metallic luster. A metallic luster means shiny like polished metal. For example cleaned polished pieces of chrome, steel, titanium, copper, and brass all exhibit metallic luster as do many other minerals. Of the nonmetallic lusters, glassy is the most common and means the surface of the mineral reflects light like glass. Pearly luster is important in identifying the feldspars, which are the most common type of mineral. Pearly luster refers to a subtle irridescence or color play in the reflected light, same way pearls reflect light. Silky means relecting light with a silk-like sheen. Greasy luster looks similar to the luster of solidified bacon grease. Minerals with dull luster reflect very little light. Identifying luster takes a little

practice. Remember to distinguish luster from color.

Different types of non-metallic luster are described in table 1.

Table 1. Six types of non-metallic luster.

Luster	Appearance
Adamantine	Sparkly
Earthy	Dull, clay-like
Pearly	Pearl-like
Resinous	Like resins, such as tree sap
Silky	Soft-looking with long fibers
Vitreous	Glassy

Can you match the minerals in figure 4 with the correct luster from table 1?



Figure 4. (a) Diamond has an adamantine luster. (b) Quartz is not sparkly and has a vitreous, or glassy, luster. (c) Sulfur reflects less light than quartz, so it has a resinous luster.

Streak



Figure 5. The streak of hematite across an unglazed porcelain plate is red-brown.

Streak is the color of a mineral's powder. Streak is a more reliable property than color because streak does not vary. Minerals that are the same color may have a different colored streak. Many minerals, such as the quartz in the figure 3, do not have streak.

To check streak, scrape the mineral across an unglazed porcelain plate (Figure 5). Yellow-gold pyrite has a blackish streak, another indicator that pyrite is not gold, which has a golden yellow streak.

Specific Gravity

Density describes how much matter is in a certain amount of space: density = mass/volume.

Mass is a measure of the amount of matter in an object. The amount of space an object takes up is described by its volume. The density of an object depends on its mass and its volume. For example, the water in a drinking glass has the same density as the water in the same volume of a swimming pool.

The specific gravity of a substance compares its density to that of water. Substances that are more dense have higher specific gravity.

Hardness

Hardness is the strength with which a mineral resists its surface being scraped or punctured. In working with hand samples without specialized tools, mineral hardness is specified by the Mohs hardness scale. The Mohs hardness scale is based 10 reference minerals, from talc the softest (Mohs hardness of 1), to diamond the hardest (Mohs hardness of 10). It is a relative, or nonlinear, scale. A hardness of 2.5 simply means that the mineral is harder than gypsum (Mohs hardness of 2) and softer than calcite (Mohs hardness of 3). To compare the hardness of two minerals see which mineral scratches the surface of the other.

Table 2. Mohs Hardness Scale

Hardness	Index Minerals	Common Objects
1	talc	
2	gypsum	2.5-fingernail
3	calcite	3.5-pure, untarnished copper
4	fluorite	
5	feldspar	5 to 5.5-stainless steel
		5.5 to 6-glass
6	apatite	6 to 6.5-hard steel file
7	quartz	
8	topaz	
9	corundum	
10	diamond	

With a Mohs scale, anyone can test an unknown mineral for its hardness. Imagine you have an unknown mineral. You find that it can scratch fluorite or even feldspar, but apatite scratches it. You know then that the mineral's hardness is between 5 and 6. Note that no other mineral can scratch diamond.

Cleavage and Fracture

Breaking a mineral breaks its chemical bonds. Since some bonds are weaker than other bonds, each type of mineral is likely to break where the bonds between the atoms are weaker. For that reason, minerals break apart in characteristic ways.

Cleavage

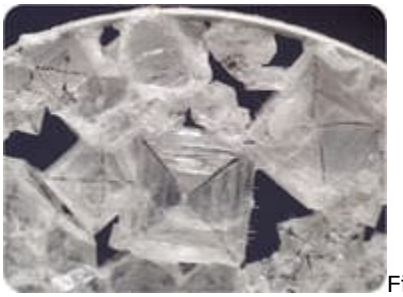


Figure 6. A close-up view of sodium chloride in a water bubble aboard the International Space Station.

Cleavage is the tendency of a mineral to break along certain planes to make smooth surfaces. Halite breaks between layers of sodium and chlorine to form cubes with smooth surfaces (figure 6).

A mineral that naturally breaks into perfectly flat surfaces is exhibiting cleavage. Not all minerals have cleavage. A cleavage represents a direction of weakness in the crystal lattice. Cleavage surfaces can be distinguished by how they consistently reflect light, as if polished, smooth, and even. The cleavage properties of a mineral are described in terms of the number of cleavages and, if more than one cleavage, the angles between the cleavages. The number of cleavages is the number or directions in which the mineral cleaves. A mineral may exhibit 100 cleavage surfaces parallel to each other. Those represent a single cleavage because the surfaces are all oriented in the same diretion. The possible number of cleavages a mineral may have are 1,2,3,4, or 6. If more than 1 cleavage is present, and a device for measuring angles is not available, simply state whether the cleavages intersect at 90° or not 90°.

To see mineral cleavage, hold the mineral up beneath a strong light and move it around, move it around some more, to see how the different sides reflect light. A cleavage direction will show up as a smooth, shiny, evenly bright sheen of light reflected by one set of parallel surfaces on the mineral.

Mica has cleavage in one direction and forms sheets (figure 7).



Figure 7. Sheets of mica.


Figure 8. This rough diamond shows its octahedral cleavage.

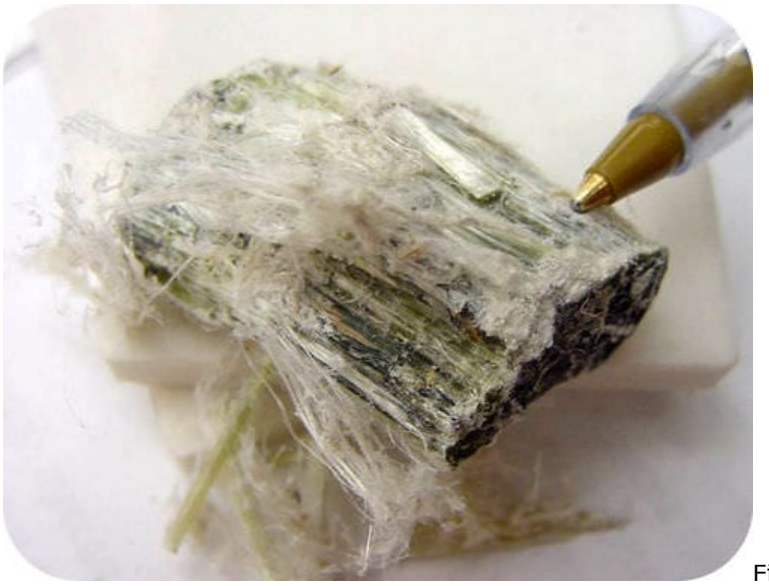
Minerals can cleave into polygons. Fluorite forms octahedrons (figure 8).

One reason gemstones are beautiful is that the cleavage planes make an attractive crystal shape with smooth faces.

Fracture

Fracture is a break in a mineral that is not along a cleavage plane. Fracture is not always the same in the same mineral because fracture is not determined by the structure of the mineral.

Minerals may have characteristic fractures (figure 9). Metals usually fracture into jagged edges. If a mineral splinters like wood, it may be fibrous. Some minerals, such as quartz, form smooth curved surfaces when they fracture.


Figure 9. Chrysotile has splintery fracture.

All minerals have fracture. Fracture is breakage, which occurs in directions that are not cleavage directions. Some minerals, such as quartz, have no cleavage whatsoever. When a mineral with no cleavage is broken apart by a hammer, it fractures in all directions. Quartz is said to exhibit conchoidal fracture. Conchoidal fracture is the way a thick piece of glass breaks with concentric, curving ridges on the broken surfaces. However, some quartz crystals have so many flaws that instead of exhibiting conchoidal fracture they simply exhibit irregular fracture. Irregular fracture is a standard term for fractures that do not exhibit any of the qualities of the other fracture types. In introductory geology, the key fracture types to remember are irregular, which most minerals exhibit, and conchoidal, seen in quartz.

Crystal Shape

All minerals are crystalline, but only some have the opportunity to exhibit the shapes of their crystals, their crystal forms. Many

minerals in an introductory geology lab do not exhibit their crystal form. If a mineral has space while it grows, it may form natural crystals, with a crystal shape reflecting the geometry of the mineral's internal crystal lattice. The shape of a crystal follows the symmetry of its crystal lattice. Quartz, for instance, forms six-sided crystals, showing the hexagonal symmetry of its crystal lattice. There are two complicating factors to remember here: (1) minerals do not always form nice crystals when they grow, and (2) a crystal face is different from a cleavage surface. A crystal face forms during the growth of the mineral. A cleavage surface is formed when the mineral is broken.

Other Identifying Characteristics

There are some properties that only help to distinguish a small number of minerals, or even just a single mineral. An example of such a special property is the effervescent reaction of calcite to a weak solution of hydrochloric acid (5% HCl). Calcite fizzes or effervesces as the HCl solution dissolves it and creates CO_2 gas. Calcite is easy to identify even without testing the reaction to HCl, by its hardness, luster and cleavage.

Another special property is magnetism. This can be tested by seeing if a small magnet responds to the mineral. The most common mineral that is strongly magnetic is the mineral magnetite. A special property that shows up in some sample of plagioclase feldspar is its tendency to exhibit striations on cleavage surfaces. Striations are perfectly straight, fine, parallel lines. Magnification may be required to see striations on plagioclase cleavage surfaces. Other special properties may be encountered on a mineral to mineral basis.

Some minerals have other unique properties, some of which are listed in table 3. Can you name a unique property that would allow you to instantly identify a mineral that's been described quite a bit in this chapter? (Hint: It is most likely found on your dinner table.)

Table 3. Some minerals have unusual properties that can be used for identification.

Property	Description	Example of Mineral
Fluorescence	Mineral glows under ultraviolet light	Fluorite
Magnetism	Mineral is attracted to a magnet	Magnetite
Radioactivity	Mineral gives off radiation that can be measured with Geiger counter	Uraninite
Reactivity	Bubbles form when mineral is exposed to a weak acid	Calcite
Smell	Some minerals have a distinctive smell	Sulfur (smells like rotten eggs)
Taste	Some minerals taste salty	Halite

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- 3.3: Mineral Identification. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/3.3/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Basics - Minerals. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/minerals.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.2.3: Physical Characteristics of Minerals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.2.4: Classifying Minerals

Classes of Minerals

Minerals are classified according to their chemical properties. Except for the native element class, the chemical basis for classifying minerals is the anion, the negatively charged ion that usually shows up at the end of the chemical formula of the mineral. For example, the sulfides are based on the sufur ion, S^{2-}. Pyrite, for example, FeS_2, is a sulfide mineral. In some cases, the anion is of a mineral class is polyatomic, such as $(CO_3)^{2-}$, the carbonate ion. The major classes of minerals are:

- silicates
- sulfides
- carbonates
- oxides
- halides
- sulfates
- phosphates
- native elements

Silicates

Based on the polyatomic anion, $(SiO_4)^{4-}$, which has a tetrahedral shape. Most minerals in the earth's crust and mantle are silicate minerals. All silicate minerals are built of silicon-oxygen tetrahedra $(SiO_4)^{4-}$ in different bonding arrangements which create different crystal lattices. You can understand the properties of a silicate mineral such as crystal shape and cleavage by knowing which type of crystal lattice it has.

- In **nesosilicates**, also called island silicates, the silicate tetrahedra are separate from each other and bonded completely to non silicate atoms. Olivine is an island silicate.
- In **sorosilicates** or paired silicates, such as epidote, the silicate tetrahedra are bonded in pairs.
- In **cyclosilicates**, also called ring silicates, the silicate tetrahedra are joined in rings. Beryl or emerald is a ring silicate.
- In **phyllosilicates** or sheet silicates, the tetrahedra are bonded at three corners to form flat sheets. Biotite is a sheet silicate.
- In **single-chain inosilicates** the silicate tetrahedra are bonded in single chains. Pyroxenes are singele-chain inosilicates.
- In **double-chain inosilicates** the silicate tetrahedra are bonded in double chains. Amphiboles are double-chain inosilicates.
- In **tectosilicates**, also known as framework silicates, all corners of the silicate tetrahedra are bonded to corners of other silicate tetrahedra, forming a complete framework of silicate tetrahedra in all directions. Feldspar, the most common mineral in earth's crust, and quartz are both framework silicates.

Sulfides

These are based on the sulfide ion, S^{2-}. Examples include pyrite, FeS_2, galena, PbS, and sphalerite, ZnS in its pure zinc form. Some sulfides are mined as sources of such metals as zinc, lead, copper, and tin.

Carbonates

These are based on the carbonate ion, $(CO_3)^{2-}$. Calcite, $CaCO_3$, and dolomite, $CaMg(CO_3)2$, are carbonate minerals. Carbonate minerals tend to dissolve relatively easily in water, especially acid water, and natural rain water is slightly acid.

Oxides

These are based on the oxygen anion, O^{2-}. Examples include iron oxides such as hematite, Fe_2O_3 and magnetite, Fe_3O_4, and pyrolusite, MgO.

Halides

These have a halogen element as the anion, whether it be fluoride, F^-, chloride, Cl^-, bromide, Br^-, iodide, I^-, or astatide, At^-. Halite, NaCl, is a halide mineral.

Sulfates

These have the polyatomic sulfate ion, $(SO_4)^{2-}$, as the anion. Anhydrite, $CaSO_4$, is a sulfate.

Phosphates

These have the polyatomic phosphate ion, $(PO_4)^{3-}$, as the anion. Fluorapatite, $Ca_5(PO_4)_3F$, which makes your teeth hard, is a phosphate mineral.

Native Elements

These are made of nothing but a single element. Gold (Au), native copper (Cu), and diamond and graphite, which are made of carbon, are all native element minerals. Recall that a mineral is defined as naturally occurring. Therefore, elements purified and crystallized in a laboratory do not qualify as minerals, unless they have also been found in nature.

Mineral Classification Tables

In tables 1–3, hardness is measured on Mohs Hardness Scale. As you read through the tables, you can click on the images of minerals to see a larger version of the photo.

Table 1. Nonmetallic Luster—Light Color

Typical Color	Hardness	Cleavage/ Fracture	Mineral Name	Photo of Mineral
colorless	7	conchodial fracture	quartz	

Table 1. Nonmetallic Luster—Light Color				
variable	7	conchodial fracture	chalcedony (chert, etc.)	
pink or white	5–6	2 planes at right angles	orthoclase (feldspar)	
white	5–6	2 planes at right angles	Na-plagioclase (feldspar)	

Table 1. Nonmetallic Luster—Light Color				
white to gray	5–6	2 planes at right angles	Ca-plagioclase (feldspar)	
variable	4	4 planes	fluorite	
colorless or white	3	3 planes at odd angles	calcite	

Table 1. Nonmetallic Luster—Light Color				
pink or white	3	3 planes at odd angles	dolomite	
colorless or white	2.5–3	3 planes at odd angles	halite	
colorless or white	2.5	1 plane	muscovite	

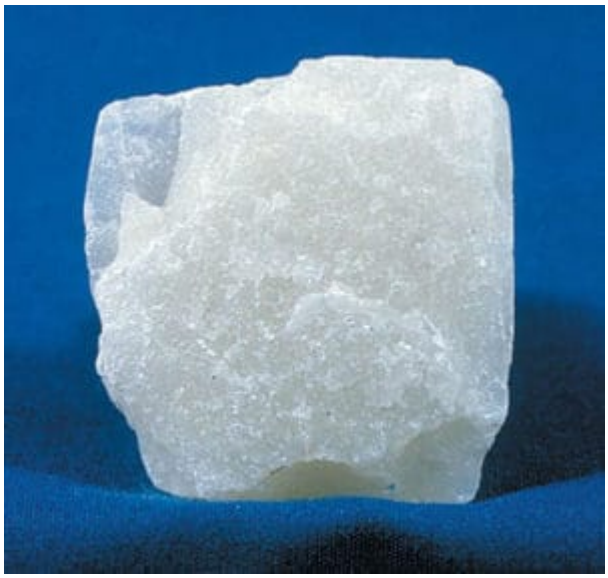
Table 1. Nonmetallic Luster—Light Color				
colorless or white	2	2 planes at right angles	gypsum	
variable	1	1 plane	talc	
white	< 1	uneven (turns to powder)	kaolinite	

Table 2. Nonmetallic Luster—Dark Color				
Typical Color	**Hardness**	**Cleavage/ Fracture**	**Mineral Name**	**Photo of Mineral**

Table 2. Nonmetallic Luster—Dark Color				
green	5–6	irregular	olivine	
red	5–6	irregular	garnet	
red	3–6	irregular	hematite	

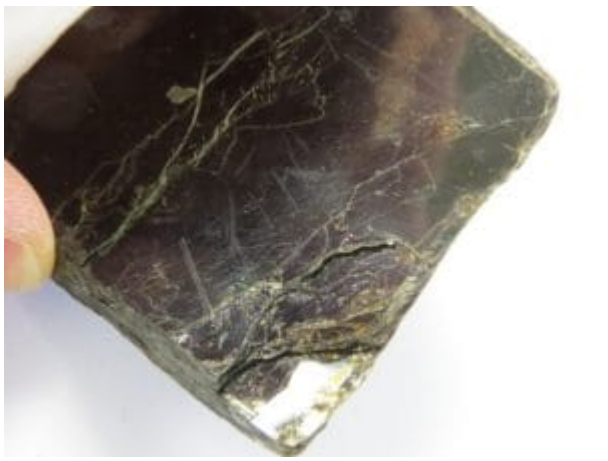
Table 2. Nonmetallic Luster—Dark Color				
dark green	3–6	2 planes at right angles	pyroxene	
black	4.5–6	2 planes at odd angles	hornblende (amphibole)	
black	2.5	1 plane	biotite	

Table 2. Nonmetallic Luster—Dark Color				
green	2	1 plane	chlorite	

Table 3. Metallic Luster				
Typical Color	**Hardness**	**Cleavage/ Fracture**	**Mineral Name**	**Photo of Mineral**
black or dark gray	6	irregular	magnetite	
brassy yellow	6	irregular	pyrite	

Table 3. Metallic Luster				
coppery yellow	4	irregular	chalcopyrite	
silver	3	3 planes at right angles	galena	

How to Identify Minerals

First, you need good light and a hand lens or magnifying glass. A hand lens is a small, double-lens magnifying glass that has a magnification power of at least 8× and can be purchased at some bookstores and nature stores.

Minerals are identified on the basis of their physical properties, which have been described in the the previous section. To identify a mineral, you look at it closely. At a glance, calcite and quartz look similar. Both are usually colorless, with a glassy luster. However, their other properties they are completely different. Quartz is much harder, hard enough to scratch glass. Calcite is soft, and will not scratch glass. Quartz has no mineral cleavage and fractures the same irregular way glass breaks. Calcite has three cleavage directions which meet at angles other than 90°, so it breaks into solid pieces with perfectly flat, smooth, shiny sides.

When identifying a mineral, you must:

1. Look at it closely on all visible sides to see how it reflects light
2. Test its hardness
3. Identify its cleavage or fracture
4. Name its luster
5. Evaluate any other physical properties necessary to determine the mineral's identity

In the minerals tables that accompanies this section, the minerals are grouped according to their luster and color. They are also classified on the basis of their hardness and their cleavage or fracture. If you can identify several of these physical properties, you can identify the mineral.

A simple lesson on how to identify minerals is seen in this video.

A link to an interactive elements can be found at the bottom of this page.

Contributors and Attributions

CC licensed content, Shared previously

- Basics - Minerals. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/minerals.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

All rights reserved content

- Rocks & Minerals : Identifying Minerals. **Authored by**: eHow. **Located at**: https://youtu.be/JeFVwqBuYl4. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

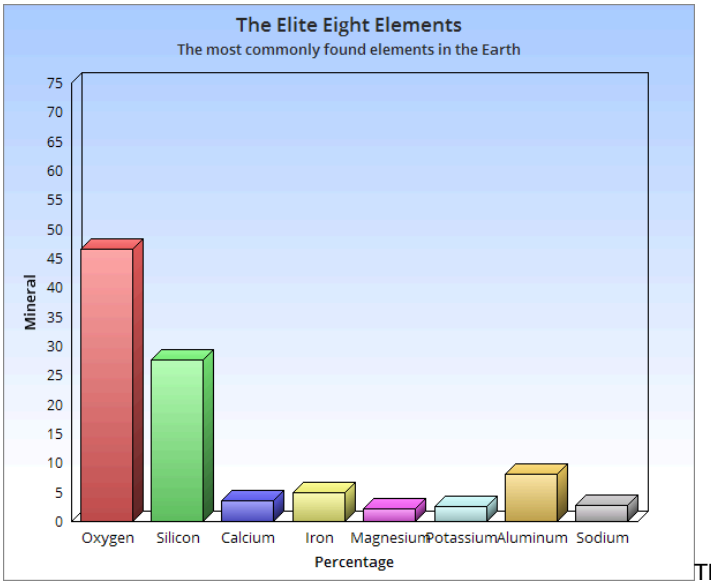
Public domain content

- Talc block. **Provided by**: USGS. **Located at**: https://en.Wikipedia.org/wiki/File:Talc_block.jpg. **License**: *Public Domain: No Known Copyright*
- Chlorite. **Provided by**: USGS. **Located at**: https://commons.wikimedia.org/wiki/File:ChloriteUSGOV.jpg. **License**: *Public Domain: No Known Copyright*

This page titled 4.2.4: Classifying Minerals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.2.5: Abundance of Elements in Earth's Crust

The table shows the **abundance of elements in Earth's crust**.



The Elite Eight Elements

Contributors and Attributions

CC licensed content, Original

- **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 4.2.5: Abundance of Elements in Earth's Crust is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.2.6: Silicate Minerals and the Silicate Tetrahedron

Most of the minerals in the earth are silicate minerals. The building block of silicate minerals—the essential component that makes them silicate minerals—is the silicate tetrahedron. The silicate tetrahedron consists of four oxygen atoms arranged as close as they can get around a central silicon atom. The result is a pyramidal shape known as a tetrahedron, with an oxygen atom at each of its four apices. (The apices are the points on the tetrahedron where three corners come together.)

The silicon atom by itself has four electrons in its outer shell. In the silicate tetrahedron each of those four electrons is being shared with one of the four attached oxygen atoms. In turn, each oxygen atom is sharing one of the 6 electrons it has in its outer shell.

The result is that the silicon at the center of the tetrahedron has, in effect, a full outer shell with eight electrons in it. Those eight electrons are shared, in pairs, with the four oxygen atoms of the tetrahedron. Each oxygen atom in the tetrahedron, in turn, will have seven electrons in its outer shell—if there is nothing more to the system than the one silicon atom bonded to the four oxygen atoms. This would leave each oxygen atom one electron short of having a full outer shell of electrons.

However, oxygen is a strongly electronegative element, which means that it has the strength to attract electrons from other elements to its nucleus in most situations. In a mineral, each oxygen atom in the silicate tetrahedron will actually have eight electrons: the six electrons that each oxygen atom had in its outer electron shell to begin with, the electron it gained by sharing a pair of electrons in a covalent bond with the silicon atom in the tetrahedron, and one more electron from another atom (or another small group of atoms) in the mineral, outside the tetrahedron.

Silicate tetrahedra are able to bond with many common elements in many different crystal lattice arrangements. In addition, silicate tetrahedra are able to bond with other silicate tetrahedra in a variety of geometric arrangements, including rings, sheets, chains, and three-dimensional networks.

Contributors and Attributions

CC licensed content, Shared previously

- Basics - Minerals. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/minerals.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.2.6: Silicate Minerals and the Silicate Tetrahedron is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.3: ROCKS AND THE ROCK CYCLE

Learning Objectives

Identify processes of the rock cycle and the different rock types associated with each step.

- Identify igneous rocks and the steps of the rock cycle related to their formation.
- Identify sedimentary rocks and the steps of the rock cycle related to their formation.
- Identify metamorphic rocks and the steps of the rock cycle related to their formation.

4.3.1: Characteristics of Igneous Rocks

4.3.2: How Are Igneous Rocks Classified?

4.3.3: Identifying Igneous Rocks

4.3.4: Uses of Igneous Rocks

4.3.5: Characteristics of Sedimentary Rocks

4.3.6: Characteristics of Metamorphic Rocks

This page titled Chapter 4.3: Rocks and the Rock Cycle is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.3.1: Characteristics of Igneous Rocks

Magma is molten rock inside the earth. It is the source of all igneous rock. Because the earth was largely molten at its origin, magma may be considered the beginning of the rock cycle. Igneous rocks contain information about how they originate. By carefully analyzing igneous rocks and interpreting the information they contain, we can deduce processes that take place within the earth and we can understand volcanic processes that take place on the earth's surface.

The study of igneous rocks enables us to understand the igneous part of geologic history. For example, at the end of the Triassic period, 245 million years ago, the greatest mass extinction ever known took place, wiping out more life forms on earth than the mass extinction that led to the demise of dinosaurs 65 million years ago at the end of the Cretaceous. At the end of the Triassic, a huge amount of basalt erupted onto the earth. Many geologists think that the gases and particles released into the atmosphere by those eruptions may have been a major factor in the end of Triassic mass extinction. Those scientists are studying the information contained in the basalts of that age to further test their hypotheses.

Igneous rocks contain three essential sources of information: their minerals, their overall chemical composition, and their igneous texture. Igneous rock names are based on specific combinations of these features. Igneous rocks also contain isotopic information that is used in determining absolute ages and in further characterizing the origin of the magma. Special equipment and expertise is required to conduct isotopic and precise chemical analyses. Fortunately, with some basic training and practice anyone can learn to identify the minerals, composition and texture of an igneous rock; name the rock; and interpret key information about its origins.

All igneous rocks, other than pure volcanic glass, contain minerals. The minerals provide details on the chemical composition of the rock, and on the conditions in which the magma originated, cooled, and solidified. Geologists conduct chemical analyses of minerals to determine the temperatures and pressures at which they formed and to identify the dissolved gases and chemical elements that were present in the magma.

Most magmas are predominantly silicate liquids, composed largely of silica tetrahedra that have not yet bonded together to become silicate minerals. The chemical composition of an igneous rock tells us about the origin of the magma, beginning with which type of rock melted within the earth to form the magma in the first place, and how deep in the earth the melting occurred. Once magma has formed inside the earth, its composition may be modified. Minerals can grow from the magma and separate from it, changing the chemistry of the remaining liquid. Or, one body of magma can mix with another that has a different composition.

Magmas come in a range of compositions, from rich in silica and poor and iron and magnesium (felsic) to moderate in silica and high in iron and magnesium (mafic). Felsic igneous rocks, as a whole rock, tend to have light colors or shades: white, pink, light brown, light gray. Mafic igneous rocks, on the whole, tend to be dark colored, commonly black or dark gray. Most mafic magma originates by melting of rocks in the mantle that are extremely rich in iron and magnesium. Felsic magma usually originates in the crust or by the shedding of mafic minerals as magma rises through the crust.

The igneous texture tells us how the magma cooled and solidified. Magma can solidify into igneous rock in several different ways, each way resulting in a different igneous texture. Magma may stay within the earth, far below ground level, and crystallize into plutonic igneous rock (also known as intrusive igneous rock). Or, magma may flow out onto surface of the earth as a lava flow. Another way that igneous rock forms is by magma erupting explosively into the air and falling to earth in pieces known as pyroclastic material, also called tephra. Lava flows and pyroclastic material are volcanic igneous rock (also known as extrusive igneous rock).

The igneous texture of a rock is not how it feels in your hand, not whether it is rough or smooth. The igneous texture describes whether the rock has mineral crystals or is glassy, the size of the mineral grains, and the rock's porosity (empty spaces).

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY-SA: Attribution-ShareAlike*

CC licensed content, Shared previously

- Basics - Igneous Rocks. **Authored by**: Ralph L. Dawes, Ph.D. and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: https://commons.wvc.edu/rdawes/G101OCL/Basics/igneous.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*
- Igneous textures. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Igneous_textures. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.3.1: Characteristics of Igneous Rocks is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.3.2: How Are Igneous Rocks Classified?

This page focuses on igneous rocks and gives you the background needed to understand the terms used in the igneous rock classification table (at the bottom of this page).

Intrusive and Extrusive Igneous Rocks

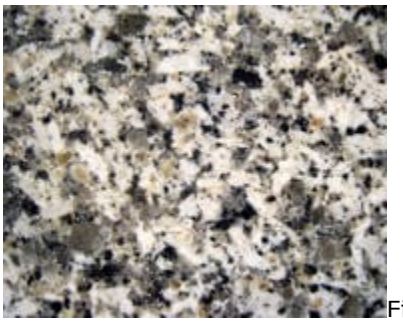


Figure 2. Granite is made of four minerals, all visible to the naked eye: feldspar (white), quartz (translucent), hornblende (black), and biotite (black, platy).

Igneous rocks are called **intrusive** when they cool and solidify beneath the surface. Intrusive rocks form plutons and so are also called plutonic. A **pluton** is an igneous intrusive rock body that has cooled in the crust. When magma cools within the Earth, the cooling proceeds slowly. Slow cooling allows time for large crystals to form, so intrusive igneous rocks have visible crystals. Granite is the most common intrusive igneous rock (see figure 2 for an example).

Igneous rocks make up most of the rocks on Earth. Most igneous rocks are buried below the surface and covered with sedimentary rock, or are buried beneath the ocean water. In some places, geological processes have brought igneous rocks to the surface. Figure 3 below shows a landscape in California's Sierra Nevada made of granite that has been raised to create mountains.

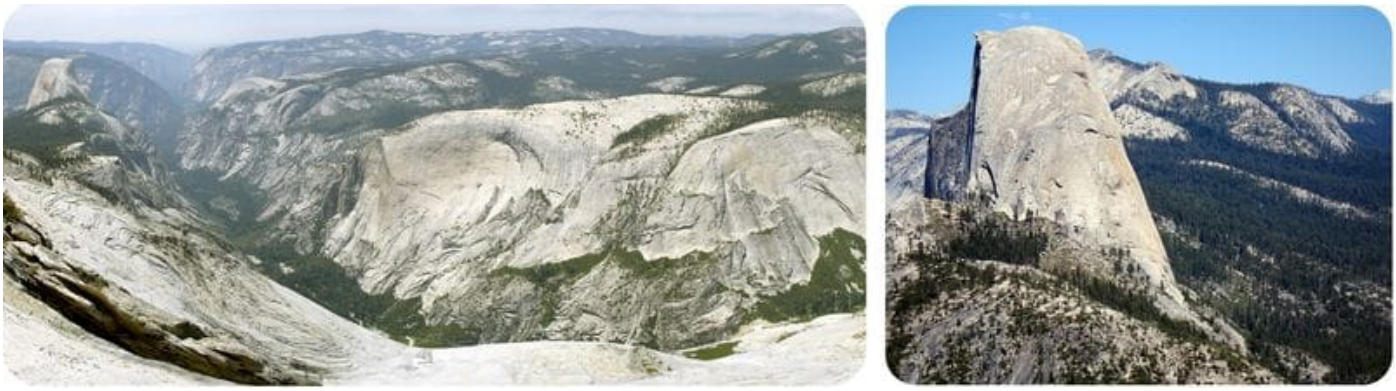


Figure 3. California's Sierra Nevada is intrusive igneous rock exposed at Earth's surface.

Igneous rocks are called **extrusive** when they cool and solidify above the surface. These rocks usually form from a volcano, so they are also called **volcanic rocks** (figure 4).



Figure 4. Extrusive igneous rocks form after lava cools above the surface.



Figure 5. Cooled lava forms basalt with no visible crystals. Why are there no visible crystals?

Extrusive igneous rocks cool much more rapidly than intrusive rocks. There is little time for crystals to form, so extrusive igneous rocks have tiny crystals (figure 5).

Classifying Igneous Rocks

As we just learned, there are two main types of igneous rocks: intrusive rocks (also known as plutonic rocks) and extrusive rocks (also known as volcanic rocks). Volcanic rocks break down into two more categories: (a) lava flows and (b) tephra (pyroclastic material).

Igneous rocks are classified on the basis of their composition and their texture. Magma, and the igneous rock it becomes, has a range of chemical compositions. For example, basalt is a mafic lava flow rock which originates from melting of the upper mantle. The way that magma turns into a solid rock gives it a distinctive igneous texture. For example, magma that becomes a pluton by slowly crystallizing (growing minerals) within the crust will develop a very different texture from magma that becomes an ash flow tuff as a result of semi-molten volcanic ash spewing across a landscape and then settling down and welding itself together into solid rock.

Igneous Rock Textures

Igneous textures include the rock textures occurring in igneous rocks. Igneous textures are used by geologists in determining the mode of origin igneous rocks and are used in rock classification. There are six main types of textures; phaneritic, aphanitic, porphyritic, glassy, pyroclastic and pegmatitic.

Aphanitic (*a* = not, *phaner* = visible) rocks in contrast to phaneritic rocks, typically form from lava which crystallize rapidly on or near Earth' surface. Because extrusive rocks make contact with the atmosphere they cool quickly, so the minerals do not have time to form large crystals. The individual crystals in an aphanitic igneous rock are not distinguishable to the naked eye. Examples of aphanitic igneous rock include basalt, andesite andrhyolite.

Glassy or *vitreous* textures occur during some volcanic eruptions when the lava is quenched so rapidly that crystallization cannot

occur. The result is a natural amorphous glass with few or no crystals. Examples includeobsidian and pumice.

Pegmatitic texture occurs during magma cooling when some minerals may grow so large that they become massive (the size ranges from a few centimetres to several metres). This is typical of pegmatites.

Phaneritic (*phaner* = visible) textures are typical of intrusive igneous rocks, these rocks crystallized slowly below Earth's surface. As magma cools slowly the minerals have time to grow and form large crystals. The minerals in a phaneritic igneous rock are sufficiently large to see each individual crystal with the naked eye. Examples of phaneritic igneous rocks are gabbro, diorite and granite.

Porphyritic textures develop when conditions during cooling of a magma change relatively quickly. The earlier formed minerals will have formed slowly and remain as large crystals, whereas, sudden cooling causes the rapid crystallization of the remainder of the melt into a fine grained (aphanitic) matrix. The result is an aphanitic rock with some larger crystals (phenocrysts) imbedded within its matrix. Porphyritic texture also occurs when magma crystallizes below a volcano but is erupted before completing crystallization thus forcing the remaining lava to crystallize more rapidly with much smaller crystals.

Pyroclastic (*pyro* = igneous, *clastic* = fragment) textures occur when explosive eruptions blast the lava into the air resulting in fragmental, typically glassy material which fall as volcanic ash, lapilli and volcanic bombs.

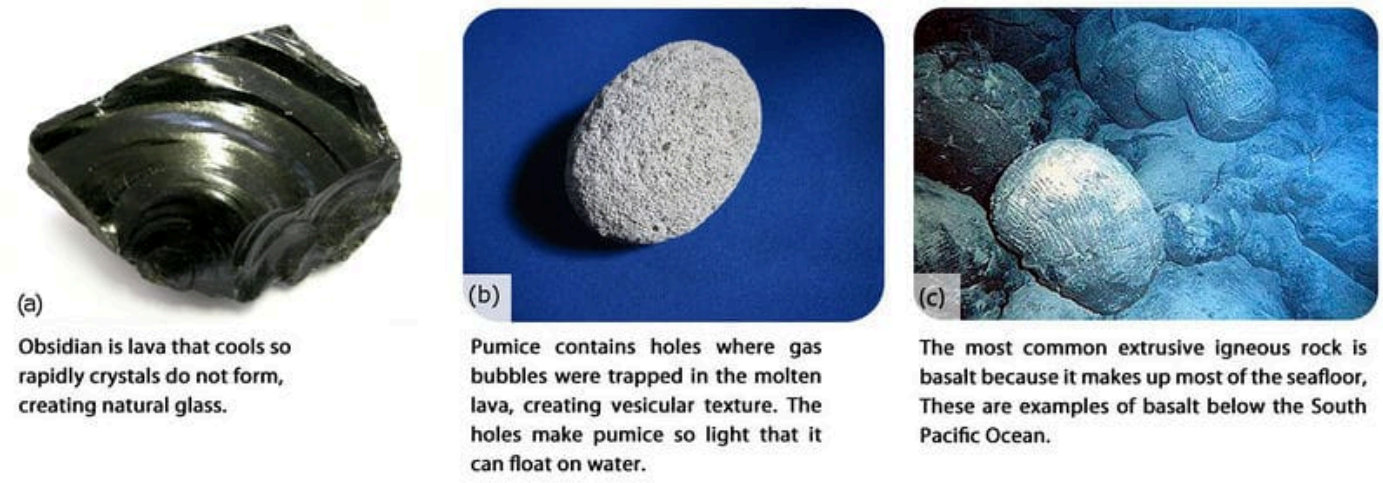


Figure 1. Different cooling rate and gas content resulted in these different textures.

Volcanic Rocks

Let us start with textures associated with rocks formed by lava flows. Magmas that erupt as lava onto the earth's surface cool and solidify rapidly. Rapid cooling results in an aphanitic igneous texture, in which few or none of the individual minerals are big enough to see with the naked eye. This is sometimes referred to as a fine-grained igneous texture.

Some lava flows, however, are not purely fine-grained. If some mineral crystals start growing while the magma is still underground and cooling slowly, those crystals grow to a large enough size to be easily seen, and the magma then erupts as a lava flow, the resulting texture will consist of coarse-grained crystals embedded in a fine-grained matrix. This texture is called porphyritic.

If lava has bubbles of gas escaping from it as it solidifies, it will end up with "frozen bubble holes" in it. These "frozen bubble holes" are called vesicles, and the texture of a rock containing them is said to be vesicular.

If so many bubbles are escaping from lava that it ends up containing more bubble holes than solid rock, the resulting texture is said to be frothy. Pumice is the name of a type of volcanic rock with a frothy texture.

If lava cools extremely quickly, and has very little water dissolved in it, it may freeze into glass, with no minerals (glass by definition is not a mineral, because it does not have a crystal lattice). Such a rock is said to have a glassy texture. Obsidian is the common rock that has a glassy texture, and is essentially volcanic glass. Obsidian is usually black.

Now let us briefly consider textures of tephra or pyroclastic rocks. Like lava flow rocks, these are also extrusive igneous rocks.

However, instead of originating from lava that flowed on the earth's surface, tephra is volcanic material that was hurled through the air during a volcanic eruption.

A pyroclastic rock made of fine-grained volcanic ash may be said to have a fine-grained, fragmental texture. Volcanic ash consists mainly of fine shards of volcanic glass. It may be white, gray, pink, brown, beige, or black in color, and it may have some other fine crystals and rock debris mixed in. The term "fine-grained, fragmental" is easy to confuse with the term fine-grained (aphanitic). An equivalent term that is less ambiguous is tuffaceous. Rocks made of volcanic ash are called tuff.

A pyroclastic rock with many big chunks of material in it that were caught up in the explosive eruption is said to have a coarse-grained, fragmental texture. However, a better word that will avoid confusion is to say it has a brecciated texture, and the rock is usually called a volcanic breccia. The bigger chunks of material in a volcanic breccia are more than 1 cm (5/8 inch) across, and sometimes are much bigger.

Plutonic Rocks

When magma cools slowly underground and solidifies there, it usually grows crystals big enough to be seen easily with the naked eye. These visible crystals comprise the whole rock, not just part of it as in a porphyritic, fine-grained igneous rock. The texture of an igneous rock made up entirely of crystals big enough to be easily seen with the naked eye is phaneritic. Phaneritic texture is sometimes referred to as coarse-grained igneous texture. Granite, the most well known example of an intrusive igneous rock, has a phaneritic texture.

Sometimes an intrusion of magma that is crystallizing slowly underground releases large amounts of hot water. The water is released from the magma as extremely hot fluid with lots of chemical elements dissolved in it. This hydrothermal fluid gets into cracks and voids in the earth's crust, and as it cools it may grow very large minerals from the dissolved chemical elements. A rock consisting of such large minerals is said to have a pegmatitic texture, which means the average mineral size is greater than 1 cm in diameter (and sometimes is much larger). The name of an igneous rock with a pegmatitic texture is pegmatite. Pegmatites are commonly found in or near the margins of bodies of granite.

Igneous Rock Compositions

The most common igneous compositions can be summarized in three words: mafic (basaltic), intermediate (andesitic), and felsic (granitic).

Felsic composition is higher in silica (SiO_2) and low in iron (Fe) and magnesium (Mg). Mafic composition is higher in iron and magnesium and lower in silica. Intermediate compositions contain silica, iron, and magnesium in amounts that are intermediate to felsic and mafic compositions.

Composition and Color

Composition influences the color of igneous rocks. Felsic rocks tend to be light in color (white, pink, tan, light brown, light gray). Mafic rocks tend to be dark in color (black, very dark brown, very dark gray, dark green mixed with black). The color distinction comes from the differences in iron and magnesium content. Iron and, to a lessor extent, magnesium give minerals a darker color. Intermediate igneous rocks tend to have intermediate shades or colors (green, gray, brown).

The association between color and composition is useful because before you can name and interpret an igneous rock you need to determine both its texture AND its composition. If you have an aphanitic igneous rock, which has no crystals big enough to see without a microscope, you can estimate its composition based on its color: pink or nearly white, felsic; medium gray, intermediate; very dark or black, mafic.

This color rule works most of the time but there are two problems that you need to keep in mind. First, the rule does not work for glassy igneous rocks. Obsidian, which is volcanic glass, is usually black, even though it has a felsic composition. That is because a tiny amount of iron, too little to color minerals very darkly, can color glass darkly.

The second problem is that when igneous rocks have been exposed to air and water for a long time, they start to weather, which changes their color. Geologists working in the field carry a rock hammer, so they can break off the weathered, outer parts of rocks to see the "fresh," unweathered rock inside.

If you can see and identify the minerals in an igneous rock, you can gain further information about the igneous composition. Igneous rocks with quartz in them are usually felsic. Igneous rocks with olivine in them are usually mafic. Igneous rocks with neither quartz nor olivine in them are most commonly intermediate.

Origins of Igneous Rocks

Once you have determined the texture and composition of an igneous rock, you can name it and you can also say something important about how it formed. For example, a coarse-grained, felsic igneous rock is not only a granite, it is an intrusive igneous rock that formed from slow cooling and crystallization of a body of magma within the earth's crust. The intrusion of large bodies of granite—batholiths—is usually part of the origin of a mountain range. Similarly, a fine-grained, mafic igneous rock is not only a basalt, it is an extrusive igneous rock that formed from rapid cooling and crystallization of a lava flow at earth's surface.

Igneous Rock Classification

Pegmatitic Texture (Extremely Coarse-Grained) ***Originates from water-rich intrusions, which cool and crystallize underground***		
Composition	**Most Common Minerals**	**Rock Name**
felsic	Na-plagioclase, orthoclase, quartz, biotite, amphibole, muscovite	pegmatite
Phanertitic Texture (Coarse-Grained) ***Originates in deep intrusions, which cool and crystallize slowly underground***		
Composition	**Most Common Minerals**	**Rock Name**
felsic	Na-plagioclase, orthoclase, quartz, biotite, amphibole, muscovite	granite
intermediate	Na-plagioclase, quartz, orthoclase, amphibole, biotite	granodiorite
	Na-plagioclase, amphibole, pyroxene, biotite	diorite
mafic	Ca-plagioclase, pyroxene, olivine, amphibole	gabbro
Aphanitic Texture (Fine-Grained) ***Originates in lava flows (or very shallow intrusions), which cool rapidly***		
Composition	**Most Common Minerals**	**Rock Name**
felsic	Na-plagioclase, orthoclase, quartz, biotite, amphibole, muscovite	rhyolite
intermediate	Na-plagioclase, quartz, orthoclase, amphibole, biotite	dacite
	Na-plagioclase, amphibole, pyroxene, biotite	andesite
mafic	Ca-plagioclase, pyroxene, olivine, amphibole	basalt
Frothy Texture (Porous, Pumiceous) ***Originates in gas-charged volcanic eruptions, commonly pyroclastic***		
Composition	**Most Common Minerals**	**Rock Name**
felsic	glass (may contain a few minerals typical of felsic rocks)	pumice
mafic	glass (may contain a few mineral typical of mafic rocks)	scoria
Note: Basalt with fewer holes, known as vesicles, is called vesicular basalt. *Scoria has more holes and may be black or red in color.*		
Glassy Texture ***Originates from cooling too rapid to allow crystal lattices to form***		
Composition	**Most Common Minerals**	**Rock Name**
felsic to mafic	glass (no minerals)	obsidian
Note: Obsidian that is transparent at thin edges and has good conchoidal fracture is probably felsic.		

Fragmental Texture—Coarse (Contains Large Rock Fragments) *Originates from pyroclastic (explosive) eruptions*		
Composition	**Most Common Minerals**	**Rock Name**
felsic to mafic	variable (depending on rock fragments and ash content)	volcanic breccia
Fragmental Texture—Fine (Mainly Volcanic Ash) *Originates from pyroclastic (explosive) eruptions*		
Composition	**Most Common Minerals**	**Rock Name**
felsic	may contain a few minerals typical of felsic rocks	rhyolitic tuff
medium	may contain a few minerals typical of intermediate rocks	andesitic tuff
mafic	may contain a few minerals typical of mafic rocks	

Contributors and Attributions

CC licensed content, Shared previously

- Basics - Igneous Rocks. **Authored by**: Ralph L. Dawes, Ph.D. and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: https://commons.wvc.edu/rdawes/G101OCL/Basics/igneous.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*
- HS Rocks. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/4.0/. **License**: *Public Domain: No Known Copyright*

This page titled 4.3.2: How Are Igneous Rocks Classified? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.3.3: Identifying Igneous Rocks

Igneous rocks can be distinguished from sedimentary rocks by the lack of beds, lack of fossils, and lack of rounded grains in igneous rocks, and the presence of igneous textures. A granite, for example, can be distinguished from a sandstone because rather than being a mixture of weathered, rounded grains compressed and cemented together, granite consists of a small number of minerals in shiny black, white, or pink colors, with excellent crystal forms, grown together into a completely interlocking pattern. Sandstones, by contrast, have sedimentary bedding (layers) and consist of rounded grains with some spaces between the grains, which you can see with a hand lens or magnifying glass.

Igneous rocks can be distinguished from most regional metamorphic rocks by the lack of foliation (layering) in igneous rocks. Unfoliated metamorphic rocks lack igneous textures and usually contain minerals not found in igneous rocks.

Granite may look like gneiss at first glance, but granite has no layering, no preferred orientation of the minerals. The minerals in a granite grow randomly in all directions, rather than tending to grow parallel to each other.

Igneous rocks are classified on the basis of their texture and their composition. See the previous sections for descriptions of the different igneous textures and compositions.

The igneous rock classification tables that accompany this section are arranged on the basis of igneous textures first, and further broken down on the basis of igneous composition. Remember that igneous composition is estimated on the basis of color: light = felsic composition, medium = intermediate composition, and dark = mafic composition.

Watch this video for an example of identifying igneous rocks:

A link to an interactive elements can be found at the bottom of this page.

Contributors and Attributions

CC licensed content, Shared previously

- Basics - Igneous Rocks. **Authored by**: Ralph L. Dawes, Ph.D. and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: https://commons.wvc.edu/rdawes/G101OCL/Basics/igneous.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

All rights reserved content

- Identifying Rocks : Identifying Igneous Rocks. **Authored by**: eHow. **Located at**: https://youtu.be/Q0XtLjE3siE. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

This page titled 4.3.3: Identifying Igneous Rocks is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.3.4: Uses of Igneous Rocks


Figure 1. This diorite statue was made in approximately 2090 BC.

Igneous rocks have a wide variety of uses. One important use is as stone for buildings and statues. Diorite was used extensively by ancient civilizations for vases and other decorative artwork and is still used for art today (Figure 1).

Granite (figure 2) is used both in building construction and for statues. It is also a popular choice for kitchen countertops. Peridotite is sometimes mined for peridot, a type of olivine that is used in jewelry.

Pumice is commonly used as an abrasive. Pumice is used to smooth skin or scrape up grime around the house. When pumice is placed into giant washing machines with newly manufactured jeans and tumbled, the result is “stone-washed” jeans. Ground up pumice stone is sometimes added to toothpaste to act as an abrasive material to scrub teeth.


Figure 2. Granite is an igneous rock used commonly in statues and building materials.

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- HS Rocks. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/4.0/. **License**: *CC BY-NC: Attribution-NonCommercial*

Public domain content

- Statue of Gudea. **Authored by**: Unknown. **Provided by**: Metropolitan Museum of Art. **Located at**: https://commons.wikimedia.org/wiki/File:Statue_of_Gudea_-_MET_-_59.2.jpg. **License**: *Public Domain: No Known Copyright*

This page titled 4.3.4: Uses of Igneous Rocks is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.3.5: Characteristics of Sedimentary Rocks



Figure 1. The White House of the USA is made of a sedimentary rock called sandstone.

The White House (shown in figure 1) is the official home and workplace of the President of the United States of America. Why do you think the White House is white? If you answered, “Because it is made of white rock,” you would be only partially correct. Construction for the White House began in 1792. Its outside walls are made of the sedimentary rock sandstone. This sandstone is very porous and is easily penetrated by rainwater. Water damage was common in the early days of construction for the building. To stop the water damage, workers covered the sandstone in a mixture of salt, rice, and glue, which help to give the White House its distinctive white color.

Lithification—Sediments to Sedimentary rocks

Sedimentary rocks are rocks made of lithified sediment. Sediments are grains of rocks, minerals, or mineraloids deposited on the surface of the earth. Reflect on the rock cycle for an indication of the relationships between the rocks that erode to become sediments and sedimentary rocks. For sediment to become sedimentary rock, it usually undergoes burial, compaction, and cementation.

Clastic sedimentary rocks are the result of weathering and erosion of source rocks, which turns them into pieces—clasts—of rocks and minerals. Once they become pieces, these clasts are free to move away from their source rock and they usually do. They are most often transported by water and deposited as layers of sediment.

The burial stage of lithification involves the deposition of more sediment layers top of those that had been deposited earlier. In a sedimentary basin where sediment is being deposited, it is common for subsidence (lowering) of the basin to be taking place, either because the crust and lithosphere beneath it are subsiding into the mantle to some extent, or because the surrounding uplands are undergoing uplift relative to the basin, or both. This allows thousands of feet of burial, in some cases tens of thousands of feet of burial, to occur.

As sediments are buried, the weight of overlying material exerts pressure, causing compaction of the sediments. The pressure, known as lithostatic pressure, “squeezes” the sediments from all sides into a smaller volume. Lithostatic pressure packs the sediment grains closer together and reduces the porosity – space between the sediment grains.

Some chemical sedimentary rocks are rock as soon as the sediments have been deposited by crystallization of minerals from substances dissolved in water, at the earth’s surface. Examples include rock salt and other evaporite deposits. These sediments of salt crystals and other minerals form sedimentary rock without having to undergo burial and compaction.

During burial and compaction, sediments will undergo some amount of cementation. Cementation refers to the growth of new minerals between the sediment grains. These new minerals bind the sediment grains together. One form of cementation is growth

of quartz rims on the surfaces of pre-existing quartz grains in the sediment. This new mineral growth is a result of water in the pore spaces that dissolves and precipiates quartz. A second common cementing mineral is hematite, a red or rust-colored iron oxide mineral, which precipitates onto the sediment grains from a combination of dissolved iron and oxygen from water in the pore spaces. A third common cementing mineral is calcite, which also precipitates from ions dissolved in the water in the pore spaces during lithification. Although there are other cementing minerals, quartz, hematite, and calcite are common cementing minerals that grow between or on the surfaces of the original sedimentary grains.

Generalized steps from source to sedimentary rock:
weathering >> erosion >> transport >> deposition >> burial >> compaction >> cementation >> sedimentary rock

Minerals and Sedimentary Rocks

Any type of rock containing any type of mineral will undergo weathering and erosion at the earth's surface. However, some minerals are more stable than others in earth's surface environments and are more likely to be found in sedimentary rocks.

Grains of clastic sediment, which are called clasts are winnowed and modified during the weathering-to-deposition process. Weathering of minerals will gradually eliminate the physically weaker and chemically more reactive minerals, increasing the relative abundance of more resistant minerals. Quartz tends to become increasingly abundant during the process, due to its common occurrence in the source rocks combined with its hardness and lack of cleavage, which makes it resistant to breaking down physically. Quartz is not easily dissolved or chemically altered, so it is resistant to breakdown by chemical reactions as well. That is why beach sand is often more rich in quartz than any other mineral. Sediments rich in quartz are considered mineralogically "mature" because they have been subjected to a longer interval of physical and chemical modification during the erosion-to-deposition process. Mature sediments are deposited farther from their source rocks in space and/or time than immature sediments.

Feldspars are the most common type of mineral in earth's crust and are also abundant in many clastic sediments and sedimentary rocks. Although feldspar is a fairly hard mineral, it does cleave (split apart) and is chemically reactive, especially in the presence of water. The most abundant product of chemical reaction of feldspar and water is clay minerals. During the erosion-to-deposition process, clastic sediments lose feldspar and gain a larger proportion of clay. Other minerals such as amphiboles, micas, and carbonates are relatively soft and chemically reactive and tend to be scarce or absent as sediment grains in mature clastic sediments, although calcite may be present in clastic sedimentary rocks as a secondary, cementing mineral that grew during lithification.

Minerals in chemical sedimentary rocks precipitate from water and usually remain in place or are not transported far before lithification. Such mineral sediments are subjected to little, if any, erosion and transportation. Therefore, the minerals in chemical sedimentary rocks are not winnowed during the weathering-to-deposition process as are the minerals in clastic sedimentary rocks. In some cases, during the formation of chemical sediments, the minerals may change as a result of chemical reactions. For example, dolostone is a chemical sedimentary rock that forms in certain coastal environments by alteration of precipitated calcite to dolomite.

Besides minerals, mineraloid solids occur in some chemical sedimentary rocks. For example, the carbonaceous material in coal is an organic mineraloid rather than a mineral. Another example, opal, is a chemical sedimentary rock that does not have a fully developed crystal lattice and therefore is a mineraloid.

Sediments



Figure 2. Water erodes the land surface in Alaska's Valley of Ten Thousand Smokes.

Sandstone is one of the common types of sedimentary rocks that form from sediments. There are many other types. Sediments may include:

- fragments of other rocks that often have been worn down into small pieces, such as sand, silt, or clay.
- **organic** materials, or the remains of once-living organisms.
- chemical precipitates, which are materials that get left behind after the water evaporates from a solution.

Rocks at the surface undergo mechanical and chemical weathering. These physical and chemical processes break rock into smaller pieces. Physical weathering simply breaks the rocks apart. Chemical weathering dissolves the less stable minerals. These original elements of the minerals end up in solution and new minerals may form.

Sediments are removed and transported by water, wind, ice, or gravity in a process called erosion (figure 2).



Figure 3. A river dumps sediments along its bed and on its banks.

Streams carry huge amounts of sediment (figure 3). The more energy the water has, the larger the particle it can carry. A rushing river on a steep slope might be able to carry boulders. As this stream slows down, it no longer has the energy to carry large sediments and will drop them. A slower moving stream will only carry smaller particles.

Sediments are deposited on beaches and deserts, at the bottom of oceans, and in lakes, ponds, rivers, marshes, and swamps. Avalanches drop large piles of sediment. Glaciers leave large piles of sediments, too. Wind can only transport sand and smaller particles. The type of sediment that is deposited will determine the type of sedimentary rock that can form. Different colors of sedimentary rock are determined by the environment where they are deposited. Red rocks form where oxygen is present. Darker sediments form when the environment is oxygen poor.

Sedimentary Rock Formation



Figure 4. This cliff is made of sandstone. Sands were deposited and then lithified.

Accumulated sediments harden into rock by **lithification**, as illustrated in figure 4. Two important steps are needed for sediments to lithify.

1. Sediments are squeezed together by the weight of overlying sediments on top of them. This is called **compaction**. Cemented, non-organic sediments become **clastic** rocks. If organic material is included, they are **bioclastic** rocks.
2. Fluids fill in the spaces between the loose particles of sediment and crystallize to create a rock by **cementation.**

The sediment size in clastic sedimentary rocks varies greatly (see table 1).

Table 1. Sedimentary rock sizes and features

Rock	Sediment Size	Other Features
Conglomerate	Large	Rounded
Breccia	Large	Angular
Sandstone	Sand-sized	
Siltstone	Silt-sized, smaller than sand	
Shale	Clay-sized, smallest	



Figure 5. The evaporite, halite, on a cobble from the Dead Sea, Israel.

When sediments settle out of calmer water, they form horizontal layers. One layer is deposited first, and another layer is deposited on top of it. So each layer is younger than the layer beneath it. When the sediments harden, the layers are preserved. Sedimentary rocks formed by the crystallization of chemical precipitates are called **chemical sedimentary rocks**. As discussed in the "Earth's Minerals" chapter, dissolved ions in fluids precipitate out of the fluid and settle out, just like the halite in figure 5.

Biochemical sedimentary rocks form in the ocean or a salt lake. Living creatures remove ions, such as calcium, magnesium, and potassium, from the water to make shells or soft tissue. When the organism dies, it sinks to the ocean floor to become a biochemical sediment, which may then become compacted and cemented into solid rock (figure 6).



Figure 6. Fossils in a biochemical rock, limestone, in the Carmel Formation in Utah.

Sedimentary Rock Textures

In clastic sediments the sedimentary texture includes the grain size, rounding, and sorting of the grains, all of which are related to what happened to the sediment during the weathering-to-deposition process.Because the processes that lead to the formation of chemical sedimentary rocks do not involve the weather-to-deposition process, there is no widely agreed-upon texture scheme that applies to chemical sedimentary rocks.

Clastic Textures

Clastic sedimentary textures are described in terms of the size of the sediment grains, how round they are, and how well they are sorted.

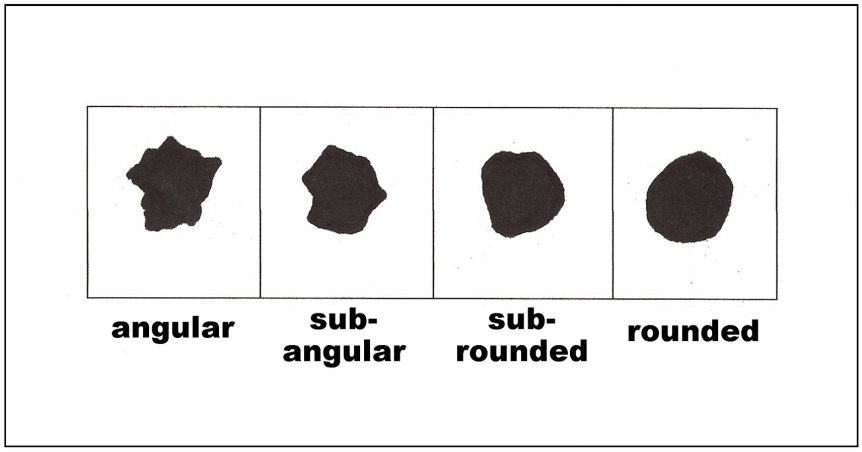
Grain Characteristics

The diameter or width of a clastic sediment grain determines its grain size. Specific ranges of grain size have specific names.

- Gravel is an overall name for large sediment grain size, which includes boulder, cobble, and pebble.
- Sand includes sediment grains ranging in size from 2mm to 0.625 mm.
- Silt is the name of a sediment grain that range in size from 0.625 mm to 0.0039 mm.
- Mud is the smallest grain size and is also known as clay. It is important to distinguish the grain size called clay from the mineral called clay. Clay sized grains are too small to see individually without the aid of a microscope.

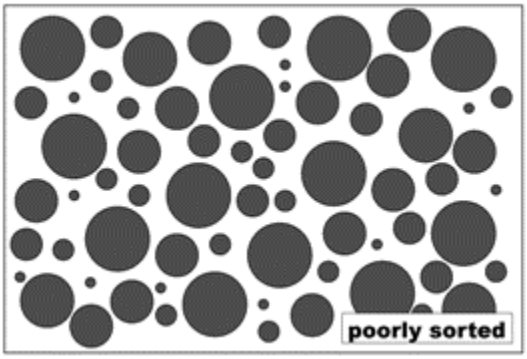
Rounding

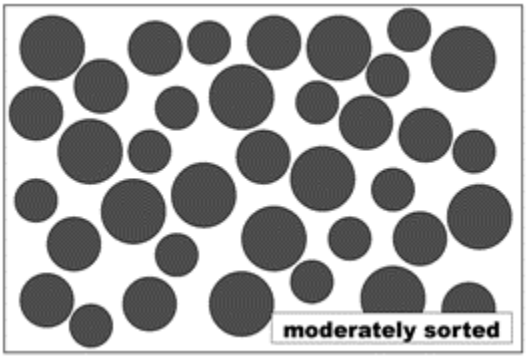
Clastic sediment grains can be round, angular, or in-between (subangular or subrounded). Breccia is a clastic sedimentary rock that by definition consists largely of angular grains of pebble size or larger. Conglomerate, another sedimentary rock, consists largely of rounded grains of pebble size or larger. The schematic diagram below shows classes of rounding, from the left: angular, subangular, subrounded, rounded. Not shown are very angular and well-rounded, which are less common.

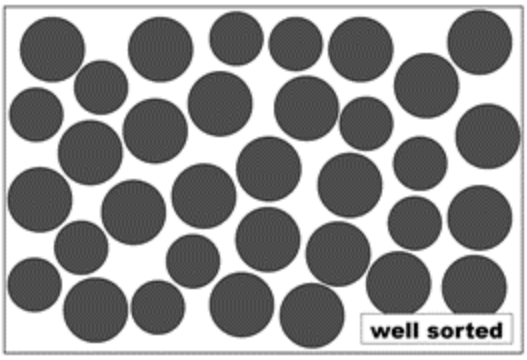


Sorting

The extent to which all the grains are the same size is known as sorting. If all the grains are the same size, they are well sorted. Some sandstones are well sorted, and some are not. Most conglomerates are poorly sorted, and consist of a mixture of grain sizes ranging from sand to pebble. The schematic diagrams below represent poorly-sorted, moderately-sorted, and well-sorted sediments.







Other Aspects of Texture

Other aspects of clastic sedimentary texture include the packing of the grains, the porosity of the rock, and the hardness of the rock as a whole.

The packing of the grains applies only to poorly sorted sediments in which the finer clasts form a matrix around the coarser clasts. If the large grains are touching each other, the packing is said to be clast-supported. If the coarse grains are separate and not touching each other, with the finer-grained sediment between them, the packing is said to be matrix-supported.

The porosity of a rock or sediment is the amount of empty space between the grains of sediment.

The hardness of the rock refers to how easily it breaks apart. Shale is harder than claystone, for example.

How to Identify Sedimentary Rocks

Clastic Textures

The common names of clastic sedimentary rocks—conglomerate, sandstone, siltstone, and shale—are based solely on grain size. However, there are more precise names within these broader categories that are based on clastic sediment features other than grain size. Breccia is a clastic sedimentary rock distinguished from conglomerate by the angularity of its clasts.

A clastic sedimentary rock simply made of sand-sized grains qualifies as sandstone, regardless of what minerals the grains are made of. More complete names for various sandstones are based on their mineral composition. Some sandstones are made almost entirely of quartz. Sandstone made virtually entirely of quartz is called arenite. Other sandstones have lots of feldspar mixed with quartz. Such sandstones are called arkose. Other sandstones are a mixture of feldspar, quartz, clay, and small fragments of “lithics,” dark-appearing minerals and rock fragments, which represent mineralogically immature sediments. Such sandstones are technically known as lithic wackes, although geologists often call them by their old name, graywackes, and the rocks are sometimes informally described as “dirty sandstones.”

In using the sedimentary rock classification table that accompanies this section, you will see that the clastic sedimentary rocks are classified on the basis of grain size. Sandstones are rocks made of sand-size grains. Shale is layered sedimentary rock made of fine mud-size grains too small to see with the naked eye.

Chemical Textures

The key to naming chemical sedimentary rocks is the minerals from which they are made. For example, all limestones consist mostly of the mineral calcite. Rock salt is made of salt minerals such as halite. Gypsum rock is made of the mineral gypsum. Chalcedony is made of microcrystalline quartz, quartz grains so tiny that they cannot be distinguished even with a standard optical microscope.

Table 2 shows some common types of sedimentary rocks.

Biochemical Textures

Biochemical sedimentary rocks form with the help of past life. This can be in the form of fossils, shells or plant remains. For example, coal is made up of plant remains.

Table 2. Common Sedimentary Rocks

Picture	Rock Name	Type of Sedimentary Rock
	Conglomerate	Clastic (fragments of non-organic sediments)
	Breccia	Clastic

	Sandstone	Clastic
	Siltstone	Clastic

	Shale	Clastic
	Rock Salt	Chemical precipitate

	Gypsum	Chemical precipitate
	Dolostone	Chemical precipitate

	Limestone	Biochemical
	Coal	Biochemical

Sedimentary Rock Classification

Clastic Sedimentary Rocks

Grain Size	Rounding	Sorting	Rock Name
gravel coarse (> 2 mm)	angular	poorly	breccia
	rounded	poorly	conglomerate
sand medium (0.0625–2 mm)	variable	variable	sandstone
	rounded to subangular	well to moderately	quartz arenite— >90% quartz
	subrounded to angular	well to moderately	arkose— >25% feldspar
	subangular to angular	moderately to poorly	graywacke—high % lithics, clay matrix

silt fine (0.004–0.0625 mm)	variable	well	siltstone
clay extra fine (< 0.004 mm)	not visible	well	shale—breaks along bedding planes
			mudstone—lacks bedding planes
			claystone—pure mineral clay, very soft
Mineral	**Distinguishing Characteristics**	**Rock Name**	**Common Depositional Environment**
calcite	softer than glass, reacts to HCl	limestone	warm shallow ocean
calcite	visible fossils, softer than glass, reacts to HCl	fossiliferous limestone (biochemical)	warm shallow ocean
dolomite	white to pink, softer than glass, reacts to HCl only when powdered	dolostone	saline lagoon
calcite	gray or brown, porous, reacts to HCl	travertine	evaporated spring or cave water
calcite	white, powdery, reacts to HCl	chalk	warm shallow ocean
microscopic quartz	conchoidal fracture	chert	deep ocean floor
microscopic quartz	red, conchoidal fracture	jasper	deep ocean floor
microscopic quartz	black, conchoidal fracture	flint	nodules in limestone
microscopic quartz	translucent, layered, conchoidal fracture	agate	nodules in volcanic rocks
microscopic quartz	white, powdery	diatomite	lakes in volcanic environments
carbon	black, soft, low density	coal (biochemical)	heavily vegetated swamp
halite	tastes like salt, cubic crystals and cleavage	rock salt	evaporating water body
gypsum	softer than fingernail	gypsum rock	evaporating water body

Uses of Sedimentary Rocks

Sedimentary rocks are used as building stones, although they are not as hard as igneous or metamorphic rocks. Sedimentary rocks are used in construction. Sand and gravel are used to make concrete; they are also used in asphalt. Many economically valuable resources come from sedimentary rocks. Iron ore and aluminum are two examples.

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: LumenLearning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- HS Rocks. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/4.0/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Basics -- Sedimentary Rocks. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/sedimentary.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.3.5: Characteristics of Sedimentary Rocks is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.3.6: Characteristics of Metamorphic Rocks

A metamorphic rock used to be some other type of rock, but it was changed inside the Earth to become a new type of rock. The word metamorphism comes from ancient Greek words for “change” (meta) and “form” (morph). The type of rock that a metamorphic rock used to be, prior to metamorphism, is called the protolith. During metamorphism the mineral content and texture of the protolith are changed due to changes in the physical and chemical environment of the rock. Metamorphism can be caused by burial, tectonic stress, heating by magma, or alteration by fluids. At advanced stages of metamorphism, it is common for a metamorphic rock to develop such a different set of minerals and such a thoroughly changed texture that it is difficult to recognize what the protolith was.

A rock undergoing metamorphism remains a solid rock during the process. Rocks do not melt during most conditions of metamorphism. At the highest grade of metamorphism, rocks begin to partially melt, at which point the boundary of metamorphic conditions is surpassed and the igneous part of the rock cycle is entered.

Even though rocks remain solid during metamorphism, fluid is generally present in the microscopic spaces between the minerals. This fluid phase may play a major role in the chemical reactions that are an important part of how metamorphism occurs. The fluid usually consists largely of water.

Metamorphic rocks provide a record of the processes that occurred inside Earth as the rock was subjected to changing physical and chemical conditions. This gives the geologist literally “inside information” on what occurs within the Earth during such processes as the formation of new mountain ranges, the collision of continents, the subduction of oceanic plates, and the circulation of sea water into hot oceanic crust. Metamorphic rocks are like probes that have gone down into the Earth and come back, bringing an record of the conditions they encountered on their journey in the depths of the Earth.



Figure 1. The platy layers in this large outcrop of metamorphic rock show the effects of pressure on rocks during metamorphism.

In the large outcrop of metamorphic rocks in figure 1, the rocks’ platy appearance is a result of the process metamorphism. Metamorphism is the addition of heat and/or pressure to existing rocks, which causes them to change physically and/or chemically so that they become a new rock. Metamorphic rocks may change so much that they may not resemble the original rock.

Metamorphism

Any type of rock—igneous, sedimentary, or metamorphic—can become a metamorphic rock. All that is needed is enough heat and/or pressure to alter the existing rock’s physical or chemical makeup without melting the rock entirely.


Figure 2. A foliated metamorphic rock.

Rocks change during metamorphism because the minerals need to be stable under the new temperature and pressure conditions. The need for stability may cause the structure of minerals to rearrange and form new minerals. Ions may move between minerals to create minerals of different chemical composition. Hornfels, with its alternating bands of dark and light crystals, is a good example of how minerals rearrange themselves during metamorphism. Hornfels is shown in table 1.

Extreme pressure may also lead to **foliation**, the flat layers that form in rocks as the rocks are squeezed by pressure (figure 2). Foliation normally forms when pressure is exerted in only one direction. Metamorphic rocks may also be non-foliated. Quartzite and limestone, shown in table 6, are nonfoliated.

The two main types of metamorphism are both related to heat within Earth:

1. **Regional metamorphism:** Changes in enormous quantities of rock over a wide area caused by the extreme pressure from overlying rock or from compression caused by geologic processes. Deep burial exposes the rock to high temperatures.
2. **Contact metamorphism:** Changes in a rock that is in contact with magma because of the magma's extreme heat.

Factors that Control Metamorphism

The reason rocks undergo metamorphism is that the minerals in a rock are only stable under a limited range of pressure, temperature, and chemical conditions. When rocks are subjected to large enough changes in these factors, the minerals will undergo chemical reactions that result in their replacement by new minerals, minerals that are stable in the new conditions.

Chemical Composition of the Protolith

The type of rock undergoes metamorphism is a major factor in determing what type of metamorphic rock it becomes. In short the identify of the protolith plays a big role the identity of the metamorphic rock. A fluid phase may introduce or remove chemical substances into or out of the rock during metamorphism, but in most metamorphic rock, most of the atoms in the protolith are be present in the metamorphic rock after metamorphism; the atoms will likely be rearranged into new mineral forms within the rock. Therefore, not only does the protolith determine the initial chemistry of the metamorphic rock, most metamorphic rocks do not change their bulk (overall) chemical compositions very much during metamorphism. The fact that most metamorphic rocks retain most of their original atoms means that even if the rock was so thoroughly metamorphosed that it no longer looks at all like the protolith, the rock can be analyzed in terms of its bulk chemical composition to determine what type of rock the protolith was.

Temperature

Temperature is another major factor of metamorphism. There are two ways to think about how the temperature of a rock can be increased as a result of geologic processes.

If rocks are buried within the Earth, the deeper they go, the higher the temperatures they experience. This is because temperature inside the Earth increases along what is called the geothermal gradient, or geotherm for short. Therefore, if rocks are simply buried deep enough enough sediment, they will experience temperatures high enough to cause metamorphism. This temperature is about 200°C (approximately 400°F).

Tectonic processes are another way rocks can be moved deeper along the geotherm. Faulting and folding the rocks of the crust, can move rocks to much greater depth than simple burial can.

Yet another way a rock in the Earth's crust can have its temperature greatly increased is by the intrusion of magma nearby. Magma intrusion subjects nearby rock to higher temperature with no increase in depth or pressure.

Pressure

Pressure is a measure of the stress, the physical force, being applied to the surface of a material. It is defined as the force per unit area acting on the surface, in a direction perpendicular to the surface.

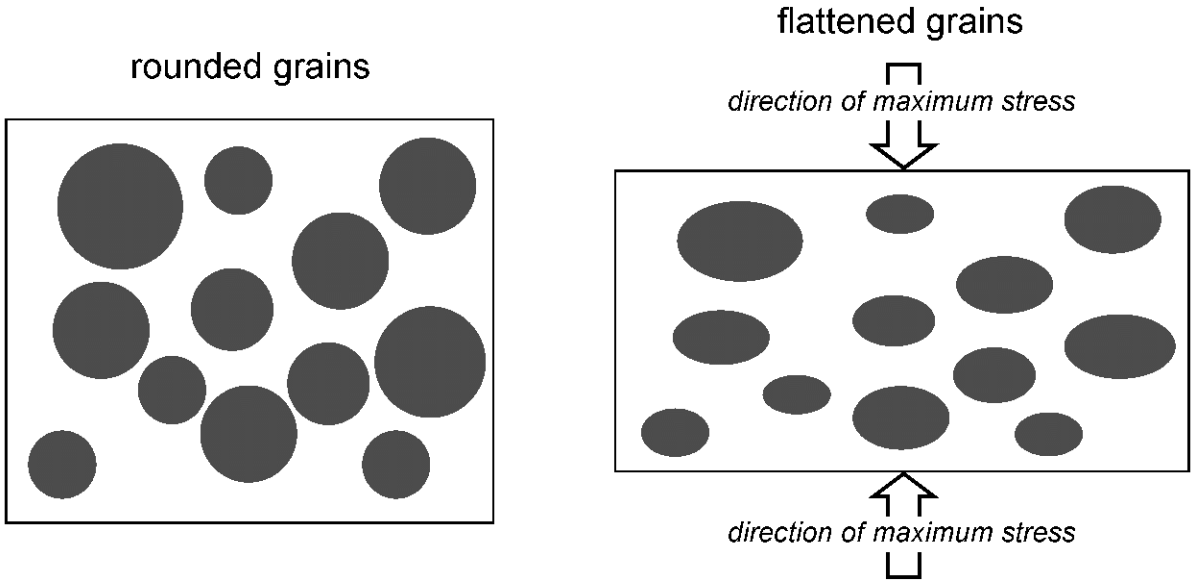
Lithostatic pressure is the pressure exerted on a rock by all the surrounding rock. The source of the pressure is the weight of all the rocks above. Lithostatic pressure increases as depth within the Earth increases and is a *uniform stress*—the pressure applies equally in all directions on the rock.

If pressure does not apply equally in all directions, differential stress occurs. There are two types of *differential stress*.

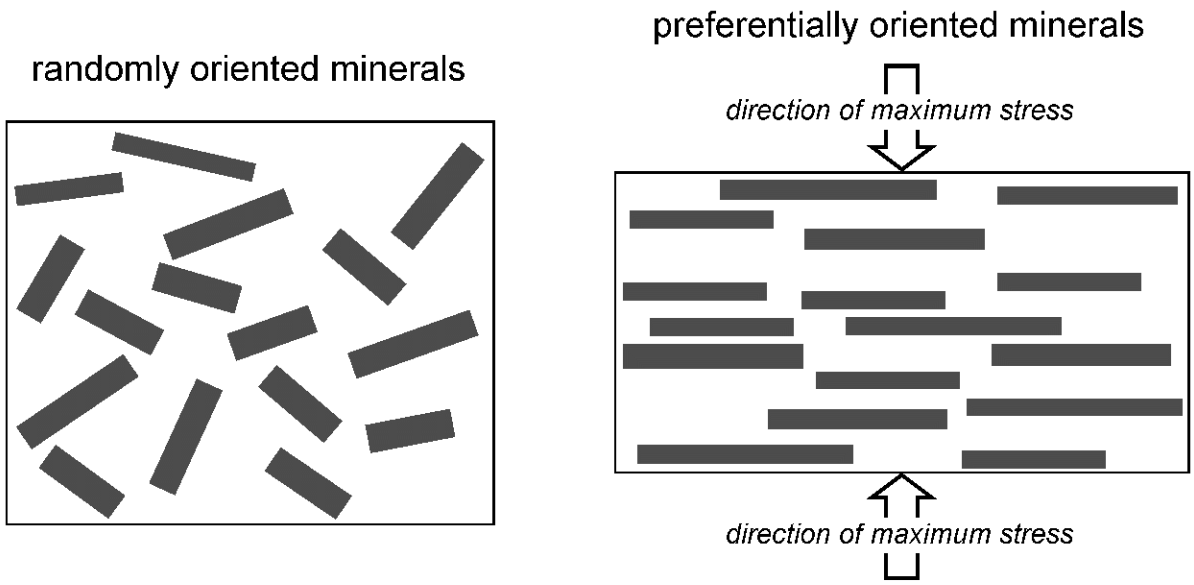
Normal stress compresses (pushes together) rock in one direction, the direction of maximum stress. At the same time, in a perpendicular direction, the rock undergoes tension (stretching), in the direction of minimum stress.

Shear stress pushes one side of the rock in a direction parallel to the side, while at the same time, the other side of the rock is being pushed in the opposite direction.

Differential stress has a major influence on the the appearance of a metamorphic rock. Differential stress can flatten pre-existing grains in the rock, as shown in the diagram below.



Metamorphic minerals that grow under differential stress will have a preferred orientation if the minerals have atomic structures that tend to make them form either flat or elongate crystals. This will be especially apparent for micas or other sheet silicates that grow during metamorphism, such as biotite, muscovite, chlorite, talc, or serpentine. If any of these flat minerals are growing under normal stress, they will grow with their sheets oriented perpendicular to the direction of maximum compression. This results in a rock that can be easily broken along the parallel mineral sheets. Such a rock is said to be foliated, or to have foliation.



Fluids

Any open space between the mineral grains in a rock, however microscopic, may contain a fluid phase. Most commonly, if there is a fluid phase in a rock during metamorphism, it will be a hydrous fluid, consisting of water and things dissolved in the water. Less commonly, it may be a carbon dioxide fluid or some other fluid. The presence of a fluid phase is a major factor during metamorphism because it helps determine which metamorphic reactions will occur and how fast they will occur. The fluid phase can also influence the rate at which mineral crystals deform or change shape. Most of this influence is due to the dissolved ions that pass in and out of the fluid phase. If during metamorphism enough ions are introduced to or removed from the rock via the fluid to change the bulk chemical composition of the rock, the rock is said to have undergone metasomatism. However, most metamorphic rocks do not undergo sufficient change in their bulk chemistry to be considered metasomatic rocks.

Time

Most metamorphism of rocks takes place slowly inside the Earth. Regional metamorphism takes place on a timescale of millions of years. Metamorphism usually involves slow changes to rocks in the solid state, as atoms or ions diffuse out of unstable minerals that are breaking down in the given pressure and temperature conditions and migrate into new minerals that are stable in those conditions. This type of chemical reaction takes a long time.

Grades of Metamorphism

Metamorphic grade refers to the general temperature and pressure conditions that prevailed during metamorphism. As the pressure and temperature increase, rocks undergo metamorphism at higher metamorphic grade. Rocks changing from one type of metamorphic rock to another as they encounter higher grades of metamorphism are said to be undergoing prograde metamorphism.

Low-grade metamorphism takes place at approximately 200–320 °C and relatively low pressure. This is not far beyond the conditions in which sediments get lithified into sedimentary rocks, and it is common for a low-grade metamorphic rock to look somewhat like its protolith. Low grade metamorphic rocks tend to characterized by an abundance of hydrous minerals, minerals that contain water within their crystal structure. Examples of low grade hydrous minerals include clay, serpentine, and chlorite. Under low grade metamorphism many of the metamorphic minerals will not grow large enough to be seen without a microscope.

Medium-grade metamorphism takes place at approximately at 320–450 °C and at moderate pressures. Low grade hydrous minerals are replaced by micas such as biotite and muscovite, and non-hydrous minerals such as garnet may grow. Garnet is an example of a mineral which may form porphyroblasts, metamorphic mineral grains that are larger in size and more equant in shape (about the same diameter in all directions), thus standing out among the smaller, flatter, or more elongate minerals.

High-grade metamorphism takes place at temperatures above about 450 °C. Micas tend to break down. New minerals such as hornblende will form, which is stable at higher temperatures. However, as metamorphic grade increases to even higher grade, all hydrous minerals, which includes hornblende, may break down and be replaced by other, higher-temperature, non-hydrous minerals such as pyroxene.

Index Minerals

Index minerals, which are indicators of metamorphic grade. In a given rock type, which starts with a particular chemical composition, lower-grade index minerals are replaced by higher-grade index minerals in a sequence of chemical reactions that proceeds as the rock undergoes prograde metamorphism. For example, in rocks made of metamorphosed shale, metamorphism may prograde through the following index minerals:

- chlorite characterizes the lowest regional metamorphic grade
- biotite replaces chlorite at the next metamorphic grade, which could be considered medium-low grade
- garnet appears at the next metamorphic grade, medium grade
- staurolite marks the next metamorphic grade, which is medium-high grade
- sillimanite is a characteristic mineral of high grade metamorphic rocks

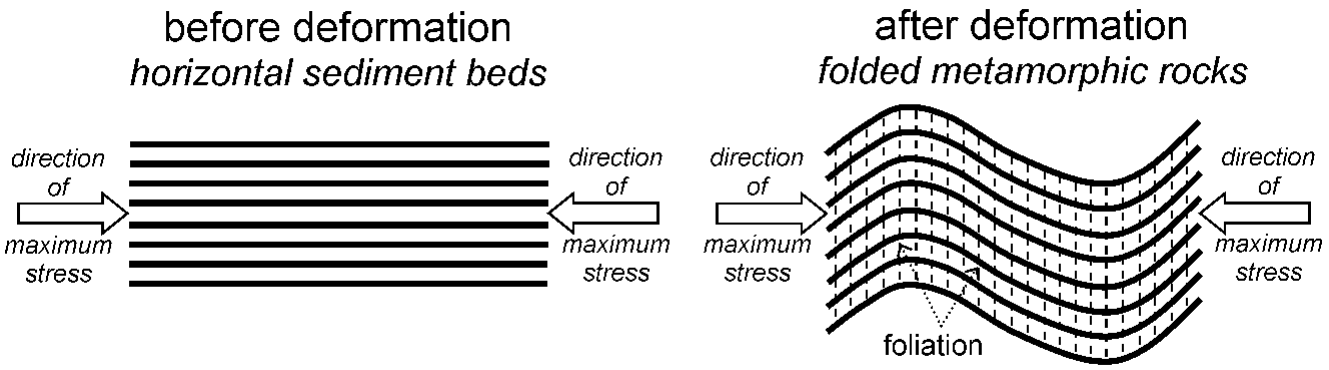
Index minerals are used by geologists to map metamorphic grade in regions of metamorphic rock. A geologist maps and collects rock samples across the region and marks the geologic map with the location of each rock sample and the type of index mineral it contains. By drawing lines around the areas where each type of index mineral occurs, the geologist delineates the zones of different metamorphic grades in the region. The lines are known as isograds.

Types of Metamorphism

Regional Metamorphism

Regional metamorphism occurs where large areas of rock are subjected to large amounts of differential stress for long intervals of time, conditions typically associated with mountain building. Mountain building occurs at subduction zones and at continental collision zones where two plates each bearing continental crust, converge upon each other.

Most foliated metamorphic rocks—slate, phyllite, schist, and gneiss—are formed during regional metamorphism. As the rocks become heated at depth in the Earth during regional metamorphism they become ductile, which means they are relatively soft even though they are still solid. The folding and deformation of the rock while it is ductile may greatly distort the original shapes and orientations of the rock, producing folded layers and mineral veins that have highly deformed or even convoluted shapes. The diagram below shows folds forming during an early stage of regional metamorphism, along with development of foliation, in response to normal stress.



The photograph below shows high-grade metamorphic rock that has undergone several stages of foliation development and folding during regional metamorphism, and may even have reached such a high temperature that it began to melt.



Contact Metamorphism

Contact metamorphism occurs to solid rock next to an igneous intrusion and is caused by the heat from the nearby body of magma. Because contact metamorphism is not caused by changes in pressure or by differential stress, contact metamorphic rocks do not become foliated. Where intrusions of magma occur at shallow levels of the crust, the zone of contact metamorphism around the intrusion is relatively narrow, sometimes only a few m (a few feet) thick, ranging up to contact metamorphic zones over 1000 m (over 3000 feet) across around larger intrusions that released more heat into the adjacent crust. The zone of contact metamorphism surrounding an igneous intrusion is called the metamorphic aureole. The rocks closest to the contact with the intrusion are heated to the highest temperatures, so the metamorphic grade is highest there and diminishes with increasing distance away from the contact. Because contact metamorphism occurs at shallow to moderate depths in the crust and subjects the rocks to temperatures up to the verge of igneous conditions, it is sometimes referred to as high-temperature, low-pressure metamorphism. Hornfels, which is a hard metamorphic rock formed from fine-grained clastic sedimentary rocks, is a common product of contact metamorphism.

Hydrothermal Metamorphism

Hydrothermal metamorphism is the result of extensive interaction of rock with high-temperature fluids. The difference in composition between the existing rock and the invading fluid drives the chemical reactions. The hydrothermal fluid may originate from a magma that intruded nearby and caused fluid to circulate in the nearby crust, from circulating hot groundwater, or from ocean water. If the fluid introduces substantal amounts of ions into the rock and removes substantial amounts of ions from it, the fluid has metasomatized the rock—changed its chemical composition.

Ocean water that penetrates hot, cracked oceanic crust and circulates as hydrothermal fluid in ocean floor basalts produces extensive hydrothermal metamorphism adjacent to mid-ocean spreading ridges and other ocean-floor volcanic zones. Much of the basalt subjected to this type of metamorphism turns into a type of metamorphic rock known as greenschist. Greenschist contains a set of minerals, some of them green, which may include chlorite, epidote, talc, Na-plagioclase, or actinolite. The fluids eventually escape through vents in the ocean floor known as black smokers, producing thick deposits of minerals on the ocean floor around the vents.

Burial Metamorphism

Burial metamorphism occurs to rocks buried beneath sediments to depths that exceed the conditions in which sedimentary rocks form. Because rocks undergoing burial metamorphism encounter the uniform stress of lithostatic pressure, not differential pressure, they do not develop foliation. Burial metamorphism is the lowest grade of metamorphism. The main type of mineral that usually grows during burial metamorphism is zeolite, a group of low-density silicate minerals. It usually requires a strong microscope see the small grains of zeolite minerals that form during burial metamorphism.

Subduction Zone Metamorphism

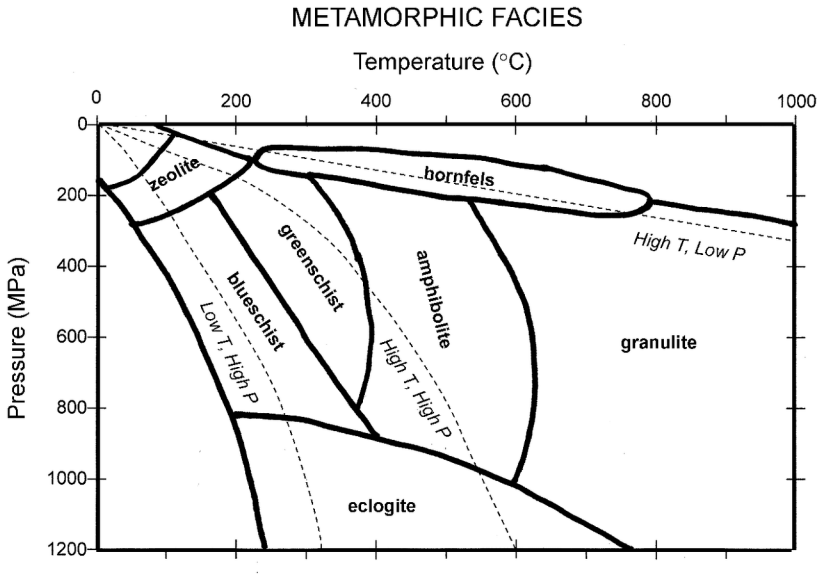
During subduction, a tectonic plate, consisting of oceanic crust and lithospheric mantle, is recycled back into the deeper mantle. In most subduction zones the subducting plate is relatively cold compared with the high temperature it had when first formed at a mid-ocean spreading ridge. Subduction takes the rocks to great depth in the Earth relatively quickly. This produces a characteristic type of metamorphism, sometimes called high-pressure, low-temperature (high-P, low-T) metamorphism, which only occurs deep in a subduction zone. In oceanic basalts that are part of a subducting plate, the high-P, low-T conditions create a distinctive set of metamorphic minerals including a type of amphibole, called glaucophane, that has a blue color. Blueschist is the name given to this type of metamorphic rock. Blueschist is generally interpreted as having been produced within a subduction zone, even if the plate boundaries have subsequently shifted and that location is no longer at a subduction zone.

Metamorphic Facies

Much as the minerals and textures of sedimentary rocks can be used as windows to see into the environment in which the sediments were deposited on the Earth's surface, the minerals and textures of metamorphic rocks provide windows through which we view the conditions of pressure, temperature, fluids, and stress that occurred inside the Earth during metamorphism. The pressure and temperature conditions under which specific types of metamorphic rocks form has been determined by a combination labratory experiments, physics-based theoretical calculations, along with evidence in the textures of the rocks and their field relations as recorded on geologic maps. The knowledge of temperatures and pressures at which particular types of metamorphic rocks form led to the concept of metamorphic facies. Each metamorphic facies is represented by a specific type of metamorphic rock that forms under a specific pressure and temperature conditions.

Even though the name of the each metamorphic facies is taken from a type of rock that forms under those conditions, that is not the only type of rock that will form in those conditions. For example, if the protolith is basalt, it will turn into greenschist under greenschist facies conditions, and that is what facies is named for. However, if the protolith is shale, a muscovite-biotite schist, which is not green, will form instead. If it can be determined that a muscovite-biotite schist formed at around 350°C temperature and 400 MPa pressure, it can be stated that the rock formed in the greenschist facies, even though the rock is not itself a greenschist.

The diagram below shows metamorphic facies in terms of pressure and temperature condiditons inside the Earth. Earth's surface conditions are near the top left corner of the graph at about 15°C which is the average temperature at Earth's surface and 0.1 MPa (megapascals), which is about the average atmospheric pressure on the Earth's surface. Just as atmospheric pressure comes from the weight of all the air above a point on the Earth's surface, pressure inside the Earth comes from the weight of all the rock above a given depth. Rocks are much denser than air and MPa is the unit most commonly uses to express pressures inside the Earth. One MPa equals nearly 10 atmospheres. A pressure of 1000 MPa corresponds to a depth of about 35 km inside the Earth. Although pressure inside the Earth is determined by the depth, temperature depends on more than depth. Temperature depends on the heat flow, which varies from location to location. The way temperature changes with depth inside the Earth is called the geothermal gradient, geotherm for short. In the diagram below, three different geotherms are marked with dashed lines. The three geotherms represent different geological settings in the Earth.



High-pressure, low-temperature geotherms occurs in subduction zones. As the diagram shows, rocks undergoing prograde metamorphism in subduction zones will be subjected to zeolite, blueschist, and ultimately eclogite facies conditions.

High-temperature, low-pressure geotherms occur in the vicinity of igneous intrusions in the shallow crust, underlying a volcanically active area. Rocks that have their pressure and temperature conditions increased along such a geotherm will metamorphose in the hornfels facies and, if it gets hot enough, in the granulite facies.

Blueschist facies and hornfels facies are associated with unusual geothermal gradients. The most common conditions in the Earth are found along geotherms between those two extremes. Most regional metamorphic rocks are formed in conditions within this range of geothermal gradients, passing through the greenschist facies to the amphibolites facies. At the maximum pressures and temperatures the rocks may encounter within the Earth in this range of geotherms, they will enter either the granulite or eclogite facies. Regionally metamorphosed rocks that contain hydrous fluids will begin to melt before they pass beyond the amphibolite facies.

Types of Metamorphic Rocks

Metamorphic rock fall into two categories, foliated and unfoliated. Most foliated metamorphic rocks originate from regional metamorphism. Some unfoliated metamorphic rocks, such as hornfels, originate only by contact metamorphism, but others can originate either by contact metamorphism or by regional metamorphism. Quartz and marble are prime examples of unfoliated that can be produced by either regional or contact metamorphism. Both rock types consist of metamorphic minerals that do not have flat or elongate shapes and thus cannot become layered even if they are produced under differential stress.

A geologist working with metamorphic rocks collects the rocks in the field and looks for the patterns the rocks form in outcrops as well as how those outcrops are related to other types of rock with which they are in contact. Field evidence is often required to know for sure whether rocks are products of regional metamorphism, contact metamorphism, or some other type of metamorphism. If only looking at rock samples in a laboratory, one can be sure of the type of metamorphism that produced a foliated metamorphic rock such as schist or gneiss, or a hornfels, which is unfoliated, but one cannot be sure of the type of metamorphism that produced an unfoliated marble or quartzite.

Foliated Metamorphic Rocks

Foliated metamorphic rocks are named for their style of foliation. However, a more complete name of each particular type of foliated metamorphic rock includes the main minerals that the rock comprises, such as biotite-garnet schist rather than just schist.

- slate—slates form at low metamorphic grade by the growth of fine-grained chlorite and clay minerals. The preferred orientation of these sheet silicates causes the rock to easily break along parallel planes, giving the rock a slaty cleavage.

Some slate breaks into such extensively flat sheets of rock that it is used as the base of pool tables, beneath a layer of rubber and felt. Roof tiles are also sometimes made of slate.

- phyllite—phyllite is a low-medium grade regional metamorphic rock in which the clay minerals and chlorite have been at least partly replaced by mica mica minerals, muscovite and biotite. This gives the surfaces of phyllite a satiny luster, much brighter than the surface of a piece of slate. It is also common for the differential stresses under which phyllite forms to have produced a set of folds in the rock, making the foliation surfaces wavy or irregular, in contrast to the often perfectly flat surfaces of slaty cleavage.
- schist—the size of mineral crystals tends to grow larger with increasing metamorphic grade. Schist is a product of medium grades of metamorphism and is characterized by visibly prominent, parallel sheets of mica or similar sheet silicates, usually either muscovite or biotite, or both. In schist, the sheets of mica are usually arranged in irregular planes rather than perfectly flat planes, giving the rock a schistose foliation (or simply schistosity). Schist often contains more than just micas among its minerals, such as quartz, feldspars, and garnet.
- amphibolite—a poorly foliated to unfoliated mafic metamorphic rock, usually consisting largely of the common black amphibole known as hornblende, plus plagioclase, plus or minus biotite and possibly other minerals; it usually does not contain any quartz. Amphibolite forms at medium-high metamorphic grades. Amphibolite is also listed below in the section on unfoliated metamorphic rocks.
- gneiss—like the word schist, the word gneiss is originated from the German language; it is pronounced “nice.” As metamorphic grade continue to increase, sheet silicates become unstable and dark minerals such as hornblende or pyroxene start to grow. The dark-colored minerals tend to form separate bands or stripes in the rock, giving it a gneissic foliation of dark and light streaks. Gneiss is a high-grade metamorphic rock. Many types of gneiss look somewhat like granite, except that the gneiss has dark and light stripes whereas in granite randomly oriented and distributed minerals with no stripes or layers.
- migmatite—a combination of high-grade regional metamorphic rock – usually gneiss or schist – and granitic igneous rock. The granitic rock in migmatite probably originated from partial melting of some of the metamorphic rock, though in some migmatites the granite may have intruded the rock from deeper in the crust. In migmatite you can see metamorphic rock that has reached the limits of metamorphism and begun transitioning into the igneous stage of the rock cycle by melting to form magma.

Names of different styles of foliation come from the common rocks that exhibit such foliation:

- slate has ***slaty*** foliation
- phyllite has ***phyllitic*** foliation
- schist has ***schistose*** foliation
- gneiss has ***gneissic*** foliation (also called gneissose foliation)

Nonfoliated Metamorphic Rocks

Nonfoliated metamorphic rocks lack a planar (oriented) fabric, either because the minerals did not grow under differential stress, or because the minerals that grew during metamorphism are not minerals that have elongate or flat shapes. Because they lack foliation, these rocks are named entirely on the basis of their mineralogy.

- hornfels—hornfels are very hard rocks formed by contact metamorphism of shale, siltstone, or sandstone. The heat from the nearby magma “bakes” the sedimentary rocks and recrystallizes the minerals in them into a new texture that no longer breaks easily along the original sedimentary bedding planes. Depending on the composition of the rock and the temperature reached, minerals indicative of high metamorphic grade such as pyroxene may occur in some hornfels, though many hornfels have minerals indicating medium grade metamorphism.
- amphibolite—amphibolites are dark-colored rocks with amphibole, usually the common black amphibole known as hornblende, as their most abundant mineral, along with plagioclase and possibly other minerals, though usually no quartz. Amphibolites are poorly foliated to unfoliated and form at medium to medium-high grades of metamorphism from basalt or gabbro.
- quartzite—quartzite is a metamorphic rock made almost entirely of quartz, for which the protolith was quartz arenite. Because quartz is stable over a wide range of pressure and temperature, little or no new minerals form in quartzite during metamorphism. Instead, the quartz grains recrystallize into a denser, harder rock than the original sandstone. If struck by a rock hammer, quartzite will commonly break right through the quartz grains, rather than around them as when quartz arenite is broken.
- marble—marble is a metamorphic rock made up almost entirely of either calcite or dolomite, for which the protolith was either limestone or dolostone, respectively. Marbles may have bands of different colors which were deformed into convoluted folds while the rock was ductile. Such marble is often used as decorative stone in buildings. Some marble, which is considered

better quality stone for carving into statues, lacks color bands.

Table 1. Common Metamorphic Rocks and Their Parent Rock

Picture	Rock Name	Type of Metamorphic Rock	Comment
	Slate	Foliated	Metamorp of shale
	Phyllite	Foliated	Metamorp of slate, b under gre heat and pressure t slate

Table 1. Common Metamorphic Rocks and Their Parent Rock

	Schist	Foliated	Often deri from metamorp of claysto shale; metamorp under mo heat and pressure t phyllite
	Gneiss	Foliated	Metamorp of various different r under extr conditions heat and pressure

Table 1. Common Metamorphic Rocks and Their Parent Rock			
	Hornfels	Non-foliated	Contact metamorp of various different r types
	Quartzite	Non-foliated	Metamorp of sandsto

Table 1. Common Metamorphic Rocks and Their Parent Rock			
	Marble	Non-foliated	Metamorp of limesto
	Metaconglomerate	Non-foliated	Metamorp of conglor

Metamorphic Rock Classification

Foliated Metamorphic Rocks				
Crystal Size	**Mineralogy**	**Protolith**	**Metamorphism**	**Rock Name**
very fine	clay minerals	shale	low grade regional	slate
fine	clay minerals, biotite, muscovite	shale	low grade regional	phyllite
medium to coarse	biotite, muscovite, quartz, garnet, plagioclase	shale, basalt	medium grade regional	schist
medium to coarse	amphibole, plagioclase, biotite	basalt	medium grade regional	amphibolite *(Note: may be unfoliated)*

medium to coarse	plagioclase, orthoclase, quartz, biotite, amphibole, pyroxene	basalt, granite, shale	high grade regional	gneiss
Unfoliated Metamorphic Rocks				
Crystal Size	**Mineralogy**	**Protolith**	**Metamorphism**	**Rock Name**
fine to coarse	quartz	sandstone	regional or contact	quartzite
fine to coarse	calcite	limestone	regional or contact	marble
fine	pyroxene, amphibole, plagioclase	shale	contact	hornfels

Note that not all minerals listed in the mineralogy column will be present in every rock of that type and that some rocks may have minerals not listed here.

Uses of Metamorphic Rocks



Figure 3. Marble is used for decorative items and in art.

Quartzite and marble are commonly used for building materials and artwork. Marble is beautiful for statues and decorative items such as vases (see an example in figure 3). Ground up marble is also a component of toothpaste, plastics, and paper.

Quartzite is very hard and is often crushed and used in building railroad tracks (see figure 4). Schist and slate are sometimes used as building and landscape materials. Graphite, the “lead” in pencils, is a mineral commonly found in metamorphic rocks.



Figure 4. Crushed quartzite is sometimes placed under railroad tracks because it is very hard and durable.

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Basics -- Metamorphic Rocks. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/metamorphic.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*
- HS Rocks. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/4.0/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.3.6: Characteristics of Metamorphic Rocks is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.4: WEATHERING AND SOIL FORMATION

Learning Objectives

Explain the formation of soils, different weathering processes and erosion.

- Describe types and processes of weathering and erosion.
- Describe and compare typical soil profiles and basic soil types
- Discuss the importance of soil conservation.

4.4.1: What Is Weathering?

4.4.2: Mechanical Weathering

4.4.3: Chemical Weathering

4.4.4: Influences on Weathering

4.4.5: Soil Texture and Composition

4.4.6: Soil Horizons and Profiles

4.4.7: Types of Soils

4.4.8: Causes of Soil Erosion

4.4.9: Avoiding Soil Loss

This page titled Chapter 4.4: Weathering and Soil Formation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.4.1: What Is Weathering?

The footprints that astronauts left on the Moon will be there forever. Why? This is because the Moon has no atmosphere and, as a result, has no weathering. Weathering is one of the forces on Earth that destroy rocks and landforms. Without weathering, geologic features would build up but would be less likely to break down.

What is Weathering?

Weathering is the process that changes solid rock into sediments. Sediments were described in the Rocks chapter. With weathering, rock is disintegrated. It breaks into pieces.

Once these sediments are separated from the rocks, erosion is the process that moves the sediments. Erosion is the next chapter's topic. The four forces of erosion are water, wind, glaciers, and gravity.

- Water is responsible for most erosion. Water can move most sizes of sediments, depending on the strength of the force.
- Wind moves sand-sized and smaller pieces of rock through the air.
- Glaciers move all sizes of sediments, from extremely large boulders to the tiniest fragments.
- Gravity moves broken pieces of rock, large or small, downslope.



Figure 1. A once smooth road surface has cracks and fractures, plus a large pothole.

While plate tectonics forces work to build huge mountains and other landscapes, the forces of weathering gradually wear those rocks and landscapes away. Together with erosion, tall mountains turn into hills and even plains. The Appalachian Mountains along the east coast of North America were once as tall as the Himalayas.

No human being can watch for millions of years as mountains are built, nor can anyone watch as those same mountains gradually are worn away. But imagine a new sidewalk or road. The new road is smooth and even. Over hundreds of years, it will completely disappear, but what happens over one year? What changes would you see (figure 1)? What forces of weathering wear down that road, or rocks or mountains over time?

Follow this link to view some animations of different types of weathering processes.

Contributors and Attributions

CC licensed content, Shared previously

- 9.1: Weathering. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/9.1/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.4.1: What Is Weathering? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.4.2: Mechanical Weathering

Mechanical weathering (also called physical weathering) breaks rock into smaller pieces. These smaller pieces are just like the bigger rock, just smaller. That means the rock has changed physically without changing its composition. The smaller pieces have the same minerals, in just the same proportions as the original rock.

There are many ways that rocks can be broken apart into smaller pieces. **Ice wedging** is the main form of mechanical weathering in any climate that regularly cycles above and below the freezing point (figure 2). Ice wedging works quickly, breaking apart rocks in areas with temperatures that cycle above and below freezing in the day and night, and also that cycle above and below freezing with the seasons.

Figure 2. Ice wedging.

Ice wedging breaks apart so much rock that large piles of broken rock are seen at the base of a hillside, as rock fragments separate and tumble down. Ice wedging is common in Earth's polar regions and mid latitudes, and also at higher elevations, such as in the mountains. **Abrasion** is another form of mechanical weathering. In abrasion, one rock bumps against another rock.



Figure 3. Rocks on a beach are worn down by abrasion as passing waves cause them to strike each other.

- Gravity causes abrasion as a rock tumbles down a mountainside or cliff.
- Moving water causes abrasion as particles in the water collide and bump against one another.
- Strong winds carrying pieces of sand can sandblast surfaces.
- Ice in glaciers carries many bits and pieces of rock. Rocks embedded at the bottom of the glacier scrape against the rocks below.

Abrasion makes rocks with sharp or jagged edges smooth and round. If you have ever collected beach glass or cobbles from a stream, you have witnessed the work of abrasion (figure 3).

Now that you know what mechanical weathering is, can you think of other ways it could happen? Plants and animals can do the work of mechanical weathering (figure 4). This could happen slowly as a plant's roots grow into a crack or fracture in rock and gradually grow larger, wedging open the crack. Burrowing animals can also break apart rock as they dig for food or to make living spaces for themselves.

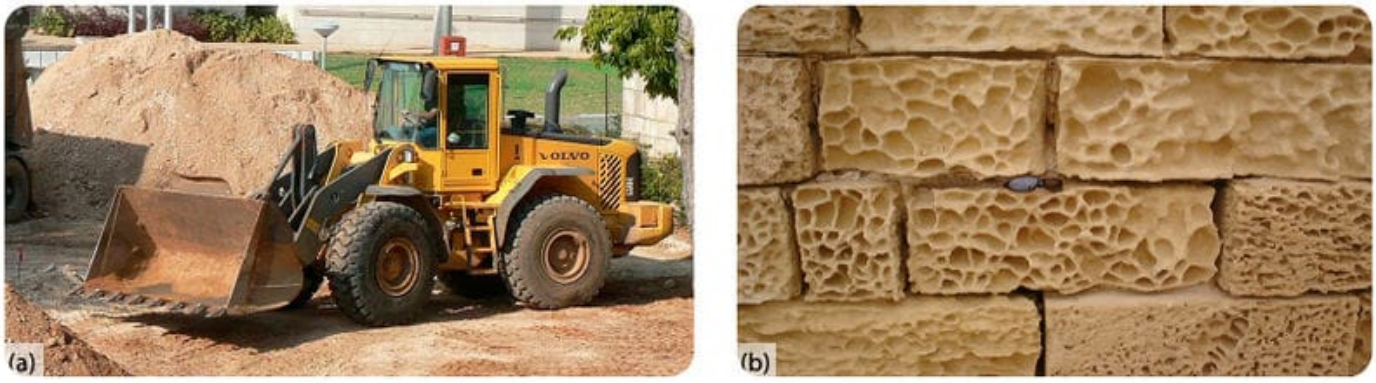


Figure 4. (a) Human activities are responsible for enormous amounts of mechanical weathering, by digging or blasting into rock to build homes, roads, subways, or to quarry stone. (b) Salt weathering of building stone on the island of Gozo, Malta.

Mechanical weathering increases the rate of chemical weathering. As rock breaks into smaller pieces, the surface area of the pieces increases figure 5. With more surfaces exposed, there are more surfaces on which chemical weathering can occur.

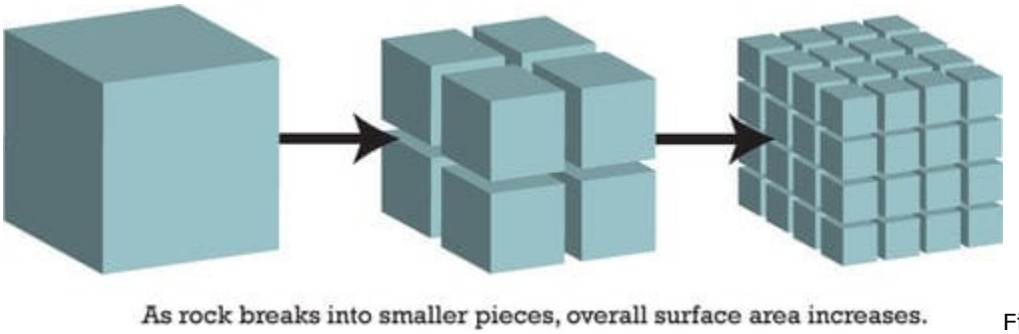


Figure 5. Mechanical weathering may increase the rate of chemical weathering.

Contributors and Attributions

CC licensed content, Shared previously

- 9.1: Weathering. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/9.1/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.4.2: Mechanical Weathering is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.4.3: Chemical Weathering

Chemical weathering is the other important type of weathering. Chemical weathering is different from mechanical weathering because the rock changes, not just in size of pieces, but in composition. That is, one type of mineral changes into a different mineral. Chemical weathering works through chemical reactions that cause changes in the minerals.



Figure 1. Deforestation in Brazil reveals the underlying clay-rich soil.

Most minerals form at high pressure or high temperatures deep in the crust, or sometimes in the mantle. When these rocks reach the Earth's surface, they are now at very low temperatures and pressures. This is a very different environment from the one in which they formed and the minerals are no longer stable. In chemical weathering, minerals that were stable inside the crust must change to minerals that are stable at Earth's surface.

Remember that the most common minerals in Earth's crust are the silicate minerals. Many silicate minerals form in igneous or metamorphic rocks. The minerals that form at the highest temperatures and pressures are the least stable at the surface. Clay is stable at the surface and chemical weathering converts many minerals to clay (figure 1).

There are many types of chemical weathering because there are many agents of chemical weathering. Water is the most important agent of chemical weathering. Two other important agents of chemical weathering are carbon dioxide and oxygen.

Chemical Weathering by Water

A water molecule has a very simple chemical formula, H_2O, two hydrogen atoms bonded to one oxygen atom. But water is pretty remarkable in terms of all the things it can do. Remember from the Earth's Minerals chapter that water is a polar molecule. The positive side of the molecule attracts negative ions and the negative side attracts positive ions. So water molecules separate the ions from their compounds and surround them. Water can completely dissolve some minerals, such as salt. Follow this link to check out this animation of how water dissolves salt.

Hydrolysis is the name of the chemical reaction between a chemical compound and water. When this reaction takes place, water dissolves ions from the mineral and carries them away. These elements have undergone **leaching**. Through hydrolysis, a mineral such as potassium feldspar is leached of potassium and changed into a clay mineral. Clay minerals are more stable at the Earth's surface.

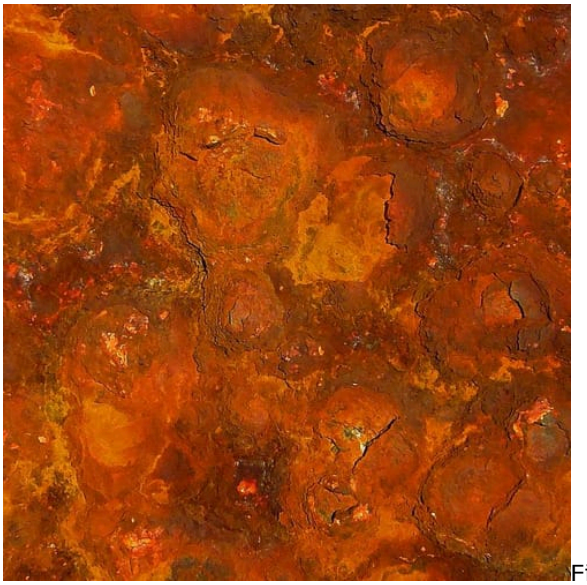

Figure 7. This statue has been damaged by acid rain.

Chemical Weathering by Acid Rain

Carbon dioxide (CO_2) combines with water as raindrops fall through the atmosphere. This makes a weak acid, called carbonic acid. Carbonic acid is a very common in nature where it works to dissolve rock.

Pollutants, such as sulfur and nitrogen, from fossil fuel burning, create sulfuric and nitric acid. Sulfuric and nitric acids are the two main components of acid rain, which accelerate chemical weathering (figure 7). Acid rain is discussed in the Human Actions and the Atmosphere chapter.

Chemical Weathering by Oxygen


Figure 8. When iron rich minerals oxidize, they produce the familiar red color found in rust.

Oxidation is a chemical reaction that takes place when oxygen reacts with another element. Oxygen is very strongly chemically reactive. The most familiar type of oxidation is when iron reacts with oxygen to create rust (figure 8). Minerals that are rich in iron break down as the iron oxidizes and forms new compounds. Iron oxide produces the red color in soils.

Now that you know what chemical weathering is, can you think of some other ways chemical weathering might occur? Chemical weathering can also be contributed to by plants and animals. As plant roots take in soluble ions as nutrients, certain elements are exchanged. Plant roots and bacterial decay use carbon dioxide in the process of respiration.

Contributors and Attributions

CC licensed content, Shared previously

- 9.1: Weathering. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/9.1/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.4.3: Chemical Weathering is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.4.4: Influences on Weathering

Weathering rates depend on several factors. These include the composition of the rock and the minerals it contains as well as the climate of a region.

Rock and Mineral Type



Figure 1. Devil's Tower is the central plug of resistant lava from which the surrounding rock weathered and eroded away.

Different rock types weather at different rates. Certain types of rock are very resistant to weathering. Igneous rocks, especially intrusive igneous rocks such as granite, weather slowly because it is hard for water to penetrate them. Other types of rock, such as limestone, are easily weathered because they dissolve in weak acids.

Rocks that resist weathering remain at the surface and form ridges or hills. Devil's Tower in Wyoming is an igneous rock from beneath a volcano (figure 1). As the surrounding less resistant rocks were worn away, the resistant center of the volcano remained behind.

Different minerals also weather at different rates. Some minerals in a rock might completely dissolve in water but the more resistant minerals remain. In this case, the rock's surface becomes pitted and rough. When a less resistant mineral dissolves, more resistant mineral grains are released from the rock.

Climate

A region's **climate** strongly influences weathering. Climate is determined by the temperature of a region plus the amount of precipitation it receives. Climate is weather averaged over a long period of time. Chemical weathering increases as:

- Temperature increases: Chemical reactions proceed more rapidly at higher temperatures. For each 10°C increase in average temperature, the rate of chemical reactions doubles.
- Precipitation increases: More water allows more chemical reactions. Since water participates in both mechanical and chemical weathering, more water strongly increases weathering.

 Figure 2. Wet, warm tropical areas have the most weathering.

So how do different climates influence weathering? A cold, dry climate will produce the lowest rate of weathering. A warm, wet climate will produce the highest rate of weathering. The warmer a climate is, the more types of vegetation it will have and the greater the rate of biological weathering (figure 2). This happens because plants and bacteria grow and multiply faster in warmer temperatures.

Some resources are concentrated by weathering processes. In tropical climates, intense chemical weathering carries away all soluble minerals, leaving behind just the least soluble components. The aluminum oxide, bauxite, forms this way and is our main source of aluminum ore.

Contributors and Attributions

CC licensed content, Shared previously

- 9.1: Weathering. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/9.1/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.4.4: Influences on Weathering is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.4.5: Soil Texture and Composition



Figure 1. A loam field.

The inorganic portion of soil is made of many different size particles, and these different size particles are present in different proportions. The combination of these two factors determines some of the properties of the soil.

- A **permeable** soil allows water to flow through it easily because the spaces between the inorganic particles are large and well connected. Sandy or silty soils are considered 'light' soils because they are permeable, water-draining types of soils.
- Soils that have lots of very small spaces are water-holding soils. For example, when clay is present in a soil, the soil is heavier, holds together more tightly, and holds water.
- When a soil contains a mixture of grain sizes, the soil is called a **loam** (figure 1).

When soil scientists want to precisely determine soil type, they measure the percentage of sand, silt, and clay. They plot this information on a triangular diagram, with each size particle at one corner (figure 2). The soil type can then be determined from the location on the diagram. At the top, a soil would be clay; at the left corner, it would be sand, and at the right corner it would be silt. Soils in the lower middle with less than 50% clay are loams.

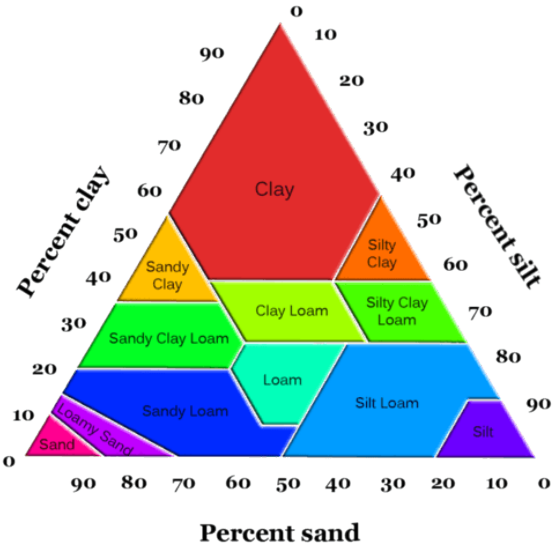


Figure 2. Soil types by particle size.

Using the chart as a guide, what is the composition of a sandy clay loam? If you would like to determine soil type by feel, here's a chart from the USDA to help you.

Contributors and Attributions

CC licensed content, Shared previously

- 9.2: Soils. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/9.2/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.4.5: Soil Texture and Composition is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.4.6: Soil Horizons and Profiles

A residual soil forms over many years, as mechanical and chemical weathering slowly change solid rock into soil. The development of a residual soil may go something like this.



Figure 1. Soil is an important resource. Each soil horizon is distinctly visible in this photograph.

1. The bedrock fractures because of weathering from ice wedging or another physical process.
2. Water, oxygen, and carbon dioxide seep into the cracks to cause chemical weathering.
3. Plants, such as lichens or grasses, become established and produce biological weathering.
4. Weathered material collects until there is soil.
5. The soil develops **soil horizons**, as each layer becomes progressively altered. The greatest degree of weathering is in the top layer. Each successive, lower layer is altered just a little bit less. This is because the first place where water and air come in contact with the soil is at the top.

A cut in the side of a hillside shows each of the different layers of soil. All together, these are called a **soil profile** (figure 1).

The simplest soils have three horizons: topsoil (A horizon), subsoil (B horizon), and C horizon.

Topsoil

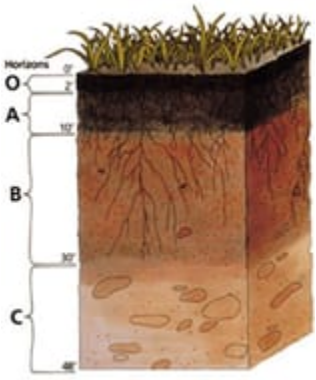


Figure 2. A soil profile is the complete set of soil layers. Each layer is called a horizon.

Called the A horizon, the **topsoil** is usually the darkest layer of the soil because it has the highest proportion of organic material. The topsoil is the region of most intense biological activity: insects, worms, and other animals burrow through it and plants stretch their roots down into it. Plant roots help to hold this layer of soil in place. In the topsoil, minerals may dissolve in the fresh water

that moves through it to be carried to lower layers of the soil. Very small particles, such as clay, may also get carried to lower layers as water seeps down into the ground.

Subsoil

The **B horizon** or **subsoil** is where soluble minerals and clays accumulate. This layer is lighter brown and holds more water than the topsoil because of the presence of iron and clay minerals. There is less organic material. Look at figure 2.

C horizon

The **C horizon** is a layer of partially altered bedrock. There is some evidence of weathering in this layer, but pieces of the original rock are seen and can be identified.

Not all climate regions develop soils, and not all regions develop the same horizons. Some areas develop as many as five or six distinct layers, while others develop only very thin soils or perhaps no soils at all.

Contributors and Attributions

CC licensed content, Shared previously

- 9.2: Soils. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/9.2/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.4.6: Soil Horizons and Profiles is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.4.7: Types of Soils

Although soil scientists recognize thousands of types of soil—each with its own specific characteristics and name—let's consider just three soil types. This will help you to understand some of the basic ideas about how climate produces a certain type of soil, but there are many exceptions to what we will learn right now (figure 1).


Figure 1. Just some of the thousands of soil types.

Pedalfer


Figure 2. A pedalfer is the dark, fertile type of soil that will form in a forested region.

Deciduous trees, the trees that lose their leaves each winter, need at least 65 cm of rain per year. These forests produce soils called **pedalfers**, which are common in many areas of the temperate, eastern part of the United States (figure 2).

The word pedalfer comes from some of the elements that are commonly found in the soil. The *Al* in ped**al**fer is the chemical symbol of the element aluminum, and the *Fe* in pedal**fe**r is the chemical symbol for iron.

Pedalfers are usually a very fertile, dark brown or black soil. Not surprising, they are rich in aluminum clays and iron oxides. Because a great deal of rainfall is common in this climate, most of the soluble minerals dissolve and are carried away, leaving the less soluble clays and iron oxides behind.

Pedocal



Figure 3. A pedocal is the alkaline type of soil that forms in grassland regions.

Pedocal soils form in drier, temperate areas where grasslands and brush are the usual types of vegetation (figure 3). The climates that form pedocals have less than 65 cm rainfall per year, so compared to pedalfers, there is less chemical weathering and less water to dissolve away soluble minerals so more soluble minerals are present and fewer clay minerals are produced. It is a drier region with less vegetation, so the soils have lower amounts of organic material and are less fertile.

A pedocal is named for the calcite enriched layer that forms. Water begins to move down through the soil layers, but before it gets very far, it begins to evaporate. Soluble minerals, like calcium carbonate, concentrate in a layer that marks the lowest place that water was able to reach. This layer is called caliche.

Laterite



Figure 4. A laterite is the type of thick, nutrient poor soil that forms in the rainforest.

In tropical rainforests where it rains literally every day, **laterite** soils form (figure 4). In these hot, wet, tropical regions, intense chemical weathering strips the soils of their nutrients. There is practically no humus. All soluble minerals are removed from the soil and all plant nutrients are carried away. All that is left behind are the least soluble materials, like aluminum and iron oxides. These soils are often red in color from the iron oxides. Laterite soils bake as hard as a brick if they are exposed to the sun.

Many climates types have not been mentioned here. Each produces a distinctive soil type that forms in the particular circumstances found there. Where there is less weathering, soils are thinner but soluble minerals may be present. Where there is intense weathering, soils may be thick but nutrient poor. Soil development takes a very long time, it may take hundreds or even thousands of years for a good fertile topsoil to form. Soil scientists estimate that in the very best soil-forming conditions, soil forms at a rate of about 1mm/year. In poor conditions, soil formation may take thousands of years!

Contributors and Attributions

CC licensed content, Shared previously

- 9.2: Soils. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/9.2/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.4.7: Types of Soils is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.4.8: Causes of Soil Erosion

The agents of soil erosion are the same as the agents of all types of erosion: water, wind, ice, or gravity. Running water is the leading cause of soil erosion, because water is abundant and has a lot of power. Wind is also a leading cause of soil erosion because wind can pick up soil and blow it far away.

Activities that remove vegetation, disturb the ground, or allow the ground to dry are activities that increase erosion. What are some human activities that increase the likelihood that soil will be eroded?

Farming

Agriculture is probably the most significant activity that accelerates soil erosion because of the amount of land that is farmed and how much farming practices disturb the ground (Figure 1). Farmers remove native vegetation and then plow the land to plant new seeds. Because most crops grow only in spring and summer, the land lies fallow during the winter. Of course, winter is also the stormy season in many locations, so wind and rain are available to wash soil away. Tractor tires make deep grooves, which are natural pathways for water. Fine soil is blown away by wind.

The soil that is most likely to erode is the nutrient-rich topsoil, which degrades the farmland.



Figure 1. (a) The bare areas of farmland are especially vulnerable to erosion. (b) Slash-and-burn agriculture leaves land open for soil erosion and is one of the leading causes of soil erosion in the world.

Grazing

Grazing animals (Figure 2) wander over large areas of pasture or natural grasslands eating grasses and shrubs. Grazers expose soil by removing the plant cover for an area. They also churn up the ground with their hooves. If too many animals graze the same land area, the animals' hooves pull plants out by their roots. A land is overgrazed if too many animals are living there.



Grazing animals can cause erosion if they are allowed to overgraze and remove too much or all of the vegetation in a pasture.

Logging and Mining



Figure 3. Logging exposes large areas of land to erosion.

Logging removes trees that protect the ground from soil erosion. The tree roots hold the soil together and the tree canopy protects the soil from hard falling rain. Logging results in the loss of **leaf litter**, or dead leaves, bark, and branches on the forest floor. Leaf litter plays an important role in protecting forest soils from erosion (Figure 3).

Much of the world's original forests have been logged. Many of the tropical forests that remain are currently the site of logging because North America and Europe have already harvested many of their trees (Figure 4). Soils eroded from logged forests clog rivers and lakes, fill estuaries, and bury coral reefs.

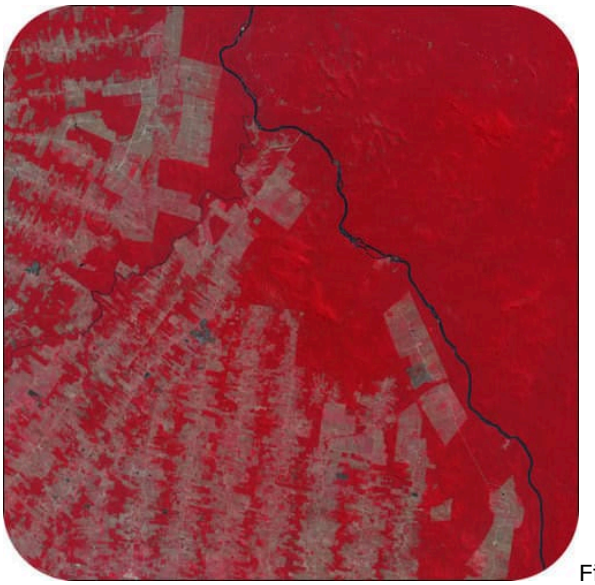


Figure 4. Deforested swatches in Brazil show up as gray amid the bright red tropical rainforest.

Surface mining disturbs the land (Figure 5) and leaves the soil vulnerable to erosion.

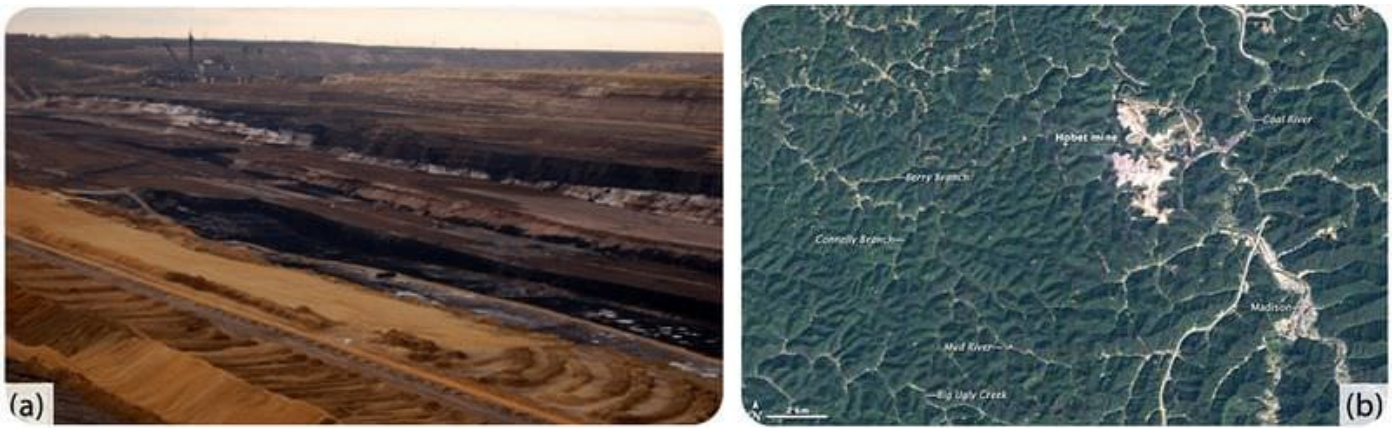


Figure 5. (a) Disturbed land at a coal mine pit in Germany. (b) This coal mine in West Virginia covers more than 10,000 acres (15.6 square miles). Some of the exposed ground is being reclaimed by planting trees.

Construction

Constructing buildings and roads churns up the ground and exposes soil to erosion. In some locations, native landscapes, such as forest and grassland, are cleared, exposing the surface to erosion (in some locations the land that will be built on is farmland). Near construction sites, dirt, picked up by the wind, is often in the air. Completed construction can also contribute to erosion (Figure 6).



Figure 6. Urban areas and parking lots result in less water entering the ground. Water runs off the parking lot onto nearby lands and speeds up erosion in those areas.

Recreational Activities

Recreational activities may accelerate soil erosion. Off-road vehicles disturb the landscape and the area eventually develops bare spots where no plants can grow. In some delicate habitats, even hikers' boots can disturb the ground, so it's important to stay on the trail (Figure 7).



Figure 7. (a) ATVs churn up the soil, accelerating erosion. (b) Hiking trails may become eroded.

Soil erosion is as natural as any other type of erosion, but human activities have greatly accelerated soil erosion. In some locations soil erosion may occur about 10 times faster than its natural rate. Since Europeans settled in North America, about one-third of the topsoil in the area that is now the United States has eroded away.

Summary

- Although soil erosion is a natural process, human activities have greatly accelerated it.
- The agents of soil erosion are the same as of other types of erosion: water, ice, wind, and gravity.
- Soil erosion is more likely where the ground has been disturbed by agriculture, grazing animals, logging, mining, construction, and recreational activities.

Contributors and Attributions

CC licensed content, Shared previously

- Soil Erosion. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Soil-Erosion/lesson/Soil-Erosion-HS-ES/?referrer=featured_content. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.4.8: Causes of Soil Erosion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.4.9: Avoiding Soil Loss



Figure 1. Terraced fields in Sa Pa, Vietnam

How does the terracing shown in Figure 1 photo prevent soil erosion?

Terracing keeps the soil from moving very far downhill since it will only get as far as the next terrace downhill. Water will also be slowed by the terraces and so will be less able to carry tremendous amounts of soil downhill. Terracing is a great way to preserve soil when farming is being done on hillsides.

Soil Erosion

Bad farming practices and a return to normal rainfall levels after an unusually wet period led to the Dust Bowl. In some regions more than 75% of the topsoil blew away. This is the most extreme example of soil erosion the United States has ever seen.



Figure 1. A farmer and his sons walk through a dust storm in Cimarron County, Oklahoma in 1936.

Still, in many areas of the world, the rate of soil erosion is many times greater than the rate at which it is forming. Drought, insect plagues, or outbreaks of disease are natural cycles of events that can negatively impact ecosystems and the soil, but there are also many ways in which humans neglect or abuse this important resource. Soils can also be contaminated if too much salt accumulates in the soil or where pollutants sink into the ground.

One harmful practice is removing the vegetation that helps to hold soil in place. Sometimes just walking or riding your bike over

the same place will kill the grass that normally grows there. Land is also deliberately cleared or deforested for wood. The loose soils then may be carried away by wind or running water.

Soil Conservation

Soil is only a renewable resource if it is carefully managed. There are many practices that can protect and preserve soil resources.

Organic Material



Figure 3. Organic material can be added to soil to help increase its fertility.

Adding organic material to the soil in the form of plant or animal waste, such as compost or manure, increases the fertility of the soil and improves its ability to hold on to water and nutrients (Figure 3). Inorganic fertilizer can also temporarily increase the fertility of a soil and may be less expensive or time consuming, but it does not provide the same long-term improvements as organic materials.

Preventing Soil Erosion

Soil is a natural resource that is vitally important for sustaining natural habitats and for growing food. Although soil is a renewable resource, it is renewed slowly, taking hundreds or thousands of years for a good fertile soil to develop.



Figure 4. Steep slopes can be terraced to make level planting areas and decrease surface water runoff and erosion.

Most of the best land for farming is already being cultivated. With human populations continuing to grow, it is extremely important to protect our soil resources. Agricultural practices such as rotating crops, alternating the types of crops planted in each row, and planting nutrient-rich cover crops all help to keep soil more fertile as it is used season after season. Planting trees as windbreaks, plowing along contours of the field, or building terraces into steeper slopes will all help to hold soil in place (Figure 4). No-till or low-tillage farming helps to keep soil in place by disturbing the ground as little as possible when planting.

The rate of topsoil loss in the United States and other developed countries has decreased recently as better farming practices have been adopted. Unfortunately, in developing nations, soil is often not protected.

The table below shows some steps that we can take to prevent erosion. Some are things that can be done by farmers or developers. Others are things that individual homeowners or community members can implement locally.

Source of Erosion	Strategies for Prevention
Agriculture	• Leave leaf litter on the ground in the winter. • Grow cover crops, special crops grown in the winter to cover the soil. • Plant tall trees around fields to buffer the effects of wind. • Drive tractors as little as possible. • Use drip irrigation that puts small amounts of water in the ground frequently. • Avoid watering crops with sprinklers that make big water drops on the ground. • Keep fields as flat as possible to avoid soil eroding down hill.
Grazing Animals	• Move animals throughout the year, so they don't consume all the vegetation in one spot. • Keep animals away from stream banks, where hills are especially prone to erosion.
Logging and Mining	• Reduce the amount of land that is logged and mined. • Reduce the number of roads that are built to access logging areas. • Avoid logging and mining on steep lands. • Cut only small areas at one time and quickly replant logged areas with new seedlings.
Development	• Reduce the amount of land area that is developed into urban areas, parking lots, etc. • Keep as much "green space" in cities as possible, such as parks or strips where plants can grow. • Invest in and use new technologies for parking lots that make them permeable to water in order to reduce runoff of water.
Recreational Activities	• Avoid using off-road vehicles on hilly lands. • Stay on designated trails.
Building Construction	• Avoid building on steep hills. • Grade surrounding land to distribute water rather than collecting it in one place. • Where water collects, drain to creeks and rivers. • Landscape with plants that minimize erosion.

Contributors and Attributions

CC licensed content, Shared previously

- Avoiding Soil Loss. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Avoiding-Soil-Loss/lesson/Avoiding-Soil-Loss-HS-ES/?referrer=featured_content. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.4.9: Avoiding Soil Loss is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.5: PLATE TECTONICS

Learning Objectives

Recognize the ongoing development of, evidence for, and large scale planetary effects of the Theory of Plate Tectonics.

- Describe Continental Drift including supporting evidence.
- Critique and interpret major types of evidence supporting the Theory of Plate Tectonics.
- Describe different geological processes in relation to plate tectonics.

4.5.1: Continental Drift

4.5.2: Wegener and the Continental Drift Hypothesis

4.5.3: Theory of Plate Tectonics

4.5.4: Developing the Theory

4.5.5: Supercontinents

4.5.6: Volcanic Arcs

4.5.7: Hot Spots

This page titled Chapter 4.5: Plate Tectonics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.5.1: Continental Drift

The Continental Drift Idea



Figure 1. The continents fit together like pieces of a puzzle. This is how they looked 250 million years ago.

Find a map of the continents and cut each one out. Better yet, use a map where the edges of the continents show the continental shelf. That's the true size and shape of a continent. Can you fit the pieces together? The easiest link is between the eastern Americas and western Africa and Europe, but the rest can fit together too (Figure 1).

Alfred Wegener proposed that the continents were once united into a single supercontinent named Pangaea, meaning *all earth* in ancient Greek. He suggested that Pangaea broke up long ago and that the continents then moved to their current positions. He called his hypothesis **continental drift.**

Evidence for Continental Drift

Besides the way the continents fit together, Wegener and his supporters collected a great deal of evidence for the continental drift hypothesis.

- Identical rocks, of the same type and age, are found on both sides of the Atlantic Ocean. Wegener said the rocks had formed side-by-side and that the land had since moved apart.
- Mountain ranges with the same rock types, structures, and ages are now on opposite sides of the Atlantic Ocean. The Appalachians of the eastern United States and Canada, for example, are just like mountain ranges in eastern Greenland, Ireland, Great Britain, and Norway (figure 2). Wegener concluded that they formed as a single mountain range that was separated as the continents drifted.



Figure 2. The similarities between the Appalachian and the eastern Greenland mountain ranges are evidences for the continental drift hypothesis.

- Ancient fossils of the same species of extinct plants and animals are found in rocks of the same age but are on continents that are now widely separated (figure 3). Wegener proposed that the organisms had lived side by side, but that the lands had moved apart after they were dead and fossilized. He suggested that the organisms would not have been able to travel across the oceans.
  - Fossils of the seed fern *Glossopteris* were too heavy to be carried so far by wind.
  - *Mesosaurus* was a swimming reptile but could only swim in fresh water.
  - *Cynognathus* and *Lystrosaurus* were land reptiles and were unable to swim.

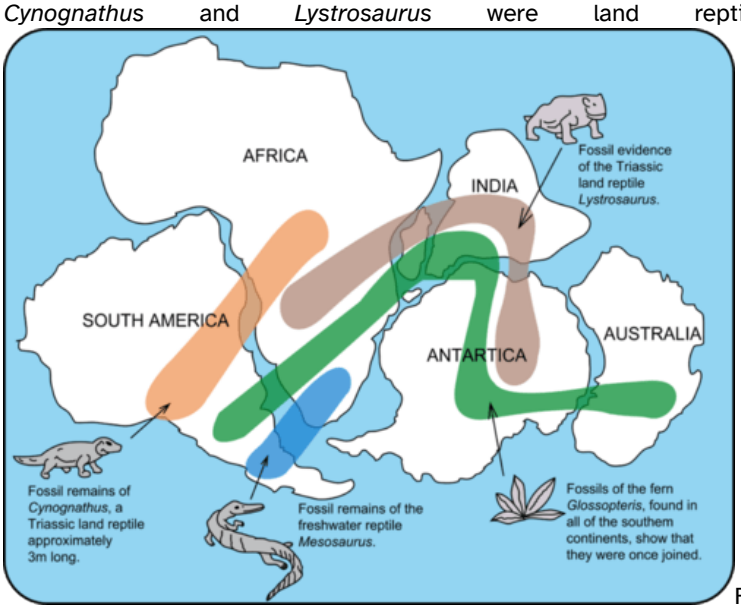


Figure 3. Wegener used fossil evidence to support his continental drift hypothesis. The fossils of these organisms are found on lands that are now far apart.

- Grooves and rock deposits left by ancient glaciers are found today on different continents very close to the equator. This would indicate that the glaciers either formed in the middle of the ocean and/or covered most of the Earth. Today glaciers only form on land and nearer the poles. Wegener thought that the glaciers were centered over the southern land mass close to the South Pole and the continents moved to their present positions later on.
- Coral reefs and coal-forming swamps are found in tropical and subtropical environments, but ancient coal seams and coral reefs are found in locations where it is much too cold today. Wegener suggested that these creatures were alive in warm climate zones and that the fossils and coal later had drifted to new locations on the continents.

Take a look at this animation showing that Earth's climate belts remain in roughly the same position while the continents move and this animation showing how the continents split up.

Although Wegener's evidence was sound, most geologists at the time rejected his hypothesis of continental drift. Why do you think they did not accept continental drift?

Scientists argued that there was no way to explain how solid continents could plow through solid oceanic crust. Wegener's idea was nearly forgotten until technological advances presented even more evidence that the continents moved and gave scientists

the tools to develop a mechanism for Wegener's drifting continents.

Magnetic Polarity Evidence

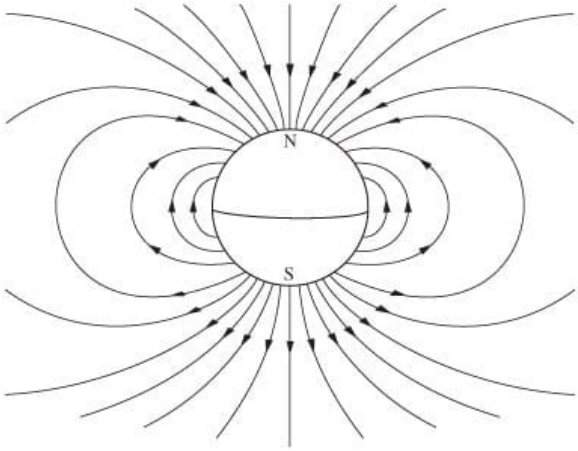


Figure 4. Earth's magnetic field is like a magnet with its north pole near the geographic North Pole and the south pole near the geographic South Pole.

Puzzling new evidence came in the 1950s from studies on the Earth's magnetic history (figure 4). Scientists used **magnetometers**, devices capable of measuring the magnetic field intensity, to look at the magnetic properties of rocks in many locations.

Magnetite crystals are like tiny magnets that point to the north magnetic pole as they crystallize from magma. The crystals record both the direction and strength of the **magnetic field** at the time. The direction is known as the field's **magnetic polarity.**

Magnetic Polarity on the Same Continent with Rocks of Different Ages

Geologists noted important things about the magnetic polarity of different aged rocks on the same continent:

- Magnetite crystals in fresh volcanic rocks point to the current magnetic north pole (figure 5) no matter what continent or where on the continent the rocks are located.

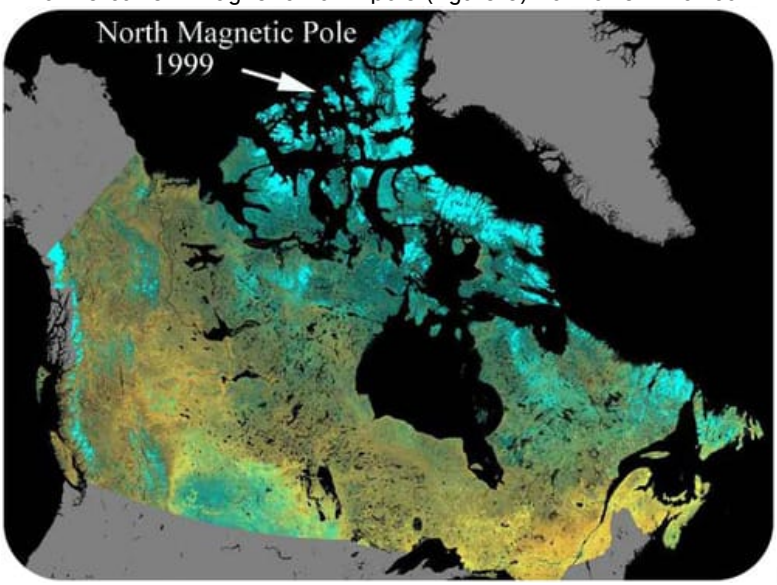


Figure 5. Earth's current north magnetic pole is in northern Canada.

- Older rocks that are the same age and are located on the same continent point to the same location, but that location is not the current north magnetic pole.
- Older rock that are of different ages do not point to the same locations or to the current magnetic north pole.

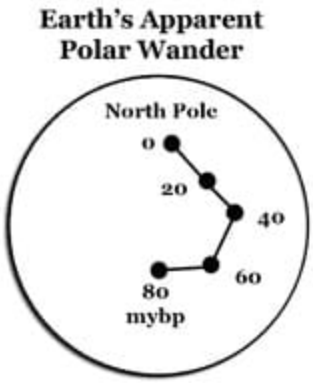


Figure 6. The location of the north magnetic north pole 80 million years before present (mybp), then 60, 40, 20, and now.

In other words, although the magnetite crystals were pointing to the magnetic north pole, the location of the pole seemed to wander. Scientists were amazed to find that the north magnetic pole changed location through time (figure 6).

There are three possible explanations for this:

1. The continents remained fixed and the north magnetic pole moved.
2. The north magnetic pole stood still and the continents moved.
3. Both the continents and the north pole moved.

Magnetic Polarity on Different Continents with Rocks of the Same Age

Geologists noted that for rocks of the same age but on different continents, the little magnets pointed to different magnetic north poles.

- 400-million-year-old magnetite in Europe pointed to a different north magnetic pole than the same-aged magnetite in North America.
- 250 million years ago, the north poles were also different for the two continents.

The scientists looked again at the three possible explanations. Only one can be correct. If the continents had remained fixed while the north magnetic pole moved, there must have been two separate north poles. Since there is only one north pole today, the only reasonable explanation is that the north magnetic pole has remained fixed but that the continents have moved.

To test this, geologists fitted the continents together as Wegener had done. It worked! There has only been one magnetic north pole and the continents have drifted (figure 7). They named the phenomenon of the magnetic pole that seemed to move but actually did not **apparent polar wander.**

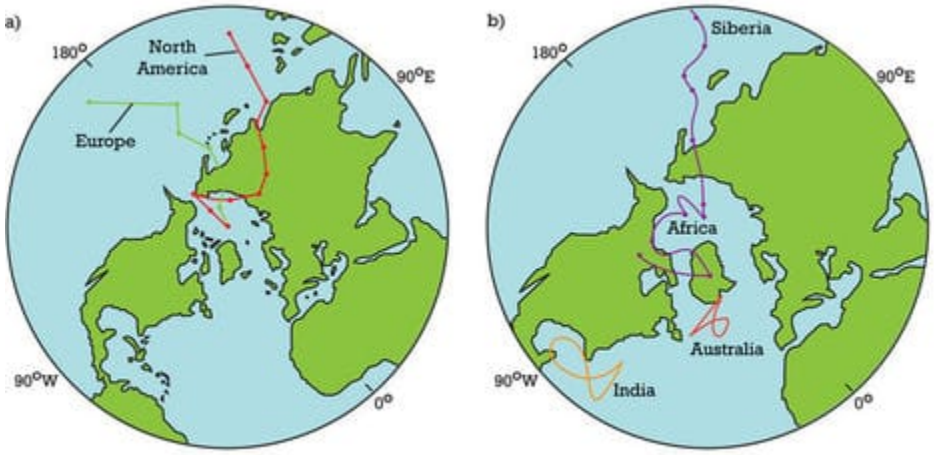


Figure 7. On the left: The apparent

north pole for Europe and North America if the continents were always in their current locations. The two paths merge into one if the continents are allowed to drift.

This evidence for continental drift gave geologists renewed interest in understanding how continents could move about on the planet's surface.

Lesson Summary

- In the early part of the 20th century, scientists began to put together evidence that the continents could move around on Earth's surface.
- The evidence for continental drift included the fit of the continents; the distribution of ancient fossils, rocks, and mountain ranges; and the locations of ancient climatic zones.
- Although the evidence for continental drift was extremely strong, scientists rejected the idea because no mechanism for how solid continents could move around on the solid earth was developed.
- The discovery of apparent polar wander renewed scientists interest in continental drift.

Contributors and Attributions

CC licensed content, Shared previously

- 6.2: Continental Drift. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/6.2/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.5.1: Continental Drift is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.5.2: Wegener and the Continental Drift Hypothesis

Scientists still do not appear to understand sufficiently that all earth sciences must contribute evidence toward unveiling the state of our planet in earlier times, and that the truth of the matter can only be reached by combing all this evidence. . . . It is only by combing the information furnished by all the earth sciences that we can hope to determine 'truth' here, that is to say, to find the picture that sets out all the known facts in the best arrangement and that therefore has the highest degree of probability. Further, we have to be prepared always for the possibility that each new discovery, no matter what science furnishes it, may modify the conclusions we draw.

— Alfred L. Wegener, *The Origins of Continents and Oceans*, first published in 1915.

Wegener put together a tremendous amount of evidence that the continents had been joined. He advocated using scientific evidence to find the "truth." As his colleague, are you convinced? Let's explore.

Wegener's Continental Drift Hypothesis



Figure 1. Alfred Wegener suggested that continental drift occurred as continents cut through the ocean floor, in the same way as this icebreaker plows through sea ice.

Wegener put his idea and his evidence together in his book *The Origin of Continents and Oceans*, first published in 1915. New editions with additional evidence were published later in the decade. In his book he said that around 300 million years ago the continents had all been joined into a single landmass he called Pangaea, meaning "all earth" in ancient Greek. The supercontinent later broke apart and the continents having been moving into their current positions ever since. He called his hypothesis **continental drift**.

The Problem with the Hypothesis

Wegener's idea seemed so outlandish at the time that he was ridiculed by other scientists. What do you think the problem was? To his colleagues, his greatest problem was that he had no plausible mechanism for how the continents could move through the oceans. Based on his polar experiences, Wegener suggested that the continents were like icebreaking ships plowing through ice sheets. The continents moved by centrifugal and tidal forces. As Wegener's colleague, how would you go about showing whether these forces could move continents? What observations would you expect to see on these continents?



Figure 2. Early hypotheses proposed that centrifugal forces moved continents. This is the same force that moves the swings outward on a spinning carnival ride.

Scientists at the time calculated that centrifugal and tidal forces were too weak to move continents. When one scientist did calculations that assumed that these forces were strong enough to move continents, his result was that if Earth had such strong forces the planet would stop rotating in less than one year. In addition, scientists also thought that the continents that had been plowing through the ocean basins should be much more deformed than they are.

Wegener answered his question of whether Africa and South America had once been joined. But a hypothesis is rarely accepted without a mechanism to drive it. Are you going to support Wegener? A very few scientists did, since his hypothesis elegantly explained the similar fossils and rocks on opposite sides of the ocean, but most did not.

Mantle Convection

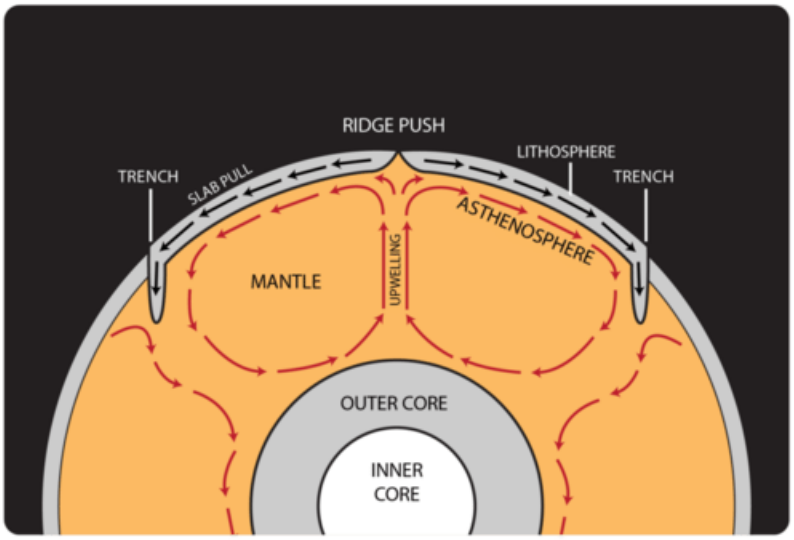


Figure 3. Thermal convection occurs as hot rock in the deep mantle rises towards the Earth's surface. This rock then spreads out and cools, sinking back towards the core, where it can be heated again. This circulation of rock through the mantle creates convection cells.

Wegener had many thoughts regarding what could be the driving force behind continental drift. Another of Wegener's colleagues, Arthur Holmes, elaborated on Wegener's idea that there is thermal convection in the mantle.

In a **convection cell**, material deep beneath the surface is heated so that its density is lowered and it rises. Near the surface it becomes cooler and denser, so it sinks. Holmes thought this could be like a conveyor belt. Where two adjacent convection cells rise to the surface, a continent could break apart with pieces moving in opposite directions. Although this sounds like a great idea, there was no real evidence for it, either.

Alfred Wegener died in 1930 on an expedition on the Greenland icecap. For the most part the continental drift idea was put to rest for a few decades, until technological advances presented even more evidence that the continents moved and gave scientists the tools to develop a mechanism for Wegener's drifting continents. Since you're on a virtual field trip, you get to go along with them as well.

Summary

- Alfred Wegener published his idea that the continents had been joined as a single landmass, which he called Pangaea, about 300 million years ago.
- Wegener's idea was mostly ridiculed, in part because Wegener could not develop a plausible mechanism for continents moving through oceanic crust.
- Calculations showed that his idea about centrifugal and tidal forces powering the continents could not be right.
- Wegener also thought about mantle convection an idea expanded on by Arthur Holmes as the driving force for continental drift. There was no evidence available to support the idea at the time.

Contributors and Attributions

CC licensed content, Shared previously

- Wegener and the Continental Drift Hypothesis. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Wegener-and-the-Continental-Drift-Hypothesis/lesson/Wegener-and-the-Continental-Drift-Hypothesis-HS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.5.2: Wegener and the Continental Drift Hypothesis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.5.3: Theory of Plate Tectonics

When the concept of seafloor spreading came along, scientists recognized that it was the mechanism to explain how continents could move around Earth's surface. Like the scientists before us, we will now merge the ideas of continental drift and seafloor spreading into the theory of plate tectonics.

Watch this video about continental drift and the mechanism of seafloor spreading create plate tectonics.

Earth's Tectonic Plates

Seafloor and continents move around on Earth's surface, but what is actually moving? What portion of the Earth makes up the "plates" in plate tectonics? This question was also answered because of technology developed during war times – in this case, the Cold War. The **plates** are made up of the lithosphere.

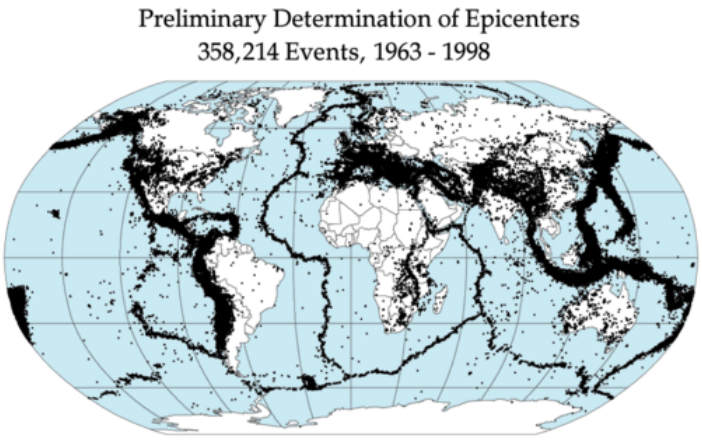


Figure 1. Earthquakes outline the plates.

During the 1950s and early 1960s, scientists set up seismograph networks to see if enemy nations were testing atomic bombs. These seismographs also recorded all of the earthquakes around the planet. The seismic records could be used to locate an earthquake's **epicenter**, the point on Earth's surface directly above the place where the earthquake occurs.

Earthquake epicenters outline the plates. Mid-ocean ridges, trenches, and large faults mark the edges of the plates, and this is where earthquakes occur (figure 1).

The lithosphere is divided into a dozen major and several minor plates (figure 2). The plates' edges can be drawn by connecting the dots that mark earthquakes' epicenters. A single plate can be made of all oceanic lithosphere or all continental lithosphere, but nearly all plates are made of a combination of both.

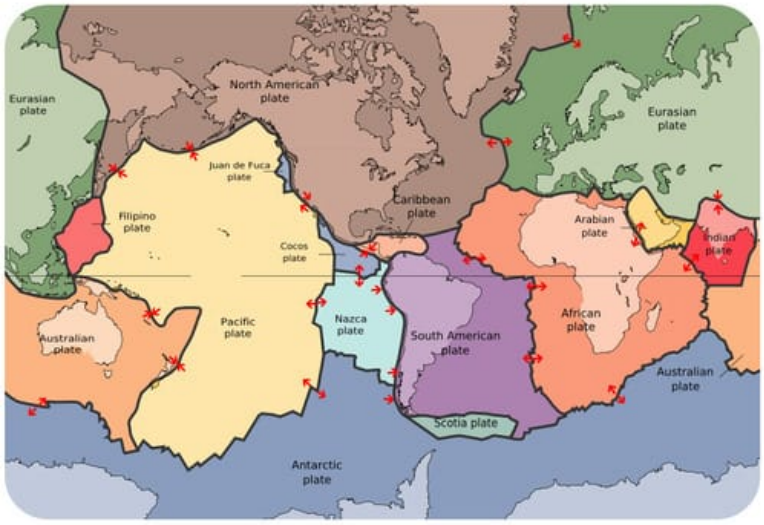


Figure 2. The lithospheric plates and their names. The arrows show whether the plates are moving apart, moving together, or sliding past each other.

Movement of the plates over Earth's surface is termed **plate tectonics**. Plates move at a rate of a few centimeters a year, about the same rate fingernails grow.

How Plates Move

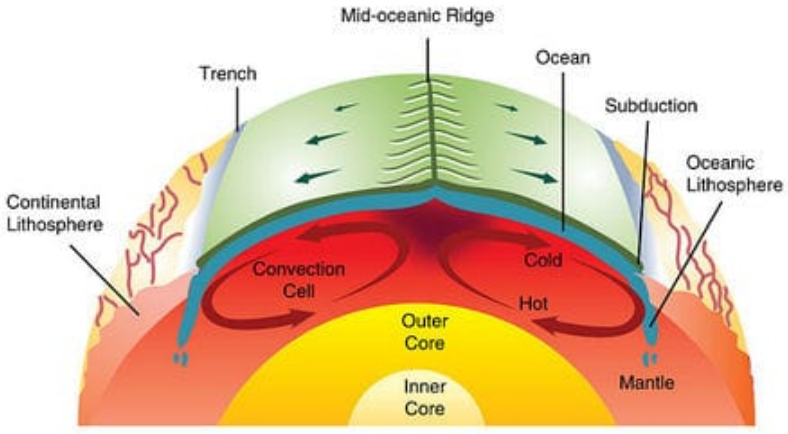


Figure 3. Mantle convection drives plate tectonics. Hot material rises at mid-ocean ridges and sinks at deep sea trenches, which keeps the plates moving along the Earth's surface.

If seafloor spreading drives the plates, what drives seafloor spreading? Picture two convection cells side-by-side in the mantle, similar to the illustration in figure 3.

1. Hot mantle from the two adjacent cells rises at the ridge axis, creating new ocean crust.
2. The top limb of the convection cell moves horizontally away from the ridge crest, as does the new seafloor.
3. The outer limbs of the convection cells plunge down into the deeper mantle, dragging oceanic crust as well. This takes place at the deep sea trenches.
4. The material sinks to the core and moves horizontally.
5. The material heats up and reaches the zone where it rises again.

Check out this animation of mantle convection and watch this video:

A link to an interactive elements can be found at the bottom of this page.

Plate Boundaries

Plate boundaries are the edges where two plates meet. Most geologic activities, including volcanoes, earthquakes, and mountain building, take place at plate boundaries. How can two plates move relative to each other?

- **Divergent plate boundaries**: the two plates move away from each other.
- **Convergent plate boundaries**: the two plates move towards each other.
- **Transform plate boundaries**: the two plates slip past each other.

The type of plate boundary and the type of crust found on each side of the boundary determines what sort of geologic activity will be found there.

Divergent Plate Boundaries

Plates move apart at mid-ocean ridges where new seafloor forms. Between the two plates is a rift valley. Lava flows at the surface cool rapidly to become basalt, but deeper in the crust, magma cools more slowly to form gabbro. So the entire ridge system is made up of igneous rock that is either extrusive or intrusive. Earthquakes are common at mid-ocean ridges since the movement of magma and oceanic crust results in crustal shaking. The vast majority of mid-ocean ridges are located deep below the sea (figure 4).

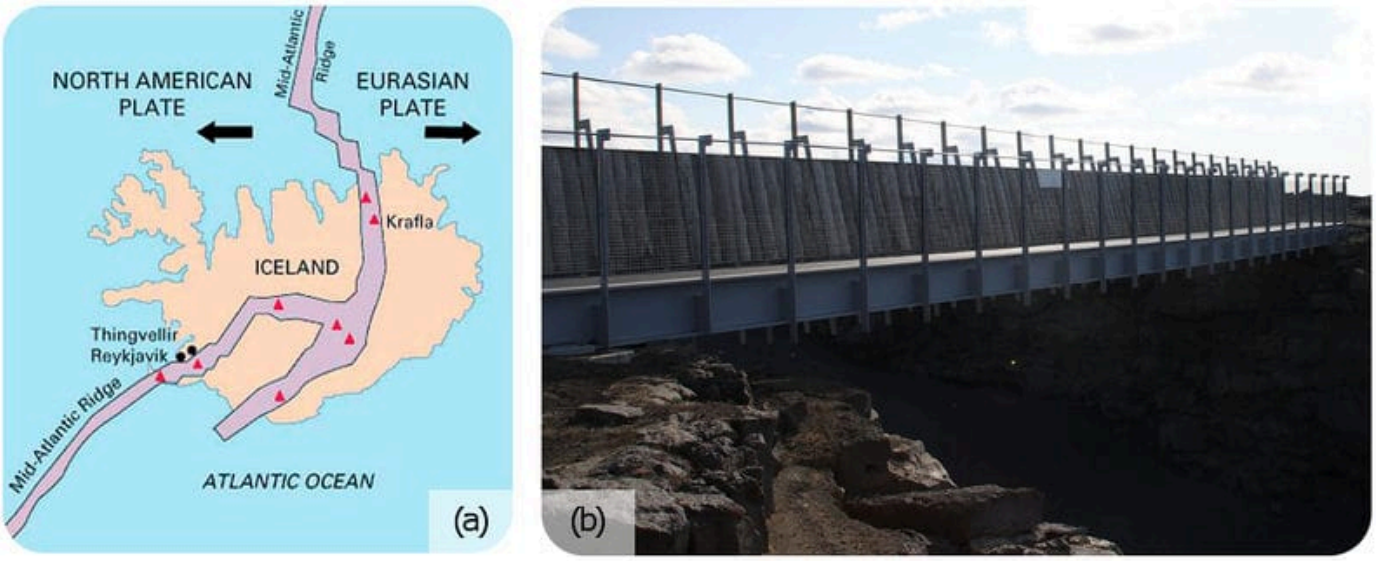


Figure 4. (a) Iceland is the one location where the ridge is located on land: the Mid-Atlantic Ridge separates the North American and Eurasian plates; (b) The rift valley in the Mid-Atlantic Ridge on Iceland.



Figure 5. The Arabian, Indian, and African plates are rifting apart, forming the Great Rift Valley in Africa. The Dead Sea fills the rift with seawater.

Check out these animations:

- Divergent plate boundary at mid-ocean ridge

- Divergent plate boundary

Can divergent plate boundaries occur within a continent? What is the result? **Incontinental rifting** (figure 5), magma rises beneath the continent, causing it to become thinner, break, and ultimately split apart. New ocean crust erupts in the void, creating an ocean between continents.

Convergent Plate Boundaries

When two plates converge, the result depends on the type of lithosphere the plates are made of. No matter what, smashing two enormous slabs of lithosphere together results in magma generation and earthquakes.

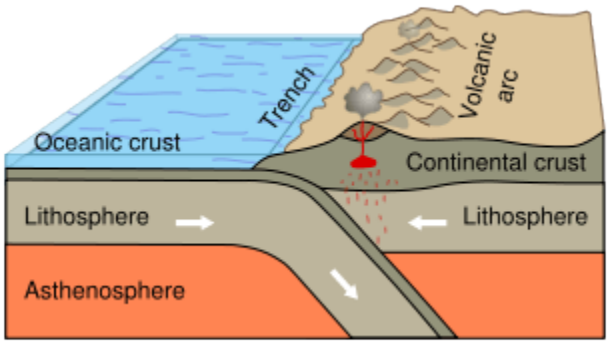


Figure 6. Subduction of an oceanic plate beneath a continental plate causes earthquakes and forms a line of volcanoes known as a continental arc.

Ocean-Continent

When oceanic crust converges with continental crust, the denser oceanic plate plunges beneath the continental plate. This process, called **subduction**, occurs at the oceanic trenches (figure 6). The entire region is known as a **subduction zone**. Subduction zones have a lot of intense earthquakes and volcanic eruptions. The subducting plate causes melting in the mantle. The magma rises and erupts, creating volcanoes. These coastal volcanic mountains are found in a line above the subducting plate (figure 7). The volcanoes are known as a **continental arc**.

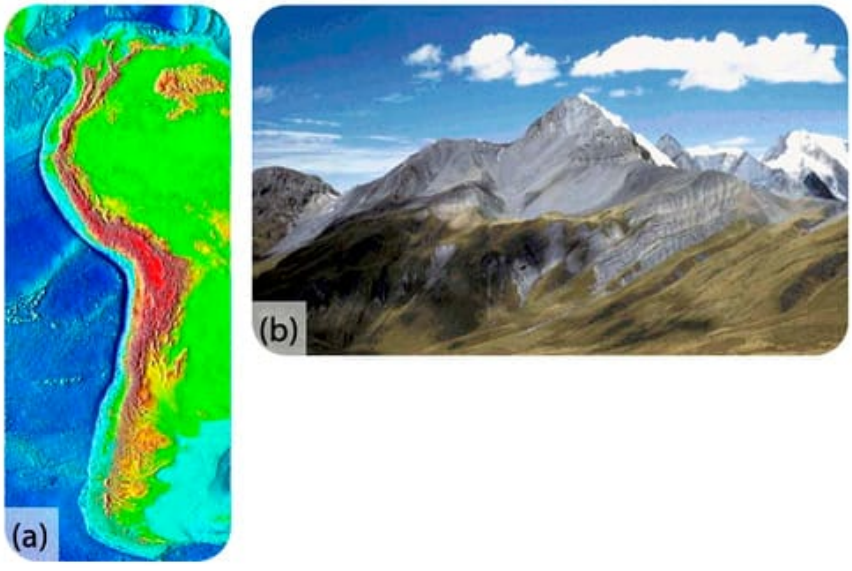


Figure 7. (a) At the trench lining the western margin of South America, the Nazca plate is subducting beneath the South American plate, resulting in the Andes Mountains (brown and red uplands); (b) Convergence has pushed up limestone in the Andes Mountains where volcanoes are common.

The movement of crust and magma causes earthquakes. Look at this map of earthquake epicenters at subduction zones. This animation shows the relationship between subduction of the lithosphere and creation of a volcanic arc.

The volcanoes of northeastern California—Lassen Peak, Mount Shasta, and Medicine Lake volcano—along with the rest of the Cascade Mountains of the Pacific Northwest are the result of subduction of the Juan de Fuca plate beneath the North American plate (figure 8). The Juan de Fuca plate is created by seafloor spreading just offshore at the Juan de Fuca ridge.

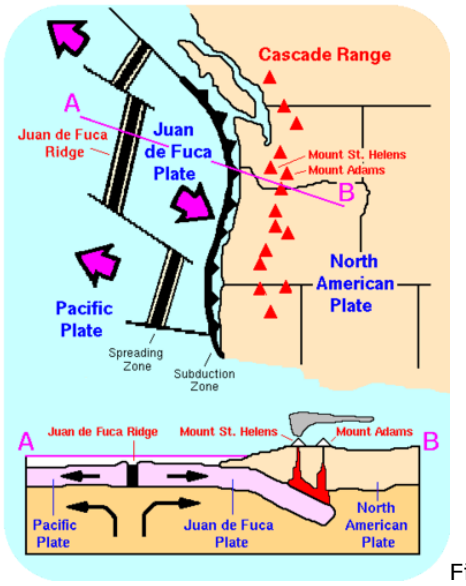


Figure 8. The Cascade Mountains of the Pacific Northwest are a continental arc.

If the magma at a continental arc is felsic, it may be too viscous (thick) to rise through the crust. The magma will cool slowly to form granite or granodiorite. These large bodies of intrusive igneous rocks are called **batholiths**, which may someday be uplifted to form a mountain range (figure 9).



Figure 9. The Sierra Nevada batholith cooled beneath a volcanic arc roughly 200 million years ago. The rock is well exposed here at Mount Whitney. Similar batholiths are likely forming beneath the Andes and Cascades today.

Ocean-Ocean

When two oceanic plates converge, the older, denser plate will subduct into the mantle. An ocean trench marks the location where the plate is pushed down into the mantle. The line of volcanoes that grows on the upper oceanic plate is an **island arc**. Do you think earthquakes are common in these regions (figure 10)?

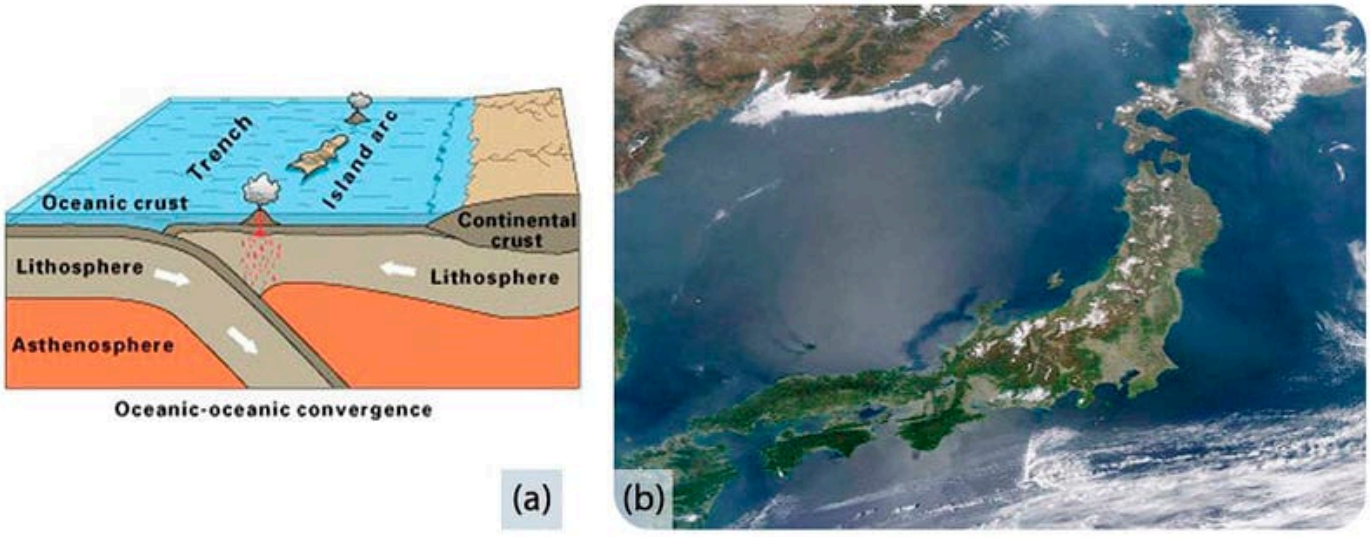


Figure 10. (a) Subduction of an ocean plate beneath an ocean plate results in a volcanic island arc, an ocean trench and many earthquakes. (b) Japan is an arc-shaped island arc composed of volcanoes off the Asian mainland, as seen in this satellite image.

Check out this animation of an ocean continent plate boundary.

Continent-Continent

Continental plates are too buoyant to subduct. What happens to continental material when it collides? Since it has nowhere to go but up, this creates some of the world's largest mountains ranges (figure 11). Magma cannot penetrate this thick crust so there are no volcanoes, although the magma stays in the crust. Metamorphic rocks are common because of the stress the continental crust experiences. With enormous slabs of crust smashing together, continent-continent collisions bring on numerous and large earthquakes.

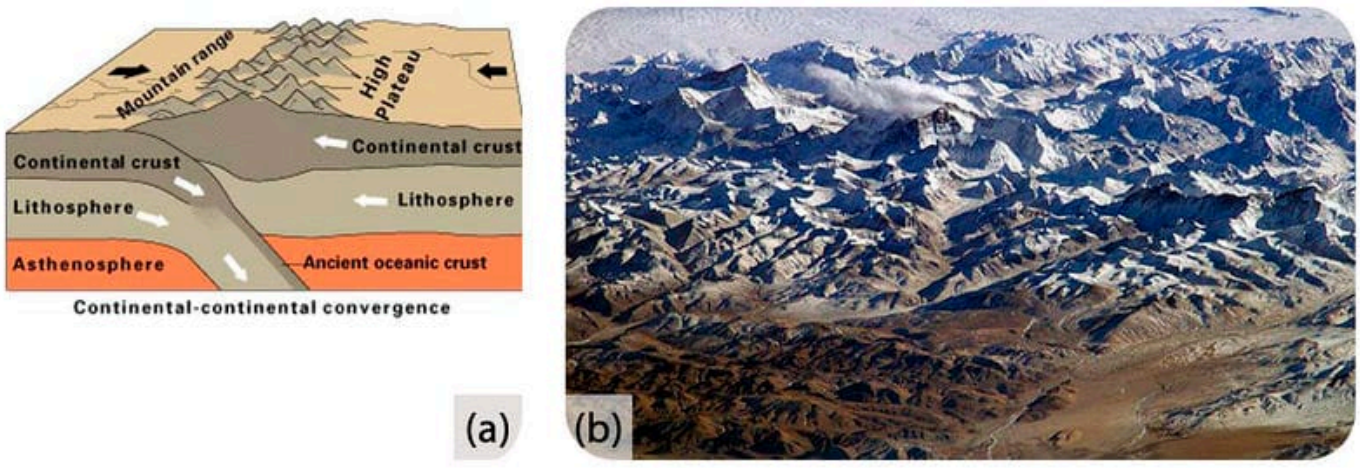
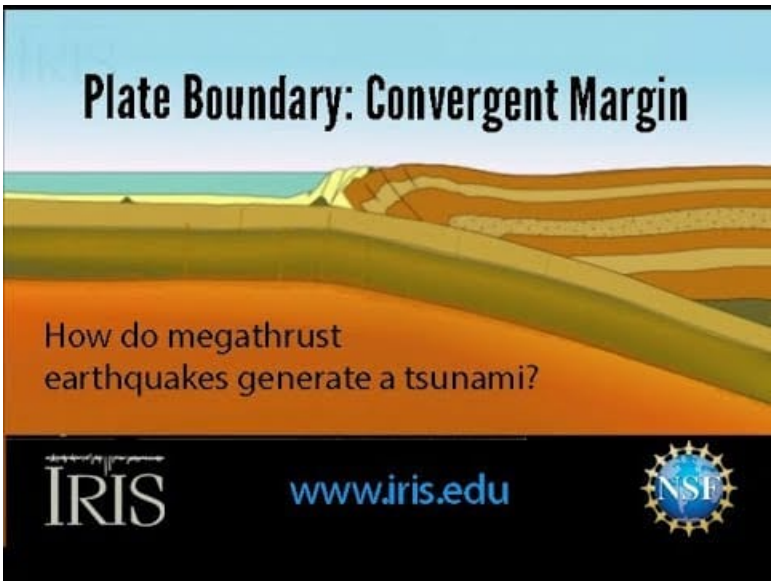


Figure 11. (a) In continent-continent convergence, the plates push upward to create a high mountain range. (b) The world's highest mountains, the Himalayas, are the result of the collision of the Indian Plate with the Eurasian Plate, seen in this photo from the International Space Station.

Check out this short animation of the Indian Plate colliding with the Eurasian Plate.

Watch this animation of the Himalaya rising.



A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/geo/?p=196

The Appalachian Mountains are the remnants of a large mountain range that was created when North America rammed into Eurasia about 250 million years ago.

Transform Plate Boundaries

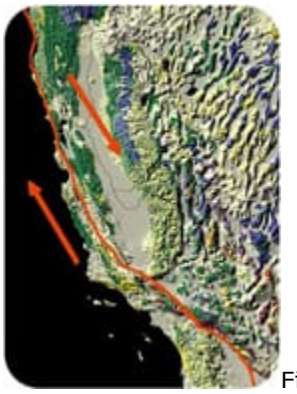


Figure 12. At the San Andreas Fault in California, the Pacific Plate is sliding northwest relative to the North American plate, which is moving southeast. At the northern end of the picture, the transform boundary turns into a subduction zone.

Transform plate boundaries are seen as **transform faults**, where two plates move past each other in opposite directions. Transform faults on continents bring massive earthquakes (figure 12).

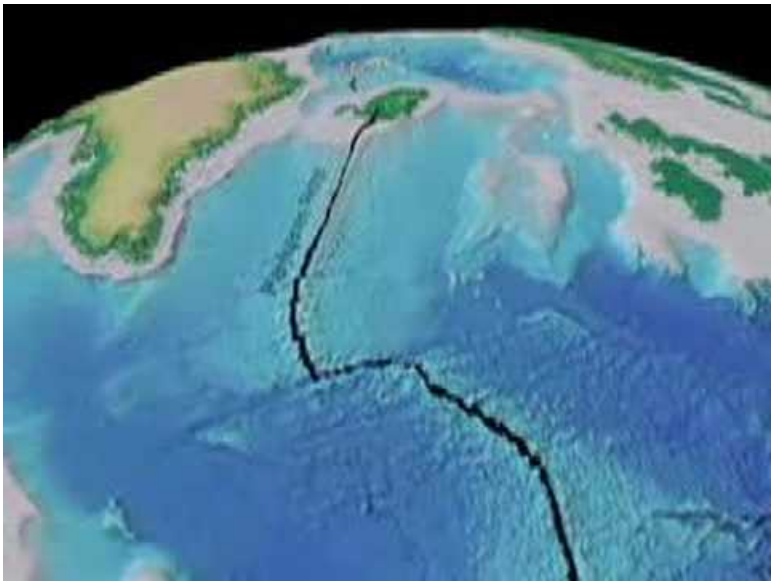
California is very geologically active. What are the three major plate boundaries in or near California (figure 13)?

1. A transform plate boundary between the Pacific and North American plates creates the San Andreas Fault, the world's most notorious transform fault.
2. Just offshore, a divergent plate boundary, Juan de Fuca ridge, creates the Juan de Fuca plate.
3. A convergent plate boundary between the Juan de Fuca oceanic plate and the North American continental plate creates the Cascades volcanoes.



Figure 13. This map shows the three major plate boundaries in or near California.

A brief review of the three types of plate boundaries and the structures that are found there is the subject of this wordless video.



A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/geo/?p=196

Earth's Changing Surface

Geologists know that Wegener was right because the movements of continents explain so much about the geology we see. Most of the geologic activity that we see on the planet today is because of the interactions of the moving plates.



Figure 14. Mountain ranges of North America.

In the map of North America (figure 14), where are the mountain ranges located? Using what you have learned about plate tectonics, try to answer the following questions:

1. What is the geologic origin of the Cascades Range? The Cascades are a chain of volcanoes in the Pacific Northwest. They are not labelled on the diagram but they lie between the Sierra Nevada and the Coastal Range.
2. What is the geologic origin of the Sierra Nevada? (Hint: These mountains are made of granitic intrusions.)
3. What is the geologic origin of the Appalachian Mountains along the Eastern US?



Figure 15. About 200 million years ago, the Appalachian Mountains of eastern North America were probably once as high as the Himalaya, but they have been weathered and eroded significantly since the breakup of Pangaea.

Remember that Wegener used the similarity of the mountains on the west and east sides of the Atlantic as evidence for his continental drift hypothesis. The Appalachian mountains formed at a convergent plate boundary as Pangaea came together (figure 15).

Before Pangaea came together, the continents were separated by an ocean where the Atlantic is now. The proto-Atlantic ocean shrank as the Pacific ocean grew. Currently, the Pacific is shrinking as the Atlantic is growing. This **supercontinent cycle** is responsible for most of the geologic features that we see and many more that are long gone (figure 16).

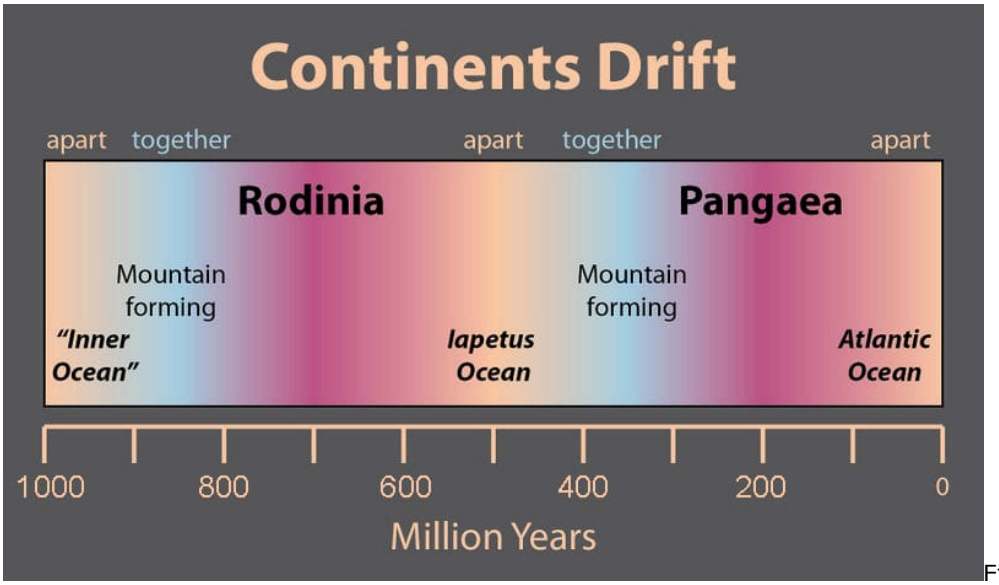


Figure 16. Scientists think that the creation and breakup of a supercontinent takes place about every 500 million years. The supercontinent before Pangaea was Rodinia. A new continent will form as the Pacific ocean disappears.

This animation shows the movement of continents over the past 600 million years beginning with the breakup of Rodinia.

Lesson Summary

- Plates of lithosphere move because of convection currents in the mantle. One type of motion is produced by seafloor spreading.
- Plate boundaries can be located by outlining earthquake epicenters.
- Plates interact at three types of plate boundaries: divergent, convergent and transform.
- Most of the Earth's geologic activity takes place at plate boundaries.
- At a divergent boundary, volcanic activity produces a mid ocean ridge and small earthquakes.
- At a convergent boundary with at least one oceanic plate, an ocean trench, a chain of volcanoes develops and many earthquakes occur.
- At a convergent boundary where both plates are continental, mountain ranges grow and earthquakes are common.
- At a transform boundary, there is a transform fault and massive earthquakes occur but there are no volcanoes.
- Processes acting over long periods of time create Earth's geographic features.

Contributors and Attributions

CC licensed content, Shared previously

- 6.4: Theory of Plate Tectonics. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/6.4/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.5.3: Theory of Plate Tectonics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.5.4: Developing the Theory

In line with other previous and contemporaneous proposals, in 1912 the meteorologist Alfred Wegener amply described what he called continental drift, expanded in his 1915 book *The Origin of Continents and Oceans*[1], and the scientific debate started that would end up fifty years later in the theory of plate tectonics.[2] Starting from the idea (also expressed by his forerunners) that the present continents once formed a single land mass (which was called Pangea later on) that drifted apart, thus releasing the continents from the Earth's mantle and likening them to "icebergs" of low density granite floating on a sea of denser basalt.[3]

Supporting evidence for the idea came from the dove-tailing outlines of South America's east coast and Africa's west coast, and from the matching of the rock formations along these edges. Confirmation of their previous contiguous nature also came from the fossil plants *Glossopteris* and *Gangamopteris*, and the therapsid or mammal-like reptile *Lystrosaurus*, all widely distributed over South America, Africa, Antarctica, India and Australia. The evidence for such an erstwhile joining of these continents was patent to field geologists working in the southern hemisphere. The South African Alex du Toit put together a mass of such information in his 1937 publication *Our Wandering Continents*, and went further than Wegener in recognising the strong links between the Gondwana fragments.

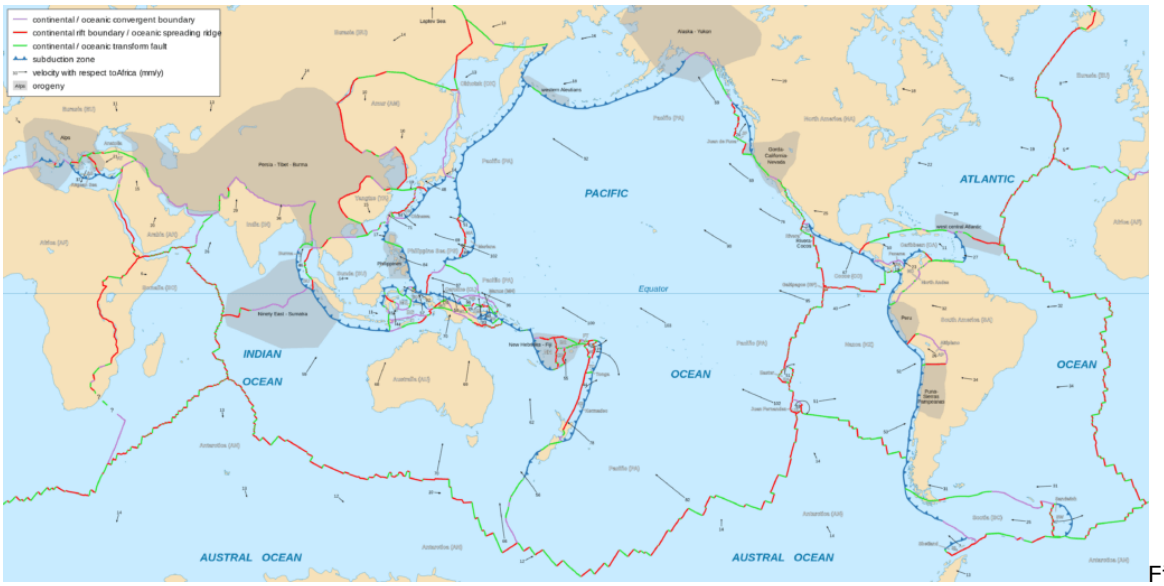


Figure 1. Detailed map showing the tectonic plates with their movement vectors. (Click on the image to open a larger version of the map.)

But without detailed evidence and a force sufficient to drive the movement, the theory was not generally accepted: the Earth might have a solid crust and mantle and a liquid core, but there seemed to be no way that portions of the crust could move around. Distinguished scientists, such as Harold Jeffreys and Charles Schuchert, were outspoken critics of continental drift.

Despite much opposition, the view of continental drift gained support and a lively debate started between "drifters" or "mobilists" (proponents of the theory) and "fixists" (opponents). During the 1920s, 1930s and 1940s, the former reached important milestones proposing that convection currents might have driven the plate movements, and that spreading may have occurred below the sea within the oceanic crust. Concepts close to the elements now incorporated in plate tectonics were proposed by geophysicists and geologists (both fixists and mobilists) like Vening-Meinesz, Holmes, and Umbgrove.

One of the first pieces of geophysical evidence that was used to support the movement of lithospheric plates came from paleomagnetism. This is based on the fact that rocks of different ages show a variable magnetic field direction, evidenced by studies since the mid-nineteenth century. The magnetic north and south poles reverse through time, and, especially important in paleotectonic studies, the relative position of the magnetic north pole varies through time. Initially, during the first half of the twentieth century, the latter phenomenon was explained by introducing what was called "polar wander" (see apparent polar wander), i.e., it was assumed that the north pole location had been shifting through time. An alternative explanation, though, was that the continents had moved (shifted and rotated) relative to the north pole, and each continent, in fact, shows its own "polar wander path". During the late 1950s it was successfully shown on two occasions that these data could show the validity of

continental drift: by Keith Runcorn in a paper in 1956,[4] and by Warren Carey in a symposium held in March 1956.[5]

The second piece of evidence in support of continental drift came during the late 1950s and early 60s from data on the bathymetry of the deep ocean floors and the nature of the oceanic crust such as magnetic properties and, more generally, with the development of marine geology which gave evidence for the association of seafloor spreading along the mid-oceanic ridges and magnetic field reversals, published between 1959 and 1963 by Heezen, Dietz, Hess, Mason, Vine & Matthews, and Morley.[6]

Simultaneous advances in early seismic imaging techniques in and around Wadati-Benioff zones along the trenches bounding many continental margins, together with many other geophysical (e.g. gravimetric) and geological observations, showed how the oceanic crust could disappear into the mantle, providing the mechanism to balance the extension of the ocean basins with shortening along its margins.

All this evidence, both from the ocean floor and from the continental margins, made it clear around 1965 that continental drift was feasible and the theory of plate tectonics, which was defined in a series of papers between 1965 and 1967, was born, with all its extraordinary explanatory and predictive power. The theory revolutionized the Earth sciences, explaining a diverse range of geological phenomena and their implications in other studies such as paleogeography and paleobiology.

Continental Drift



Figure 2. Alfred Wegener in Greenland in the winter of 1912-13.

In the late nineteenth and early twentieth centuries, geologists assumed that the Earth's major features were fixed, and that most geologic features such as basin development and mountain ranges could be explained by vertical crustal movement, described in what is called the geosynclinal theory. Generally, this was placed in the context of a contracting planet Earth due to heat loss in the course of a relatively short geological time.

It was observed as early as 1596 that the opposite coasts of the Atlantic Ocean—or, more precisely, the edges of the continental shelves—have similar shapes and seem to have once fitted together.[7]

Since that time many theories were proposed to explain this apparent complementarity, but the assumption of a solid Earth made these various proposals difficult to accept.[8]

The discovery of radioactivity and its associated heating properties in 1895 prompted a re-examination of the apparent age of the Earth.[9] This had previously been estimated by its cooling rate and assumption the Earth's surface radiated like a black body.[10] Those calculations had implied that, even if it started at red heat, the Earth would have dropped to its present temperature in a few tens of millions of years. Armed with the knowledge of a new heat source, scientists realized that the Earth would be much older, and that its core was still sufficiently hot to be liquid.

By 1915, after having published a first article in 1912,[11] Alfred Wegener was making serious arguments for the idea of continental drift in the first edition of *The Origin of Continents and Oceans*.[12] In that book (re-issued in four successive editions up to the final one in 1936), he noted how the east coast of South America and the west coast of Africa looked as if they were once attached. Wegener was not the first to note this (Abraham Ortelius, Antonio Snider-Pellegrini, Eduard Suess, Roberto Mantovani and Frank Bursley Taylor preceded him just to mention a few), but he was the first to marshal significant fossil and paleo-topographical and climatological evidence to support this simple observation (and was supported in this by researchers such as Alex du Toit). Furthermore, when the rock strata of the margins of separate continents are very similar it suggests that these rocks were formed in the same way, implying that they were joined initially. For instance, parts of Scotland and Ireland contain rocks very similar to those found in Newfoundland and New Brunswick. Furthermore, the Caledonian Mountains of Europe and parts of the Appalachian Mountainsof North America are very similar in structure and lithology.

However, his ideas were not taken seriously by many geologists, who pointed out that there was no apparent mechanism for continental drift. Specifically, they did not see how continental rock could plow through the much denser rock that makes up oceanic crust. Wegener could not explain the force that drove continental drift, and his vindication did not come until after his death in 1930.

Floating Continents, Paleomagnetism, and Seismicity Zones

As it was observed early that although granite existed on continents, seafloor seemed to be composed of denser basalt, the prevailing concept during the first half of the twentieth century was that there were two types of crust, named “sial” (continental type crust) and “sima” (oceanic type crust). Furthermore, it was supposed that a static shell of strata was present under the continents. It therefore looked apparent that a layer of basalt (sial) underlies the continental rocks.

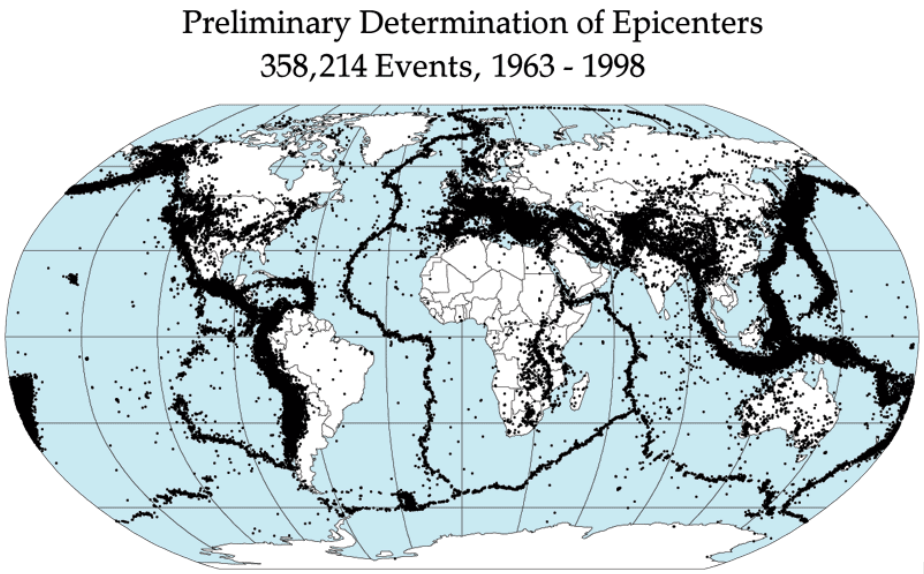


Figure 3. Global earthquake epicenters, 1963–1998

However, based on abnormalities in plumb line deflection by the Andes in Peru, Pierre Bouguer had deduced that less-dense mountains must have a downward projection into the denser layer underneath. The concept that mountains had “roots” was confirmed by George B. Airy a hundred years later, during study of Himalayan gravitation, and seismic studies detected corresponding density variations. Therefore, by the mid-1950s, the question remained unresolved as to whether mountain roots were clenched in surrounding basalt or were floating on it like an iceberg.

During the 20th century, improvements in and greater use of seismic instruments such as seismographs enabled scientists to learn that earthquakes tend to be concentrated in specific areas, most notably along the oceanic trenches and spreading ridges. By the late 1920s, seismologists were beginning to identify several prominent earthquake zones parallel to the trenches that typically were inclined 40–60° from the horizontal and extended several hundred kilometers into the Earth. These zones later became known as Wadati-Benioff zones, or simply Benioff zones, in honor of the seismologists who first recognized them, Kiyoo Wadati of Japan and Hugo Benioff of the United States. The study of global seismicity greatly advanced in the 1960s with the establishment of the Worldwide Standardized Seismograph Network (WWSSN)[13] to monitor the compliance of the 1963 treaty

banning above-ground testing of nuclear weapons. The much improved data from the WWSSN instruments allowed seismologists to map precisely the zones of earthquake concentration worldwide.

Meanwhile, debates developed around the phenomena of polar wander. Since the early debates of continental drift, scientists had discussed and used evidence that polar drift had occurred because continents seemed to have moved through different climatic zones during the past. Furthermore, paleomagnetic data had shown that the magnetic pole had also shifted during time. Reasoning in an opposite way, the continents might have shifted and rotated, while the pole remained relatively fixed. The first time the evidence of magnetic polar wander was used to support the movements of continents was in a paper by Keith Runcorn in 1956,[14] and successive papers by him and his students Ted Irving (who was actually the first to be convinced of the fact that paleomagnetism supported continental drift) and Ken Creer.

This was immediately followed by a symposium in Tasmania in March 1956.[15] In this symposium, the evidence was used in the theory of an expansion of the global crust. In this hypothesis the shifting of the continents can be simply explained by a large increase in size of the Earth since its formation. However, this was unsatisfactory because its supporters could offer no convincing mechanism to produce a significant expansion of the Earth. Certainly there is no evidence that the moon has expanded in the past 3 billion years; other work would soon show that the evidence was equally in support of continental drift on a globe with a stable radius.

During the thirties up to the late fifties, works by Vening-Meinesz, Holmes, Umbgrove, and numerous others outlined concepts that were close or nearly identical to modern plate tectonics theory. In particular, the English geologist Arthur Holmes proposed in 1920 that plate junctions might lie beneath the sea, and in 1928 that convection currents within the mantle might be the driving force.[16] Often, these contributions are forgotten because:

- At the time, continental drift was not accepted.
- Some of these ideas were discussed in the context of abandoned fixistic ideas of a deforming globe without continental drift or an expanding Earth.
- They were published during an episode of extreme political and economic instability that hampered scientific communication.
- Many were published by European scientists and at first not mentioned or given little credit in the papers on sea floor spreading published by the American researchers in the 1960s.

Mid-Oceanic Ridge Spreading and Convection

In 1947, a team of scientists led by Maurice Ewing utilizing the Woods Hole Oceanographic Institution's research vessel *Atlantis* and an array of instruments, confirmed the existence of a rise in the central Atlantic Ocean, and found that the floor of the seabed beneath the layer of sediments consisted of basalt, not the granite which is the main constituent of continents. They also found that the oceanic crust was much thinner than continental crust. All these new findings raised important and intriguing questions.[17]

The new data that had been collected on the ocean basins also showed particular characteristics regarding the bathymetry. One of the major outcomes of these datasets was that all along the globe, a system of mid-oceanic ridges was detected. An important conclusion was that along this system, new ocean floor was being created, which led to the concept of the "Great Global Rift." This was described in the crucial paper of Bruce Heezen (1960),[18] which would trigger a real revolution in thinking. A profound consequence of seafloor spreading is that new crust was, and still is, being continually created along the oceanic ridges. Therefore, Heezen advocated the so-called "expanding Earth" hypothesis of S. Warren Carey (see above). So, still the question remained: how can new crust be continuously added along the oceanic ridges without increasing the size of the Earth? In reality, this question had been solved already by numerous scientists during the forties and the fifties, like Arthur Holmes, Vening-Meinesz, Coates and many others: The crust in excess disappeared along what were called the oceanic trenches, where so-called "subduction" occurred. Therefore, when various scientists during the early sixties started to reason on the data at their disposal regarding the ocean floor, the pieces of the theory quickly fell into place.

The question particularly intrigued Harry Hammond Hess, a Princeton University geologist and a Naval Reserve Rear Admiral, and Robert S. Dietz, a scientist with the U.S. Coast and Geodetic Survey who first coined the term *seafloor spreading*. Dietz and Hess (the former published the same idea one year earlier in *Nature*,[19] but priority belongs to Hess who had already distributed an unpublished manuscript of his 1962 article by 1960)[20]were among the small handful who really understood the broad implications of sea floor spreading and how it would eventually agree with the, at that time, unconventional and unaccepted

ideas of continental drift and the elegant and mobilistic models proposed by previous workers like Holmes.

In the same year, Robert R. Coats of the U.S. Geological Survey described the main features of island arc subduction in the Aleutian Islands. His paper, though little noted (and even ridiculed) at the time, has since been called “seminal” and “prescient.” In reality, it actually shows that the work by the European scientists on island arcs and mountain belts performed and published during the 1930s up until the 1950s was applied and appreciated also in the United States.

If the Earth’s crust was expanding along the oceanic ridges, Hess and Dietz reasoned like Holmes and others before them, it must be shrinking elsewhere. Hess followed Heezen, suggesting that new oceanic crust continuously spreads away from the ridges in a conveyor belt–like motion. And, using the mobilistic concepts developed before, he correctly concluded that many millions of years later, the oceanic crust eventually descends along the continental margins where oceanic trenches—very deep, narrow canyons—are formed, e.g. along the rim of the Pacific Ocean basin. The important step Hess made was that convection currents would be the driving force in this process, arriving at the same conclusions as Holmes had decades before with the only difference that the thinning of the ocean crust was performed using Heezen’s mechanism of spreading along the ridges. Hess therefore concluded that the Atlantic Ocean was expanding while the Pacific Ocean was shrinking. As old oceanic crust is “consumed” in the trenches (like Holmes and others, he thought this was done by thickening of the continental lithosphere, not, as now understood, by underthrusting at a larger scale of the oceanic crust itself into the mantle), new magma rises and erupts along the spreading ridges to form new crust. In effect, the ocean basins are perpetually being “recycled,” with the creation of new crust and the destruction of old oceanic lithosphere occurring simultaneously. Thus, the new mobilistic concepts neatly explained why the Earth does not get bigger with sea floor spreading, why there is so little sediment accumulation on the ocean floor, and why oceanic rocks are much younger than continental rocks.

Magnetic Striping

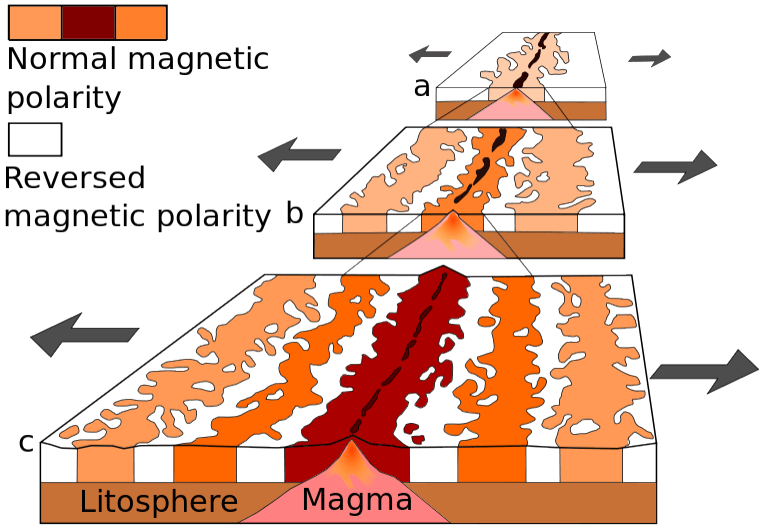
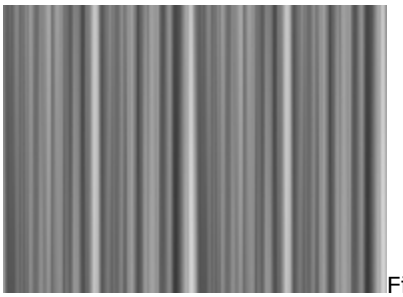


Figure 4. Seafloor magnetic striping.

Beginning in the 1950s, scientists like Victor Vacquier, using magnetic instruments (magnetometers) adapted from airborne devices developed during World War II to detect submarines, began recognizing odd magnetic variations across the ocean floor. This finding, though unexpected, was not entirely surprising because it was known that basalt—the iron-rich, volcanic rock making up the ocean floor—contains a strongly magnetic mineral (magnetite) and can locally distort compass readings. This distortion was recognized by Icelandic mariners as early as the late eighteenth century. More important, because the presence of magnetite gives the basalt measurable magnetic properties, these newly discovered magnetic variations provided another means to study the deep ocean floor. When newly formed rock cools, such magnetic materials recorded the Earth’s magnetic field at the time.


Figure 5. A demonstration of magnetic striping. (The darker the color is, the closer it is to normal polarity)

As more and more of the seafloor was mapped during the 1950s, the magnetic variations turned out not to be random or isolated occurrences, but instead revealed recognizable patterns. When these magnetic patterns were mapped over a wide region, the ocean floor showed a zebra-like pattern: one stripe with normal polarity and the adjoining stripe with reversed polarity. The overall pattern, defined by these alternating bands of normally and reversely polarized rock, became known as magnetic striping, and was published by Ron G. Mason and co-workers in 1961, who did not find, though, an explanation for these data in terms of sea floor spreading, like Vine, Matthews and Morley a few years later.[21]

The discovery of magnetic striping called for an explanation. In the early 1960s scientists such as Heezen, Hess and Dietz had begun to theorise that mid-ocean ridges mark structurally weak zones where the ocean floor was being ripped in two lengthwise along the ridge crest (see the previous paragraph). New magma from deep within the Earth rises easily through these weak zones and eventually erupts along the crest of the ridges to create new oceanic crust. This process, at first denominated the "conveyer belt hypothesis" and later called seafloor spreading, operating over many millions of years continues to form new ocean floor all across the 50,000 km-long system of mid-ocean ridges.

Only four years after the maps with the "zebra pattern" of magnetic stripes were published, the link between sea floor spreading and these patterns was correctly placed, independently byLawrence Morley, and by Fred Vine and Drummond Matthews, in 1963,[22] now called the Vine-Matthews-Morley hypothesis. This hypothesis linked these patterns to geomagnetic reversals and was supported by several lines of evidence:[23]

1. the stripes are symmetrical around the crests of the mid-ocean ridges; at or near the crest of the ridge, the rocks are very young, and they become progressively older away from the ridge crest;
2. the youngest rocks at the ridge crest always have present-day (normal) polarity;
3. stripes of rock parallel to the ridge crest alternate in magnetic polarity (normal-reversed-normal, etc.), suggesting that they were formed during different epochs documenting the (already known from independent studies) normal and reversal episodes of the Earth's magnetic field.

By explaining both the zebra-like magnetic striping and the construction of the mid-ocean ridge system, the seafloor spreading hypothesis (SFS) quickly gained converts and represented another major advance in the development of the plate-tectonics theory. Furthermore, the oceanic crust now came to be appreciated as a natural "tape recording" of the history of the geomagnetic field reversals (GMFR) of the Earth's magnetic field. Today, extensive studies are dedicated to the calibration of the normal-reversal patterns in the oceanic crust on one hand and known timescales derived from the dating of basalt layers in sedimentary sequences (magnetostratigraphy) on the other, to arrive at estimates of past spreading rates and plate reconstructions.

Definition and Refining of the Theory

After all these considerations, Plate Tectonics (or, as it was initially called "New Global Tectonics") became quickly accepted in the scientific world, and numerous papers followed that defined the concepts:

- In 1965, Tuzo Wilson who had been a promotor of the sea floor spreading hypothesis and continental drift from the very beginning[24] added the concept of transform faults to the model, completing the classes of fault types necessary to make the mobility of the plates on the globe work out.[25]
- A symposium on continental drift was held at the Royal Society of London in 1965 which must be regarded as the official start of the acceptance of plate tectonics by the scientific community, and which abstracts are issued as Blacket, Bullard & Runcorn (1965). In this symposium, Edward Bullard and co-workers showed with a computer calculation how the continents along both sides of the Atlantic would best fit to close the ocean, which became known as the famous "Bullard's Fit".

- In 1966 Wilson published the paper that referred to previous plate tectonic reconstructions, introducing the concept of what is now known as the “Wilson Cycle.”[26]
- In 1967, at the American Geophysical Union’s meeting, W. Jason Morgan proposed that the Earth’s surface consists of 12 rigid plates that move relative to each other.[27]
- Two months later, Xavier Le Pichon published a complete model based on 6 major plates with their relative motions, which marked the final acceptance by the scientific community of plate tectonics.[28]
- In the same year, McKenzie and Parker independently presented a model similar to Morgan’s using translations and rotations on a sphere to define the plate motions.[29]

---

1. Wegener, Alfred (1929). *Die Entstehung der Kontinente und Ozeane* (4 ed.). Braunschweig: Friedrich Vieweg & Sohn Akt. Ges. ↩
2. Hughes, Patrick (8 February 2001). "Alfred Wegener (1880–1930): A Geographic Jigsaw Puzzle." *On the shoulders of giants*. Earth Observatory, NASA. Retrieved 2007-12-26. ... on January 6, 1912, Wegener... proposed instead a grand vision of drifting continents and widening seas to explain the evolution of Earth's geography. ↩
3. Wegener, Alfred (1966). *The origin of continents and oceans*. Biram John (translator). Courier Dover. p. 246; Hughes, Patrick (8 February 2001). "Alfred Wegener (1880-1930): The origin of continents and oceans." On the Shoulders of Giants. Earth Observatory, NASA. Retrieved 2007-12-26. "By his third edition (1922), Wegener was citing geological evidence that some 300 million years ago all the continents had been joined in a supercontinent stretching from pole to pole. He called it Pangaea (all lands),..." ↩
4. Runcorn, S.K. (1956). "Paleomagnetic comparisons between Europe and North America". *Proceedings, Geological Association of Canada* **8** (1088): 7785. ↩
5. Carey, S. W. (1958). "The tectonic approach to continental drift." In Carey, S.W. *Continental Drift—A symposium, held in March 1956*. Hobart: Univ. of Tasmania. pp. 177–363. Expanding Earth from p. 311 to p. 349. ↩
6. Korgen, Ben J. (1995). "A voice from the past: John Lyman and the plate tectonics story" (PDF). Oceanography (The Oceanography Society) **8** (1): 19–20; Spiess, Fred; Kuperman, William (2003). "The Marine Physical Laboratory at Scripps" (PDF). *Oceanography* (The Oceanography Society) **16** (3): 45–54. ↩
7. Kious, W. Jacquelyne; Tilling, Robert I. (February 2001) [1996]. "Historical perspective." *This Dynamic Earth: the Story of Plate Tectonics* (Online ed.). U.S. Geological Survey. Retrieved 2008-01-29. "Abraham Ortelius in his work Thesaurus Geographicus... suggested that the Americas were 'torn away from Europe and Africa . . . by earthquakes and floods. . . . The vestiges of the rupture reveal themselves, if someone brings forward a map of the world and considers carefully the coasts of the three [continents].'" ↩
8. Frankel, H. (1987). "The Continental Drift Debate." In H.T. Engelhardt Jr and A.L. Caplan. *Scientific Controversies: Case Solutions in the resolution and closure of disputes in science and technology*. Cambridge University Press. ↩
9. Joly, John (1909). *Radioactivity and Geology: An Account of the Influence of Radioactive Energy on Terrestrial History*. London: Archibald Constable. p. 36. ↩
10. Thomson, W (1863). "On the secular cooling of the earth." Philosophical Magazine 4 (25): 1–14 (inactive 2015-01-09). ↩
11. Wegener, Alfred (6 January 1912). "Die Herausbildung der Grossformen der Erdrinde (Kontinente und Ozeane), auf geophysikalischer Grundlage" (PDF). *Petermanns Geographische Mitteilungen* **63**: 185–195, 253–256, 305–309. ↩
12. Wegener 1929. ↩
13. Stein, Seth; Wysession, Michael (2009). *An Introduction to Seismology, Earthquakes, and Earth Structure*. Chichester: John Wiley & Sons. ↩
14. Runcorn, 1956 ↩
15. Carey, S. W. (1958). "The tectonic approach to continental drift". In Carey, S.W. *Continental Drift—A symposium, held in March 1956*. Hobart: Univ. of Tasmania. pp. 177–363. Expanding Earth from p. 311 to p. 349; see also Quilty, Patrick G.; Banks, Maxwell R. (2003). "Samuel Warren Carey, 1911–2002." *Biographical memoirs*. Australian Academy of Science. Retrieved 2010-06-19. This memoir was originally published in *Historical Records of Australian Science* (2003) **14** (3). ↩
16. Holmes, Arthur (1928). "Radioactivity and Earth movements." *Transactions of the Geological Society of Glasgow* **18**: 559–606; see also Holmes, Arthur (1978). *Principles of Physical Geology* (3 ed.). Wiley. pp. 640–641; Frankel, Henry (July 1978). "Arthur Holmes and continental drift." *The British Journal for the History of Science* **11** (2): 130–150. doi: 10.1017/S0007087400016551. ↩
17. Lippsett, Laurence (2001). "Maurice Ewing and the Lamont-Doherty Earth Observatory." *Living Legacies*. Retrieved 2008-03-04; Lippsett, Laurence (2006). "Maurice Ewing and the Lamont-Doherty Earth Observatory". In William Theodore De Bary, Jerry Kisslinger and Tom Mathewson. *Living Legacies at Columbia*. Columbia University Press. pp. 277–297. ↩
18. Heezen, B. (1960). "The rift in the ocean floor." Scientific American 203 (4): 98–110. doi: 10.1038/scientificamerican1060-98. ↩
19. Dietz, Robert S. (June 1961). "Continent and Ocean Basin Evolution by Spreading of the Sea Floor". *Nature* **190** (4779):

854–857. ↩
20. Hess, H. H. (November 1962). "History of Ocean Basins" (PDF). In A. E. J. Engel, Harold L. James, and B. F. Leonard. *Petrologic studies: a volume to honor of A. F. Buddington*. Boulder, CO: Geological Society of America. pp. 599–620. ↩
21. Mason, Ronald G.; Raff, Arthur D. (1961). "Magnetic survey off the west coast of the United States between 32°N latitude and 42°N latitude." *Bulletin of the Geological Society of America***72** (8): 1259–1266. ↩
22. Vine, F.J.; Matthews, D.H. (1963). "Magnetic anomalies over oceanic ridges." *Nature***199** (4897): 947–949. ↩
23. See summary in Heirtzler, James R.; Le Pichon, Xavier; Baron, J. Gregory (1966). "Magnetic anomalies over the Reykjanes Ridge." *Deep-Sea Research***13** (3): 427–432. ↩
24. Wilson, J.T. (8 June 1963). "Hypothesis on the Earth's behaviour." *Nature***198** (4884): 849–865. ↩
25. Wilson, J. Tuzo (July 1965). "A new class of faults and their bearing on continental drift." *Nature***207** (4995): 343–347. ↩
26. Wilson, J. Tuzo (13 August 1966). "Did the Atlantic close and then re-open?" *Nature***211** (5050): 676–681. ↩
27. Morgan, W. Jason (1968). "Rises, Trenches, Great Faults, and Crustal Blocks." *Journal of Geophysical Research***73** (6): 1959–1982. ↩
28. Le Pichon, Xavier (15 June 1968). "Sea-floor spreading and continental drift." *Journal of Geophysical Research***73** (12): 3661–3697. ↩
29. Mc Kenzie, D.; Parker, R.L. (1967). "The North Pacific: an example of tectonics on a sphere." *Nature***216** (5122): 1276–1280. ↩

Contributors and Attributions

CC licensed content, Shared previously

- Plate tectonics. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Plate_tectonics#Development_of_the_theory. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.5.4: Developing the Theory is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.5.5: Supercontinents

Early Continents

The first crust was made of basaltic rock, like the current ocean crust. Partial melting of the lower portion of the basaltic crust began more than 4 billion years ago. This created the silica-rich crust that became the felsic continents.

Craton



Figure 1. Ice age glaciers scraped the Canadian Shield down to the 4.28 billion year old greenstone in Northwestern Quebec.

The earliest felsic continental crust is now found in the ancient cores of continents, called the **cratons**. Rapid plate motions meant that cratons experienced many continental collisions. Little is known about the **paleogeography**, or the ancient geography, of the early planet, although smaller continents could have come together and broken up.

Geologists can learn many things about the Pre-Archean by studying the rocks of the cratons.

- Cratons also contain felsic igneous rocks, which are remnants of the first continents.
- Cratonic rocks contain rounded sedimentary grains. Of what importance is this fact? Rounded grains indicate that the minerals eroded from an earlier rock type and that rivers or seas also existed.
- One common rock type in the cratons is **greenstone**, a metamorphosed volcanic rock (Figure 1). Since greenstones are found today in oceanic trenches, what does the presence of greenstones mean? These ancient greenstones indicate the presence of subduction zones.

Shield

Places the craton crops out at the surface is known as a **shield**. Cratons date from the Precambrian and are called Precambrian shields. Many Precambrian shields are about 570 million years old (Figure 2).



Figure 2. The Canadian Shield is the ancient flat part of Canada that lies around Hudson Bay, the northern parts of Minnesota, Wisconsin and Michigan and much of Greenland.

Platform

In most places the cratons were covered by younger rocks, which together are called a **platform**. Sometimes the younger rocks eroded away to expose the Precambrian craton (Figure3).



Figure 3. The Precambrian craton is exposed in the Grand Canyon where the Colorado River has cut through the younger sedimentary rocks.

Early Convection

During the Pre-Archean and Archean, Earth's interior was warmer than today. Mantle convection was faster and plate tectonics processes were more vigorous. Since subduction zones were more common, the early crustal plates were relatively small.

Since the time that it was completely molten, Earth has been cooling. Still, about half the internal heat that was generated when Earth formed remains in the planet and is the source of the heat in the core and mantle today.

The presence of water on ancient Earth is revealed in a zircon crystal:

A link to an interactive elements can be found at the bottom of this page.

Supercontinent Cycle

Look at Figure 4; is this Earth?

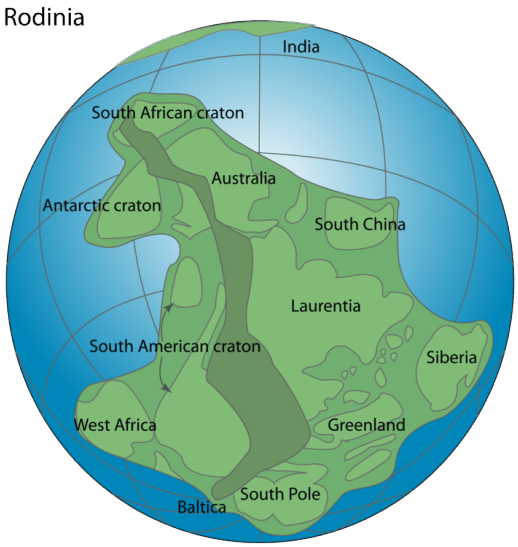


Figure 4. Rodinia

The existence of Wegener's supercontinent Pangaea is completely accepted by geologists today. The movements of continents explain so much about the geological activity we see. But did it all begin with Pangaea? Or were there other supercontinents that came before? What does the future of the continents hold?

Pangaea

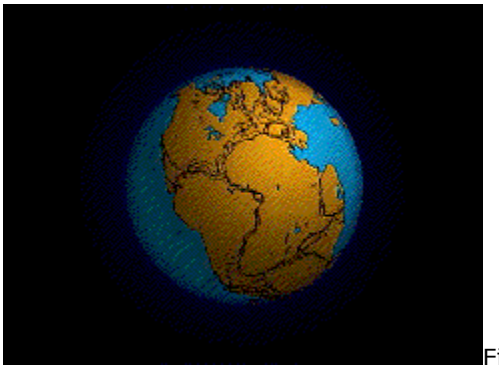


Figure 5. Pangea broke up to become our modern continents.

Wegener had lots of evidence for his continental drift hypothesis. One line of evidence was the similarity of the mountains on the west and east sides of the Atlantic. Those mountains rose at convergent plate boundaries. The continents on both sides of the ocean (where the Atlantic is now) smashed together to create Pangaea. The proto-Atlantic ocean shrank as the Pacific Ocean grew.

The Appalachian mountains of eastern North America formed at this convergent plate boundary (Figure 6a). About 200 million years ago, they were probably as high as the Himalayas.

Pangaea has been breaking apart since about 250 million years ago. Divergent plate boundaries formed within the continents to cause them to rift apart. The continents are still moving apart. The Pacific is shrinking as the Atlantic is growing. The Appalachians (Figure 6b) are now on a passive margin. The mighty mountains have weathered and eroded to what they are today.



Figure 6. (a) The Appalachian Mountains in New Hampshire. (b) The Appalachians along the eastern U.S. These mountains began when North America and Eurasia collided as Pangaea came together.

The Supercontinent Cycle

Back before Pangaea, there were earlier supercontinents. Rodinia existed 750 million to 1.1 billion years ago. Columbia existed 1.5 to 1.8 billion years ago. If the continents continue in their current directions, they will come together to create a supercontinent on the other side of the planet in around 200 million years.

This is known as the **supercontinent cycle**. The continents smash together on opposite side of the planet around every 500 million years. The creation of supercontinents is responsible for most of the geologic features that we see. It is responsible for many features that are long gone.

This animation shows the movement of continents over the past 600 million years, beginning with the breakup of Rodinia.

Summary

- The ancient core of a continent, at and beneath the surface, is its craton.
- The cratonic rock that is seen at the surface is called the shield. Where the shield is covered by younger sediments is the platform.
- Convection on early Earth was faster and so plate tectonics was faster. Since then, Earth has been cooling.
- Pangaea came together as a set of continent-continent convergent plate boundaries.
- Pangaea is still breaking up as the continents move apart. The Atlantic Ocean is getting bigger, and the Pacific Ocean is getting smaller.
- The continents come together and break apart about every 500 million years. This is called the supercontinent cycle.

Contributors and Attributions

CC licensed content, Shared previously

- Precambrian Continents. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Precambrian-Continents/lesson/Precambrian-Continents-HS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Supercontinent Cycle. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Supercontinent-Cycle/lesson/Supercontinent-Cycle-MS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*

All rights reserved content

- Earth's First Rocks. **Authored by**: National Geographic. **Located at**: https://youtu.be/V21hFmZP5zM. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

Public domain content

- Pangea Breaks Up. **Provided by**: USGS. **Located at**: http://geomaps.wr.usgs.gov/parks/pltec/pangea.html. **License**: *Public Domain: No Known Copyright*

This page titled 4.5.5: Supercontinents is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.5.6: Volcanic Arcs

All subduction zones have, at some distance in from the edge of the upper plate, arcs or chains of composite cone volcanoes. The subducting plate, as it goes down deep into the mantle, releases water. This changes the chemistry of the already hot rocks in the mantle and causes them to melt, forming magma. The magma is less dense than the solid rocks around it, so it rises upward, culminating in volcanic eruptions at the earth's surface.

The volcanic arc at an ocean-continent subduction zone is not only a chain of volcanoes. The stress of plate convergence compresses the crust there, causing it to thicken through a combination of folds and thrust faults. Igneous intrusions and volcanic eruptions also thicken the crust there. Deep within the crust, the igneous intrusions solidify into batholiths of rocks such as granite, and the pre-existing rocks that are intruded by the batholiths are regionally metamorphosed into new rocks. The result is a high mountain range with granitic and metamorphic rock at its core, folded and faulted sedimentary and volcanic around its margins, and a chain of composite cone volcanoes distributed along the crest of the range.

Contributors and Attributions

CC licensed content, Shared previously

- Basic -- Plate Tectonics. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/platetectonics.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.5.6: Volcanic Arcs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.5.7: Hot Spots

In geology, the places known as **hotspots** or **hot spots** are volcanic regions thought to be fed by underlying mantle that is anomalously hot compared with the surrounding mantle. They may be on, near to, or far from tectonic plate boundaries. Currently, there are two hypotheses that attempt to explain their origins. One suggests that they are due to hot mantle plumes that rise as thermal diapirs from the core-mantle boundary. An alternative hypothesis postulates that it is not high temperature that causes the volcanism, but lithospheric extension that permits the passive rising of melt from shallow depths. This hypothesis considers the term “hotspot” to be a misnomer, asserting that the mantle source beneath them is, in fact, not anomalously hot at all. Well known examples include Hawaii and Yellowstone.

Background

The origins of the concept of hotspots lie in the work of J. Tuzo Wilson, who postulated in 1963 that the Hawaiian Islands result from the slow movement of a tectonic plate across a hot region beneath the surface. It was later postulated that hotspots are fed by narrow streams of hot mantle rising from the Earth’score-mantle boundary in a structure called amantle plume. Whether or not such mantle plumes exist is currently the subject of a major controversy in Earth science. Estimates for the number of hotspots postulated to be fed by mantle plumes has ranged from about 20 to several thousands, over the years, with most geologists considering a few tens to exist. Hawaii, Réunion, Yellowstone, Galápagos, and Iceland are some of the currently most active volcanic regions to which the hypothesis is applied.

Most hotspot volcanoes are basaltic (e.g., Hawaii, Tahiti). As a result, they are less explosive than subduction zone volcanoes, in which water is trapped under the overriding plate. Where hotspots occur in continental regions, basaltic magma rises through the continental crust, which melts to form rhyolites. These rhyolites can form violent eruptions. For example, the Yellowstone Caldera was formed by some of the most powerful volcanic explosions in geologic history. However, when the rhyolite is completely erupted, it may be followed by eruptions of basaltic magma rising through the same lithospheric fissures (cracks in the lithosphere). An example of this activity is the Ilgachuz Range in British Columbia, which was created by an early complex series of trachyte and rhyolite eruptions, and late extrusion of a sequence of basaltic lava flows.

The hotspot hypothesis is now closely linked to the mantle plume hypothesis.

Comparison with island arc volcanoes

Hotspot volcanoes are considered to have a fundamentally different origin from island arc volcanoes. The latter form over subduction zones, at converging plate boundaries. When one oceanic plate meets another, the denser plate is forced downward into a deep ocean trench. This plate, as it is subducted, releases water into the base of the over-riding plate, and this water mixes with the rock, thus changing its composition causing some rock to melt and rise. It is this that fuels a chain of volcanoes, such as the Aleutian Islands, near Alaska.

Hotspot volcanic chains

The joint mantle plume/hotspot hypothesis envisages the feeder structures to be fixed relative to one another, with the continents and seafloor drifting overhead. The hypothesis thus predicts that time-progressive chains of volcanoes are developed on the surface. Examples are Yellowstone, which lies at the end of a chain of extinct calderas, which become progressively older to the west. Another example is the Hawaiian archipelago, where islands become progressively older and more deeply eroded to the northwest.

Geologists have tried to use hotspot volcanic chains to track the movement of the Earth’s tectonic plates. This effort has been vexed by the lack of very long chains, by the fact that many are not time-progressive (e.g. the Galápagos) and by the fact that hotspots do not appear to be fixed relative to one another (e.g., Hawaii and Iceland.)

Postulated Hotspot Volcano Chains

 Figure 2. Kilauea is the most active shield volcano in the world. The volcano has erupted nonstop since 1983 and it is part of the Hawaiian-Emperor seamount chain.

- Hawaiian-Emperor seamount chain (Hawaii hotspot)
- Louisville seamount chain (Louisville hotspot)
- Walvis Ridge (Gough and Tristan hotspot)
- Kodiak–Bowie Seamount chain (Bowie hotspot)
- Cobb-Eickelberg Seamount chain (Cobb hotspot)
- New England Seamount chain (New England hotspot)
- Anahim Volcanic Belt (Anahim hotspot)
- Mackenzie dike swarm (Mackenzie hotspot)
- Great Meteor hotspot track (New England hotspot)
- St. Helena Seamount Chain – Cameroon Volcanic Line (Saint Helena hotspot)
- Southern Mascarene Plateau–Chagos-Maldives-Laccadive Ridge (Réunion hotspot)
- Ninety East Ridge (Kerguelen hotspot)
- Tuamotu–Line Island chain (Easter hotspot)
- Austral–Gilbert–Marshall chain (Macdonald hotspot)
- Juan Fernández Ridge (Juan Fernández hotspot)

Intraplate Activity

A small amount of geologic activity, known as **intraplate activity**, does not take place at plate boundaries but within a plate instead. Mantle plumes are pipes of hot rock that rise through the mantle. The release of pressure causes melting near the surface to form a **hotspot**. Eruptions at the hotspot create a volcano. Hotspot volcanoes are found in a line (figure 3). Can you figure out why? Hint: The youngest volcano sits above the hotspot and volcanoes become older with distance from the hotspot.

Here is an animation of the creation of a hotspot chain.

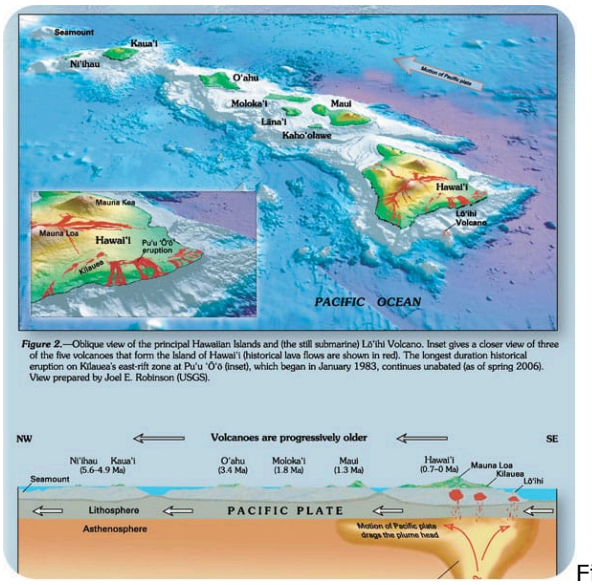


Figure 3. The Hawaiian Islands are a beautiful example of a hotspot chain. Kilauea volcano lies above the Hawaiian hotspot. Mauna Loa volcano is older than Kilauea and is still erupting, but at a lower rate. The islands get progressively older to the northwest because they are further from the hotspot. Loihi, the youngest volcano, is still below the sea surface.

Geologists use some hotspot chains to tell the direction and the speed a plate is moving (figure 4).

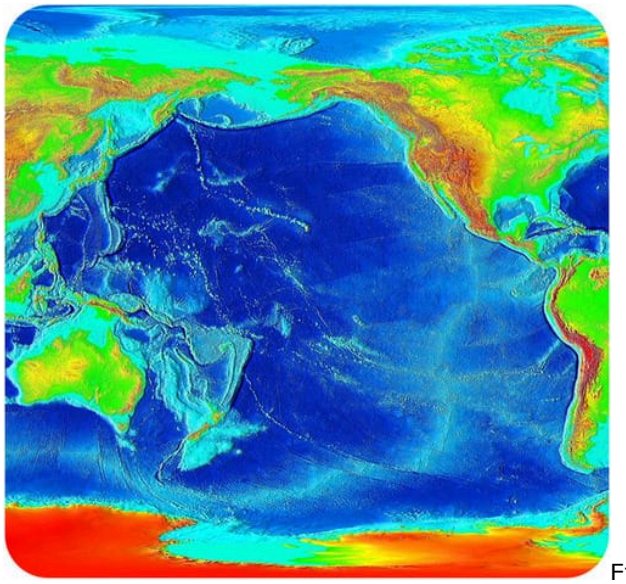


Figure 4. The Hawaiian chain continues into the Emperor Seamounts. The bend in the chain was caused by a change in the direction of the Pacific plate 43 million years ago. Using the age and distance of the bend, geologists can figure out the speed of the Pacific plate over the hotspot.

Hotspot magmas rarely penetrate through thick continental crust. One exception is the Yellowstone hotspot (figure 5).

Figure 5. Volcanic activity above the Yellowstone hotspot on the North American Plate can be traced from 15 million years ago to its present location.

Contributors and Attributions

CC licensed content, Original

- Revision and adaptation. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Hotspot (geology). **Provided by**: Wikipeida. **Located at**: http://en.Wikipedia.org/wiki/Hotspot_%28geology%29. **License**: *CC BY-SA: Attribution-ShareAlike*
- Theory of Plate Tectonics. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/6.4/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.5.7: Hot Spots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.6: EARTH'S INTERIOR

Learning Objectives

Describe the Earth including the characteristics of the various layers; learn the implications of the different layers.

- Analyze and compare the properties, material, and layers within the Earth's geosphere.
- Understand how we know about the Earth's interior and its magnetism.

4.6.1: Characteristics of the Crust

4.6.2: Characteristics of the Mantle

4.6.3: Characteristics of the Inner and Outer Core

4.6.4: The Lithosphere and Asthenosphere

4.6.5: Studying the Earth's Interior

4.6.6: Understanding the Earth's Interior

This page titled Chapter 4.6: Earth's Interior is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.6.1: Characteristics of the Crust

How does a loaf of bread resemble Earth?



A loaf of homemade bread could almost resemble Earth. The raised parts of the crust are the continents and the depressed parts are the oceans. The inside is gooier than the brittle exterior, but it's still solid. How is a loaf of bread not like Earth?

Earth's outer surface is its crust, a cold, thin, brittle outer shell made of rock. The crust is very thin relative to the radius of the planet. There are two very different types of crust, each with its own distinctive physical and chemical properties, which are summarized in Table 1.

Crust	Thickness	Density	Composition	Rock Types
Oceanic	5–12 km (3–8 mi)	3.0 g/cm^3	Mafic	Basalt and gabbro
Continental	Avg. 35 km (22 mi)	2.7 g/cm^3	Felsic	All types

Oceanic Crust



Figure 1. Gabbro from ocean crust

Oceanic crust is composed of mafic magma that erupts on the seafloor to create basalt lava flows or cools deeper down to create the intrusive igneous rock gabbro (Figure 1).

Sediments, primarily mud and the shells of tiny sea creatures, coat the seafloor. Sediment is thickest near the shore, where it comes off the continents in rivers and on wind currents.

The oceanic crust is relatively thin and lies above the mantle. The cross section of oceanic crust in Figure 2 shows the layers that grade from sediments at the top to extrusive basalt lava, to the sheeted dikes that feed lava to the surface, to deeper intrusive gabbro, and finally to the mantle.

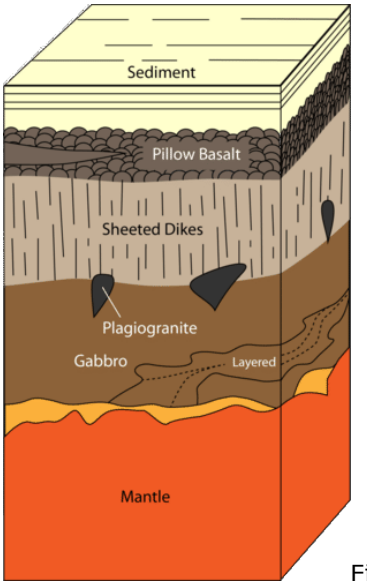


Figure 2. A cross-section of oceanic crust.

Continental Crust



Figure 3. This granite from Missouri is more than 1 billion years old.

Continental crust is made up of many different types of igneous, metamorphic, and sedimentary rocks. The average composition is granite, which is much less dense than the mafic rocks of the oceanic crust (Figure 3). Because it is thick and has relatively low density, continental crust rises higher on the mantle than oceanic crust, which sinks into the mantle to form basins. When filled with water, these basins form the planet's oceans.

Summary

- Oceanic crust is thinner and denser than continental crust.
- Oceanic crust is more mafic, continental crust is more felsic.
- Crust is very thin relative to Earth's radius.

Interactive Practice

Visit Annenberg Learner's *Dynamic Earth* and click on the crust to answer the questions below:

1. Describe the crust.
2. Where is the crust the thickest?

Contributors and Attributions

CC licensed content, Shared previously

- Earth's Crust. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Earths-Crust/lesson/Earths-Crust-HS-ES/?referrer=featured_content. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.6.1: Characteristics of the Crust is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.6.2: Characteristics of the Mantle

What is a diamond delivery system?



Some events happened when Earth was younger and hotter that do not happen any more. Kimberlite pipes shot up from deep in the mantle. These pipes are the most important source of diamonds, which form at very high pressure. Most kimberlites surfaced long ago.

The two most important things about the mantle are as follows:

1. It is made of semi-solid rock.
2. It is hot.

Solid Rock



Figure 1. Peridotite is formed of crystals of olivine (green) and pyroxene (black).

Scientists know that the mantle is made of rock based on evidence from seismic waves, heat flow, and meteorites. The properties fit the ultramafic rock peridotite, which is made of the iron- and magnesium-rich silicate minerals (Figure 1). Peridotite is rarely found at Earth's surface.

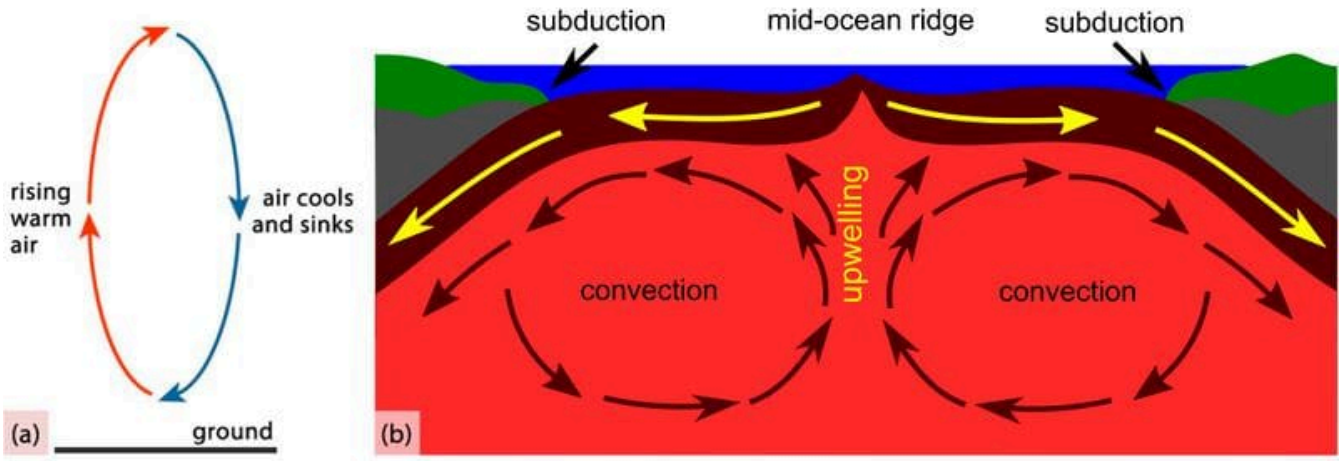
Heat Flow

Scientists know that the mantle is extremely hot because of the heat flowing outward from it and because of its physical properties.

Heat flows in two different ways within the Earth:

1. **Conduction**: Heat is transferred through rapid collisions of atoms, which can only happen if the material is solid. Heat flows from warmer to cooler places until all are the same temperature. The mantle is hot mostly because of heat conducted from the core.
2. **Convection**: If a material is able to move, even if it moves very slowly, convection currents can form.

Convection in the mantle is the same as convection in a pot of water on a stove. Convection currents within Earth's mantle form as material near the core heats up. As the core heats the bottom layer of mantle material, particles move more rapidly, decreasing its density and causing it to rise. The rising material begins the convection current. When the warm material reaches the surface, it spreads horizontally. The material cools because it is no longer near the core. It eventually becomes cool and dense enough to sink back down into the mantle. At the bottom of the mantle, the material travels horizontally and is heated by the core. It reaches the location where warm mantle material rises, and the mantle convection cell is complete (Figure 2).



In a convection cell, warm material rises and cool material sinks. In mantle convection, the heat source is the core.

Diagram of convection within Earth's mantle.

Figure 2. Convection

Contributors and Attributions

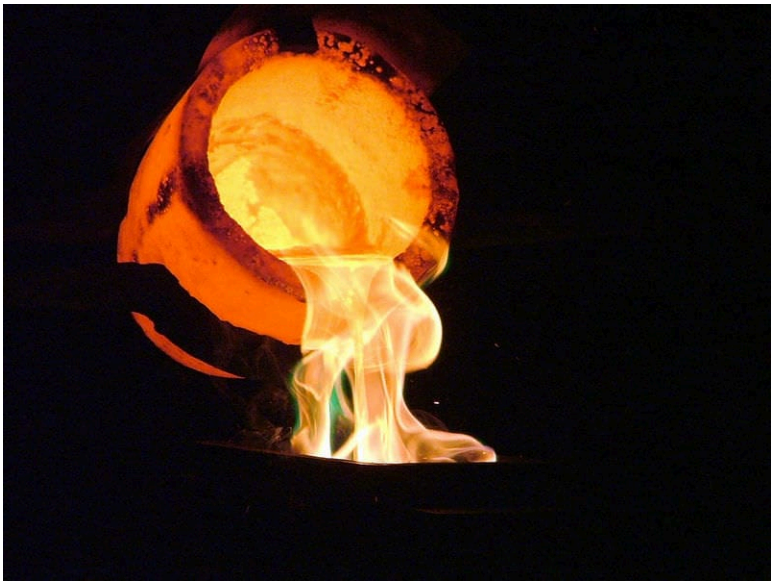
CC licensed content, Shared previously

- Earth's Mantle. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Earths-Mantle/lesson/Earths-Mantle-HS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.6.2: Characteristics of the Mantle is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.6.3: Characteristics of the Inner and Outer Core

Do you want to take a journey to the center of the earth?



Jules Verne's imagined core was fiery. But we know that the outer core is molten metal, as seen above. As hot as a journey to Verne's center of the earth might have been, a visit to the real location would be worse.



Figure 1. An iron meteorite is the closest thing to the Earth's core that we can hold in our hands.

At the planet's center lies a dense metallic core. Scientists know that the core is metal because:

1. The density of Earth's surface layers is much less than the overall density of the planet, as calculated from the planet's rotation. If the surface layers are less dense than average, then the interior must be denser than average. Calculations indicate that the core is about 85% iron metal with nickel metal making up much of the remaining 15%.
2. Metallic meteorites are thought to be representative of the core. The 85% iron/15% nickel calculation above is also seen in metallic meteorites (Figure 1).

If Earth's core were not metal, the planet would not have a magnetic field. Metals such as iron are magnetic, but rock, which makes up the mantle and crust, is not.

Scientists know that the outer core is liquid and the inner core is solid because:

1. S-waves do not go through the outer core.
2. The strong magnetic field is caused by convection in the liquid outer core. Convection currents in the outer core are due to heat from the even hotter inner core.

The heat that keeps the outer core from solidifying is produced by the breakdown of radioactive elements in the inner core.

Explore More

Use this resource to answer the questions that follow.

A link to an interactive elements can be found at the bottom of this page.

1. What materials can P-waves travel through?
2. What materials can S-waves travel through?
3. How do we know the outer core is liquid?
4. What happens to P-waves when they go through a liquid?
5. What do P-waves tell about the inner core?

Contributors and Attributions

CC licensed content, Shared previously

- Earth's Core. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Earths-Core/lesson/Earths-Core-HS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*

All rights reserved content

- INGE LEHMANN : How Inge Lehmann used Earthquakes to discover the Earthu2019s inner core. **Authored by**: Chandresh S. **Located at**: https://youtu.be/ELZ3B4flqNk. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

This page titled 4.6.3: Characteristics of the Inner and Outer Core is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.6.4: The Lithosphere and Asthenosphere

Can you think of a solid that can flow?



You use one twice a day! Toothpaste is a solid that can flow. Is the asthenosphere made of toothpaste? Only if the toothpaste is ultramafic in composition, and then it would only be able to flow if it were really, really hot. Still the toothpaste analogy gives you a good image of how the asthenosphere might behave if you squeezed it!

Lithosphere

The lithosphere is composed of both the crust and the portion of the upper mantle that behaves as a brittle, rigid solid. The lithosphere is the outermost mechanical layer, which behaves as a brittle, rigid solid. The lithosphere is about 100 kilometers thick. How are crust and lithosphere different from each other?

The definition of the lithosphere is based on how Earth materials behave, so it includes the crust and the uppermost mantle, which are both brittle. Since it is rigid and brittle, when stresses act on the lithosphere, it breaks. This is what we experience as an earthquake.

Although we sometimes refer to Earth's plates as being plates of crust, the plates are actually made of lithosphere.

Asthenosphere

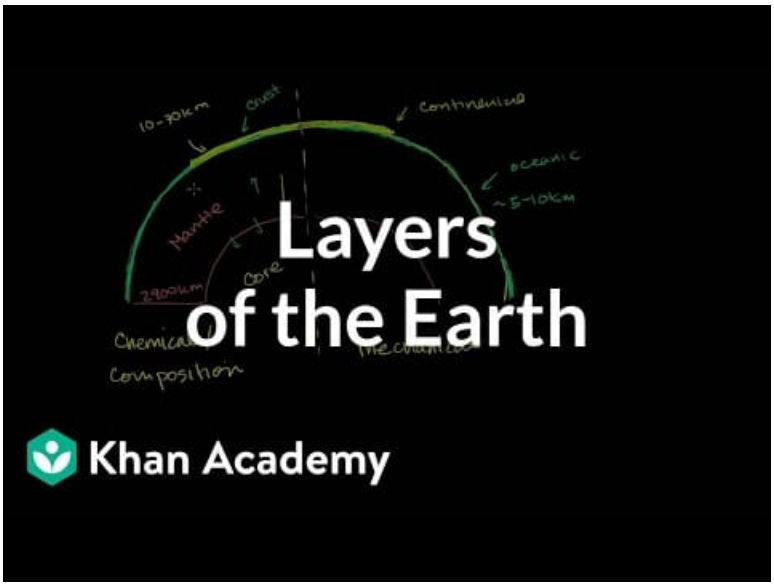
The asthenosphere is solid upper mantle material that is so hot that it behaves plastically and can flow. The lithosphere rides on the asthenosphere.

Summary

- The lithosphere is the brittle crust and uppermost mantle.
- The asthenosphere is a solid but it can flow, like toothpaste.
- The lithosphere rests on the asthenosphere.

Explore More

Watch this video to answer the questions that follow:



A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/geo/?p=230

1. What does he mean by the mechanical properties of a layer?
2. In the compositional model: What is the outermost layer? What are the two types of this layer and what are their main features?
3. What is the next layer down? What are its main features?
4. What is the deepest layer? Why is this the densest layer?
5. What is the composition of this layer?
6. What is the lithosphere?
7. What are the mechanical properties of the material below the lithosphere and what is the layer called?
8. What is the composition and mechanical property of the mesosphere relative to the asthenosphere?

Contributors and Attributions

CC licensed content, Shared previously

- Lithosphere and Asthenosphere. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Lithosphere-and-Asthenosphere/lesson/Lithosphere-and-Asthenosphere-HS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Compositional and mechanical layers of the earth. **Authored by**: Sal Khan. **Provided by**: Khan Academy. **Located at**: https://www.khanacademy.org/science/cosmology-and-astronomy/earth-history-topic/plate-techtonics/v/compositional-and-mechanical-layers-of-the-earth. **License**: *CC BY-NC-SA: Attribution-NonCommercial-ShareAlike*

This page titled 4.6.4: The Lithosphere and Asthenosphere is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.6.5: Studying the Earth's Interior

The following table summarizes the physical layers of the earth.

Physical Layers of Earth		
Layer	**Physical Behavior**	**Thickness**
Lithosphere	rigid, brittle at shallow depths	5–200 km
Asthenosphere	ductile	100–300 km
Upper Mesosphere	rigid, not brittle, rapid increase in density with depth	300–400 km
Lower Mesosphere	denser and more rigid than upper mesosphere	2,300 km
Outer Core	liquid	2,300 km
Inner Core	rigid, not brittle	1,200 km

Earth's Magnetic Field Originates in the Core

The liquid outer core is the source of the earth's magnetic field, as a result of its metallic nature, which means it contains electrons not attached to particular nuclei. Heat is transferred upward to the mantle from the inner core via convective cells, in which the liquid in the outer core flows in looping patterns. The combination of the loose electrons and looping convective flow with the rotation of the earth results in a geodynamo that produces a magnetic field. Because the magnetic field is generated by a dynamically convecting and rotating sphere of liquid, it is unstable. Every now and then, after several hundred thousand to several million years, the earth's magnetic field becomes unstable to the point that it temporarily shuts down. When it restarts, its north and south magnetic poles must inevitably be reversed, according to the physics of magnetic fields produced spontaneously from geodyamos. (For comparison, the magnetic field of the Sun, which is also produces by convecting electrical charges in a rotating sphere, becomes magnetically unstable and reverses its magnetic field on a more regular basis, every 11 years.)

Given that the inner core is a solid metallic sphere, made mostly of iron and nickel, surrounded entirely by liquid, it can be pictured as a giant ball bearing spinning in a pressurized fluid. Detailed studies of earthquake waves passing though the inner core have found evidence that it is spinning – rotating – just slightly faster than the rest of the earth.

Beyond Simple Layers

The interior of the earth is not simply layered. Some of the layers, particularly the crust and lithosphere, are highly variable in thickness. The boundaries between layers are rough and irregular. Some layers penetrate other layers at certain places. Variations in the thickness of the earth's layers, irregularities in layer boundaries, and interpenetrations of layers, reflect the dynamic nature of the earth.

For example, the lithosphere penetrates deep into the mesosphere at subduction zones. Although it is still a matter of research and debate, there is some evidence that subducted plates may penetrate all the way into the lower mesosphere. If so, plate tectonics is causing extensive mixing and exchange of matter in the earth, from the bottom of the mantle to the top of the crust.

As another example, hot spots may be places where gases and fluids rise from the core-mantle boundary, along with heat. Studies of helium isotopes in hot spot volcanic rocks find evidence that much of the helium comes from deep in the earth, probably from the lower mesosphere.

How Do We Know?

We humans have no hands-on access to samples of the earth's interior from deeper than the upper mantle. The earth's core is so dense and so deep, it is completely inaccessible. Contrary to a popular misconception, lava does not come from the earth's core. Magma and lava come from only the lithosphere and asthenosphere, the upper 200 km of earth's 6,400 km thickness. Attempts

have been made to drill through the crust to reach the mantle, without success. Given the lack of actual pieces of the earth from deeper than the asthenosphere, how do we know about the internal layers of the earth, what they are made of, and what their properties and processes are?

Igneous Rocks and Fault Blocks

There are two sources of rock samples from the lower lithosphere and asthenosphere, igneous rocks and fault blocks. Some igneous rocks contain xenoliths, pieces of solid rock that were adjacent to the body of magma, became incorporated into the magma, and were carried upward in the magma. From xenoliths in plutonic and volcanic igneous rocks, many samples of the lower crust and upper mantle have been identified and studied.

Another source of pieces of the lower crust and upper mantle is fault zones and exposed orogenic zones (root zones of mountains that have been exposed after much uplift and erosion). Some slabs of thrust-faulted rock contain lithospheric mantle rock. In ophiolites, ultramafic rock from the mantle part of the lithosphere is a defining attribute. Most ophiolites and thrust-faulted slices of rock that contain pieces of the upper mantle are related to either subduction zones or transform plate boundaries.

Seismic Waves

The energy from earthquakes travels in waves. The study of seismic waves is known as seismology. Seismologists use seismic waves to learn about earthquakes and also to learn about the Earth's interior.

One ingenious way scientists learn about Earth's interior is by looking at earthquake waves. Seismic waves travel outward in all directions from where the ground breaks and are picked up by seismographs around the world. Two types of seismic waves are most useful for learning about Earth's interior.

Body Waves

P-waves and S-waves are known as body waves because they move through the solid body of the Earth. P-waves travel through solids, liquids, and gases. S-waves only move through solids (Figure 1). Surface waves only travel along Earth's surface. In an earthquake, body waves produce sharp jolts. They do not do as much damage as surface waves.

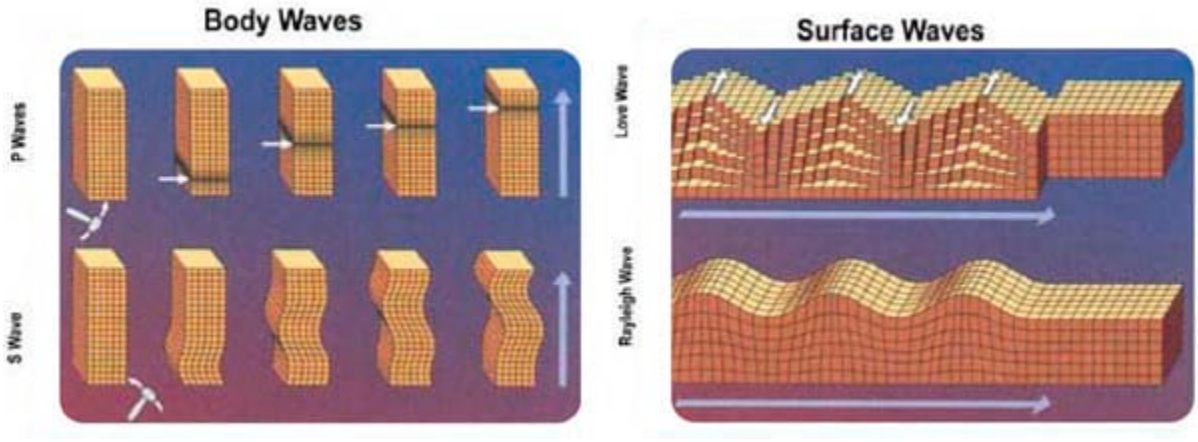
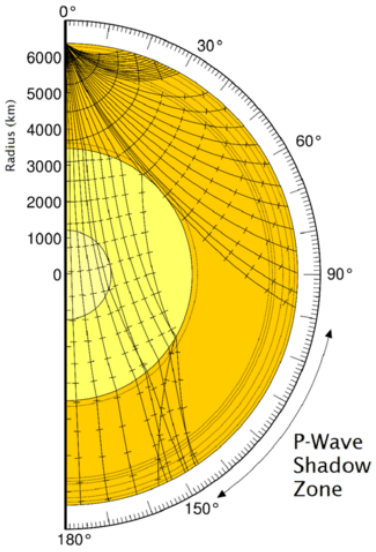


Figure 1. Body



and Surface Waves

Figure 2. How P-waves travel through Earth's interior.

- **P-waves** (primary waves) are fastest, traveling at about 6 to 7 kilometers (about 4 miles) per second, so they arrive first at the seismometer. P-waves move in a compression/expansion type motion, squeezing and unsqueezing Earth materials as they travel. This produces a change in volume for the material. P-waves bend slightly when they travel from one layer into another. Seismic waves move faster through denser or more rigid material. As P-waves encounter the liquid outer core, which is less rigid than the mantle, they slow down. This makes the P-waves arrive later and further away than would be expected. The result is a P-wave shadow zone. No P-waves are picked up at seismographs 104° to 140° from the earthquakes focus.
- **S-waves** (secondary waves) are about half as fast as P-waves, traveling at about 3.5 km (2 miles) per second, and arrive second at seismographs. S-waves move in an up and down motion perpendicular to the direction of wave travel. This produces a change in shape for the Earth materials they move through. Only solids resist a change in shape, so S-waves are only able to propagate through solids. S-waves cannot travel through liquid.

Where seismic waves speed up or slow down, they refract, changing the direction in which they are traveling. Where seismic waves encounter an abrupt boundary between two very different layers, some of the seismic wave energy is reflected, bouncing back at the same angle it struck. The reflections and refractions of seismic waves allow the layers and boundaries within the earth to be located and studied.

By tracking seismic waves, scientists have learned what makes up the planet's interior (figure 2).

- P-waves slow down at the mantle core boundary, so we know the outer core is less rigid than the mantle.
- S-waves disappear at the mantle core boundary, so the outer core is liquid.

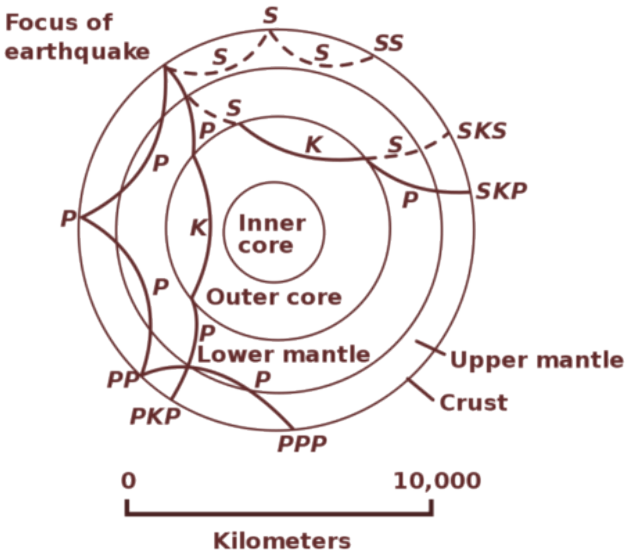


Figure 3. Letters describe the path of an individual P-wave or S-wave. Waves traveling through the core take on the letter K.

This animation shows a seismic wave shadow zone.

Here are some examples of what we have been able to distinguish in the earth's interior from the study of seismic waves and how they travel through the layers of the earth:

1. The thickness of the crust. This is a measure of the thickness of the crust based on the abrupt increase in speed of seismic waves that occurs when they enter the mantle. The boundary between the crust and mantle, as inferred from the change in the speed of P- and S-waves, is called the Mohorovicic discontinuity, named after the Croatian seismologist who first discerned it; usually it is referred to simply as the Moho. It is mainly from seismic waves that we know how thin oceanic crust is and how thick continental crust is.
2. The thickness of the lithosphere. Where seismic waves pass down from the lithosphere into the asthenosphere, they slow down. This is because of the lower rigidity and compressibility of the rocks in the layer below the lithosphere. The zone below the lithosphere where seismic waves travel more slowly is called the low velocity zone. The low velocity zone is probably coincident with the asthenosphere.
3. The boundary between the upper and lower mesosphere (upper and lower mantle). This shows up as an increase in seismic wave speed at a depth of 660 km.
4. The boundary between the mantle and the core. This is marked by S-waves coming to an abrupt stop, presumably because the outer core is liquid, and a sudden large reduction in the speed of P-waves, as they enter the liquid core where there is no rigidity to contribute to P-wave speed.
5. The inner core. This was first recognized by refraction of P-waves passing through this part of the core, due to an abrupt increase in their speed, which was not shown by P-waves traveling through only the outer part of the core.
6. Seismic tomography: imaging slabs and masses at various orientations in the earth, not just in layers. By combining data from many seismometers, three-dimensional images of zones in the earth that have higher or lower seismic wave speeds can be constructed. Seismic tomography shows that in some places there are masses of what may be subducted plates that have penetrated below the asthenosphere into the mesosphere and, in some cases, penetrated into the lower mesosphere, the deepest part of the mantle. In other places, subducted plates appear to have piled up at the base of the upper mesosphere without penetrating into the lower mesosphere.

Gravity

Isaac Newton was the first to calculate the total mass of the earth. This gives us an important constraint on what the earth is made of, because, by dividing the mass of the earth by the volume of the earth, we know the average density of the earth. Whatever the earth is made of, it must add up to the correct amount of mass. Gravity measurements, and the earth's mass, tell us that the interior of the earth must be denser than the crust, because the average density of earth is much higher than the density of the crust.

Because different parts of the crust, mantle, and core have different thicknesses and densities, the strength of gravity over

particular points on earth varies slightly. These variations from the average strength of earth's gravity are called gravity anomalies. Mapping and analyzing gravity anomalies, in some cases by using satellites, and also be measuring the effect of gravity anomalies on the surface shape of the ocean, has given us much insight into subduction zones, mid-ocean spreading ridges, and mountain ranges, including constraints on the depths of their roots.

Moment of Inertia

The earth's gravity tells us how much total mass the earth has, but does not tell us how the mass is distributed within the earth. A property known as moment of inertia, which is the resistance (inertia) of an object to changes in its spin (rotation), is determined by exactly how matter is distributed in a spinning object, from its core to its surface. The earth's moment of inertia is measured by its effect on other objects with which it interacts gravitationally, including the Moon, and satellites. Knowing the earth's moment of inertia provides a way of checking and refining our understanding of the mass and density of each of the earth's internal layers.

Meteorites

Studies of meteorites, which are pieces of asteroids that have landed on earth, along with astronomical studies of what the Sun, the other planets, and orbiting asteroids are made of, give us a model for the general chemical composition of objects in the inner solar system, which are made mainly of elements that form rocks and metals, as opposed to the outer planets such as Jupiter, which are made mostly of light, gas-forming elements. The general compositional model of the rocky and metallic part of the solar system has much higher percentages of iron, nickel, and magnesium than is found in the earth's crust.

If the earth's mantle is made of ultramafic rock, as is found in actual samples of the upper mantle in xenoliths and ophiolites, that would account for part of the missing iron, nickel, and magnesium. But much more iron and nickel would still be missing. If the core is made mostly of iron, and abundant nickel as well, it would give the earth an overall composition similar to the composition of other objects in the inner solar system, and similar to the proportions of rock and metal-forming elements measured in the Sun.

A mantle with an ultramafic composition, and a core made mostly of iron plus nickel, would make earth's composition match the composition of the rest of the solar system, and give those layers the right densities to account for the earth's moment of inertia and total mass.

Experiments

Geology, like other sciences, is based on experiment along with observation and theory. earth scientists and physicists have developed experimental methods to study how materials behave at the pressures and temperatures of the earth's interior, including core temperatures and pressures. They can measure such properties as the density, the state of matter (liquid or solid), the rigidity, the compressibility, and the speed at which seismic waves pass through these materials at high pressures and temperatures. These studies allow further refinement of our knowledge of what the interior of the earth is made of and how it behaves. These experiments support the theory that the mantle is ultramafic and the core is mostly iron and nickel, because they show that materials with those compositions have the same density and seismic wave speeds as have been observed in the earth.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Earth's Interior. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/earthinterior.html#magnetic. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*
- Seismic Waves. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Seismic-Waves/lesson/Seismic-Waves-HS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.6.5: Studying the Earth's Interior is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.6.6: Understanding the Earth's Interior

Check Your Understanding

Answer the question(s) below to see how well you understand the topics covered in the previous section. This short quiz does **not** count toward your grade in the class, and you can retake it an unlimited number of times.

Use this quiz to check your understanding and decide whether to (1) study the previous section further or (2) move on to the next section.

https://assessments.lumenlearning.co...sessments/1099

Contributors and Attributions

CC licensed content, Original

- **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 4.6.6: Understanding the Earth's Interior is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.7: CRUSTAL DEFORMATION

Learning Objectives

Correlate structural geology as it pertains to crustal deformation.

- Compare and contrast stress versus strain in the Earth's crust.
- Identify geologic structures created by deformation.

4.7.1: Stress In Earth's Crust

4.7.2: Strain

4.7.3: Folds

4.7.4: Faults

4.7.5: Mountains

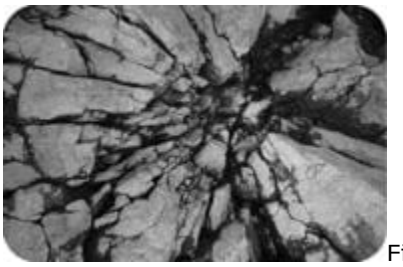
This page titled Chapter 4.7: Crustal Deformation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.7.1: Stress In Earth's Crust

Introduction

Enormous slabs of lithosphere move unevenly over the planet's spherical surface, resulting in earthquakes. This chapter deals with two types of geological activity that occur because of plate tectonics: mountain building and earthquakes. First, we will consider what can happen to rocks when they are exposed to stress.

Causes and Types of Stress


Figure 1. Stress caused these rocks to fracture.

Stress is the force applied to an object. In geology, stress is the force per unit area that is placed on a rock. Four types of stresses act on materials.

- A deeply buried rock is pushed down by the weight of all the material above it. Since the rock cannot move, it cannot deform. This is called **confining stress.**
- **Compression** squeezes rocks together, causing rocks to fold or fracture (break) (figure 1). Compression is the most common stress at convergent plate boundaries.
- Rocks that are pulled apart are under **tension**. Rocks under tension lengthen or break apart. Tension is the major type of stress at divergent plate boundaries.
- When forces are parallel but moving in opposite directions, the stress is called **shear** (figure 2). Shear stress is the most common stress at transform plate boundaries.


Figure 2. Shearing in rocks. The white quartz vein has been elongated by shear.

When stress causes a material to change shape, it has undergone **strain** or**deformation.** Deformed rocks are common in geologically active areas.

A rock's response to stress depends on the rock type, the surrounding temperature, and pressure conditions the rock is under, the length of time the rock is under stress, and the type of stress.

Rocks have three possible responses to increasing stress (illustrated in figure 3):

- **elastic deformation**: the rock returns to its original shape when the stress is removed.
- **plastic deformation**: the rock does not return to its original shape when the stress is removed.

- **fracture**: the rock breaks.

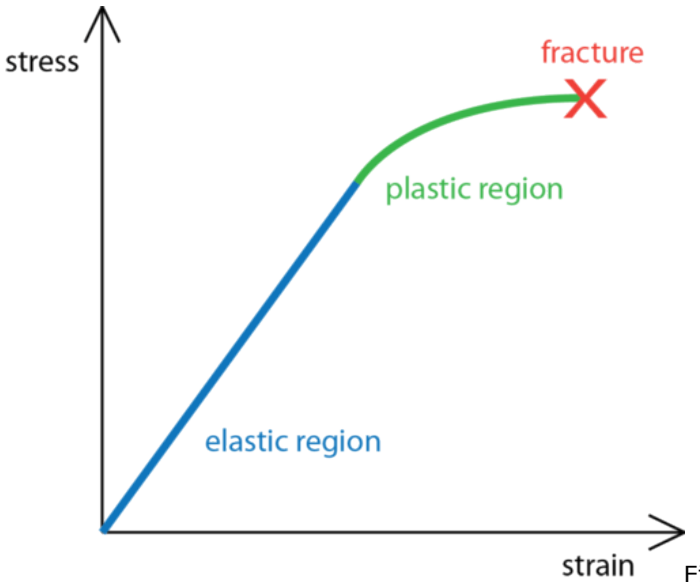


Figure 3. With increasing stress, the rock undergoes: (1) elastic deformation, (2) plastic deformation, and (3) fracture.

Under what conditions do you think a rock is more likely to fracture? Is it more likely to break deep within Earth's crust or at the surface? What if the stress applied is sharp rather than gradual?

- At the Earth's surface, rocks usually break quite quickly, but deeper in the crust, where temperatures and pressures are higher, rocks are more likely to deform plastically.
- Sudden stress, such as a hit with a hammer, is more likely to make a rock break. Stress applied over time often leads to plastic deformation.

Geologic Structures

Sedimentary rocks are important for deciphering the geologic history of a region because they follow certain rules.

1. Sedimentary rocks are formed with the oldest layers on the bottom and the youngest on top.
2. Sediments are deposited horizontally, so sedimentary rock layers are originally horizontal, as are some volcanic rocks, such as ash falls.
3. Sedimentary rock layers that are not horizontal are deformed.

You can trace the deformation a rock has experienced by seeing how it differs from its original horizontal, oldest-on-bottom position (figure 4a). This deformation produces geologic structures such as folds, joints, and faults that are caused by stresses (figure 4b). Using the rules listed above, try to figure out the geologic history of the geologic column below.

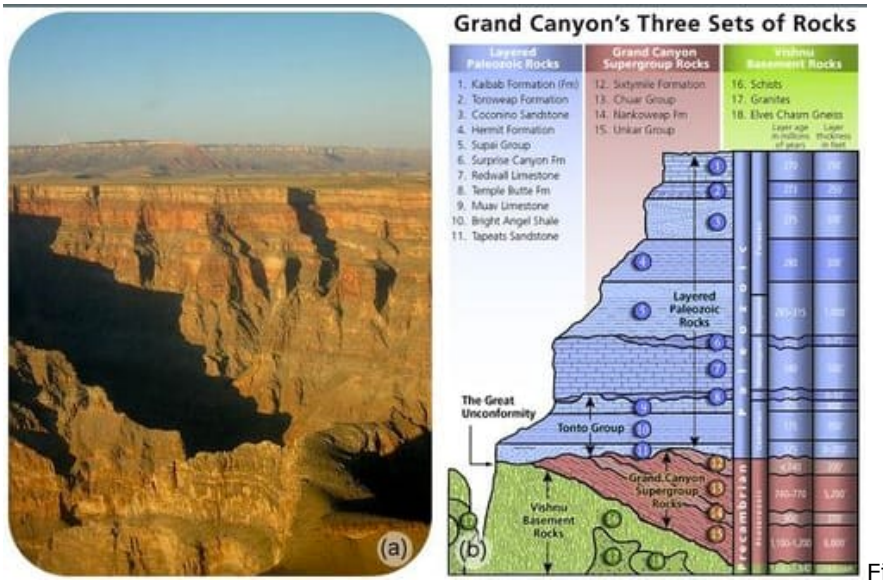


Figure 4. (a) In the Grand Canyon, the rock layers are exposed like a layer cake. Each layer is made of sediments that were deposited in a particular environment – perhaps a lake bed, shallow offshore region, or a sand dune. (b) In this geologic column of the Grand Canyon, the sedimentary rocks of the "Layered Paleozoic Rocks" column (layers 1 through 11) are still horizontal. Grand Canyon Supergroup rocks (layers 12 through 15) have been tilted. Vishnu Basement Rocks are not sedimentary (rocks 16 through 18). The oldest layers are on the bottom and youngest are on the top.

Folds

Rocks deforming plastically under compressive stresses crumple into **folds** (figure 5). They do not return to their original shape. If the rocks experience more stress, they may undergo more folding or even fracture.


Figure 5. Snow accentuates the fold exposed in these rocks in Provo Canyon, Utah.

Three types of folds are seen.

- Mononcline: A **monocline** is a simple bend in the rock layers so that they are no longer horizontal (see figure 6 for



Figure 6. At Colorado National Monument, the rocks in a monocline plunge toward the ground.

an example).

- Anticline: An **anticline** is a fold that arches upward. The rocks dip away from the center of the fold (figure 7). The oldest rocks are at the center of an anticline and the youngest are draped over them.

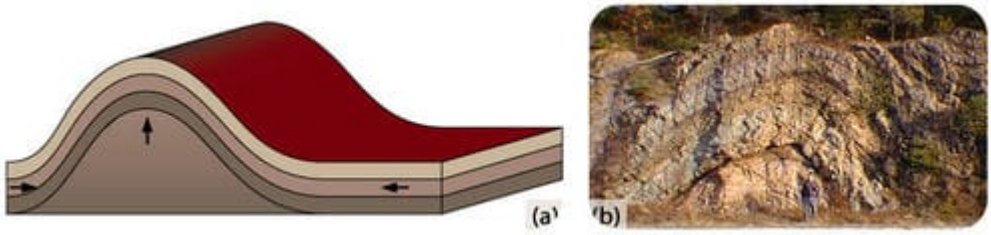


Figure 7. (a) Schematic of an anticline. (b) An anticline exposed in a road cut in New Jersey.

When rocks arch upward to form a circular structure, that structure is called a **dome.** If the top of the dome is sliced off, where are the oldest rocks located?

- Syncline: A **syncline** is a fold that bends downward. The youngest rocks are at the center and the oldest are at the outside (figure 8).

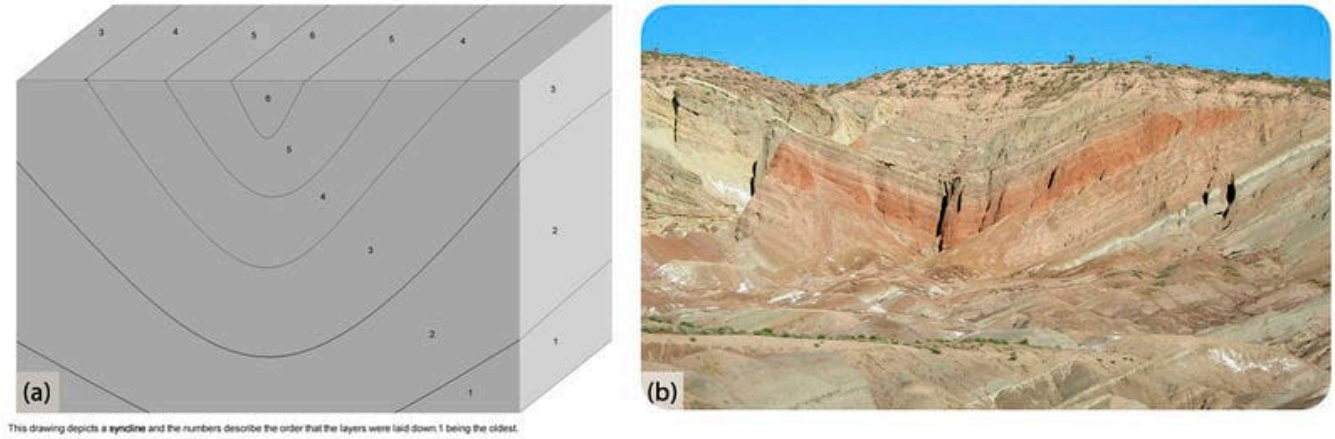


Figure 8. (a) Schematic of a syncline. (b) This syncline is in Rainbow Basin, California.

When rocks bend downward in a circular structure, that structure is called a **basin** (figure 9). If the rocks are exposed at the surface, where are the oldest rocks located?



Figure 9. Basins can be enormous. This is a geologic map of the Michigan Basin, which is centered in the state of Michigan but extends into four other states and a Canadian province.

Faults

A rock under enough stress will fracture. If there is no movement on either side of a fracture, the fracture is called a **joint**, as shown in (figure 10).

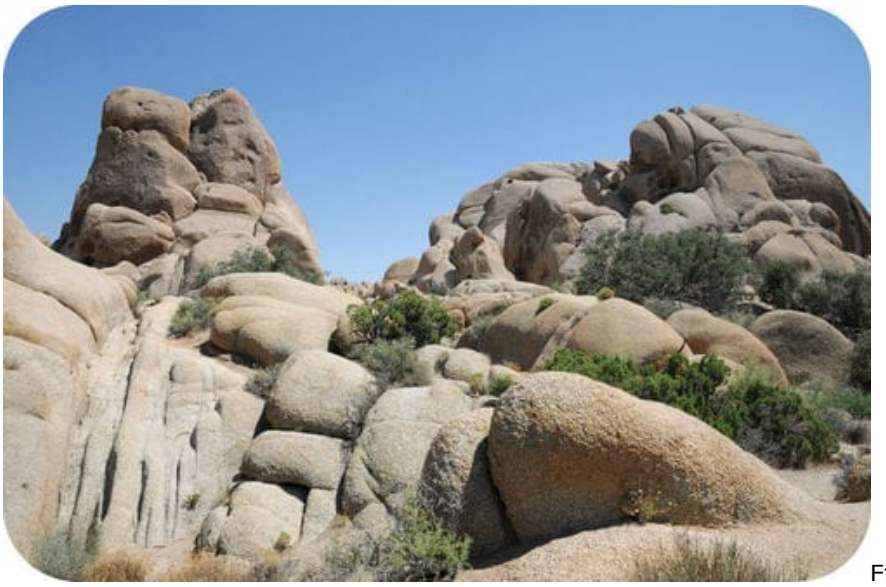


Figure 10. Granite rocks in Joshua Tree National Park showing horizontal and vertical jointing. These joints formed when the confining stress was removed from the granite.

If the blocks of rock on one or both sides of a fracture move, the fracture is called a **fault** (figure 11). Sudden motions along faults cause rocks to break and move suddenly. The energy released is an earthquake.



Figure 11. Faults are easy to recognize as they cut across bedded rocks.

Slip is the distance rocks move along a fault. Slip can be up or down the fault plane. Slip is relative, because there is usually no way to know whether both sides moved or only one. Faults lie at an angle to the horizontal surface of the Earth. That angle is called the fault's **dip.** The dip defines which of two basic types a fault is. If the fault's dip is inclined relative to the horizontal, the fault is a **dip-slip fault** (figure 12). There are two types of dip-slip faults. In **normal faults**, the hanging wall drops down relative to the footwall. In **reverse faults**, the footwall drops down relative to the hanging wall.

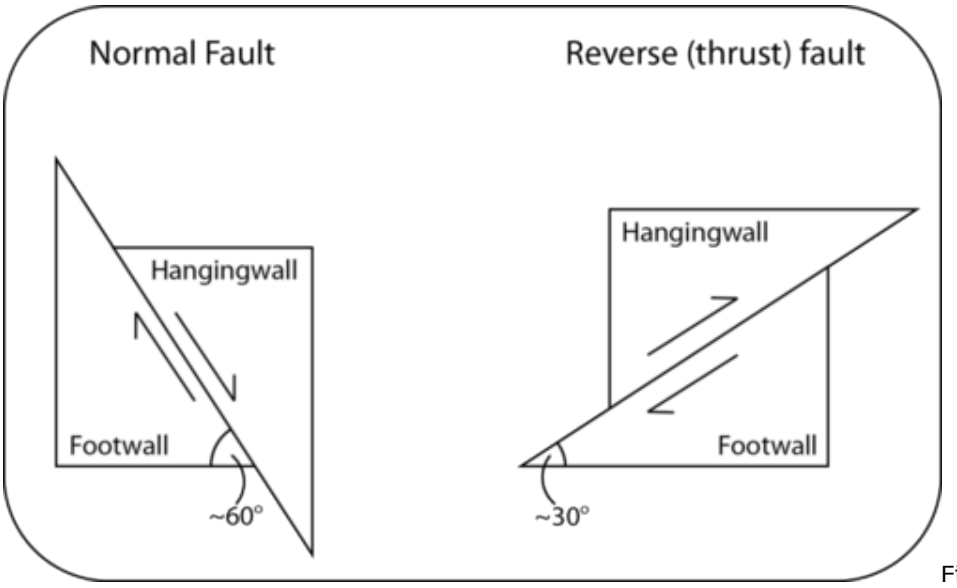


Figure 12. This diagram illustrates the two types of dip-slip faults: normal faults and reverse faults. Imagine miners extracting a resource along a fault. The hanging wall is where miners would have hung their lanterns. The footwall is where they would have walked.

Here is an animation of a normal fault.

A **thrust fault** is a type of reverse fault in which the fault plane angle is nearly horizontal. Rocks can slip many miles along thrust faults (**Figure** 13).



Figure 13. At Chief Mountain in Montana, the upper rocks at the Lewis Overthrust are more than 1 billion years older than the lower rocks. How could this happen?

Here is an animation of a thrust fault.

Normal faults can be huge. They are responsible for uplifting mountain ranges in regions experiencing tensional stress (figure 14).



Figure 14. The Teton Range in Wyoming rose up along a normal fault.

A **strike-slip fault** is a dip-slip fault in which the dip of the fault plane is vertical. Strike-slip faults result from shear stresses (figure 15).

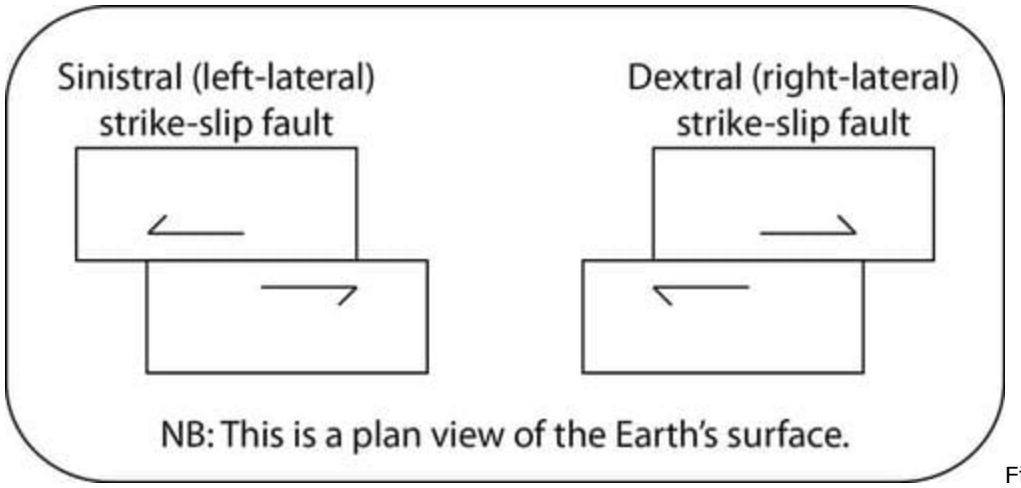


Figure 15. Imagine placing one foot on either side of a strike-slip fault. One block moves toward you. If that block moves toward your right foot, the fault is a right-lateral strike-slip fault; if that block moves toward your left foot, the fault is a left-lateral strike-slip fault.

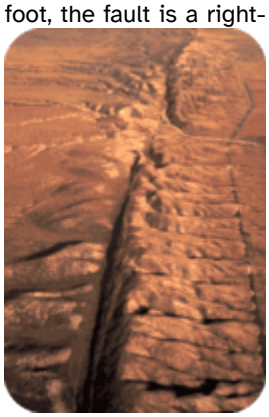


Figure 16. The San Andreas is a massive transform fault.

California's San Andreas Fault is the world's most famous strike-slip fault. It is a right-lateral strike slip fault (figure 16).

Here is a strike-slip fault animation.

People sometimes say that California will fall into the ocean someday, which is not true. This animation shows movement on the San Andreas into the future.

Stress and Mountain Building

Two converging continental plates smash upwards to create mountain ranges (figure 17). Stresses from this **uplift** cause folds, reverse faults, and thrust faults, which allow the crust to rise upwards.

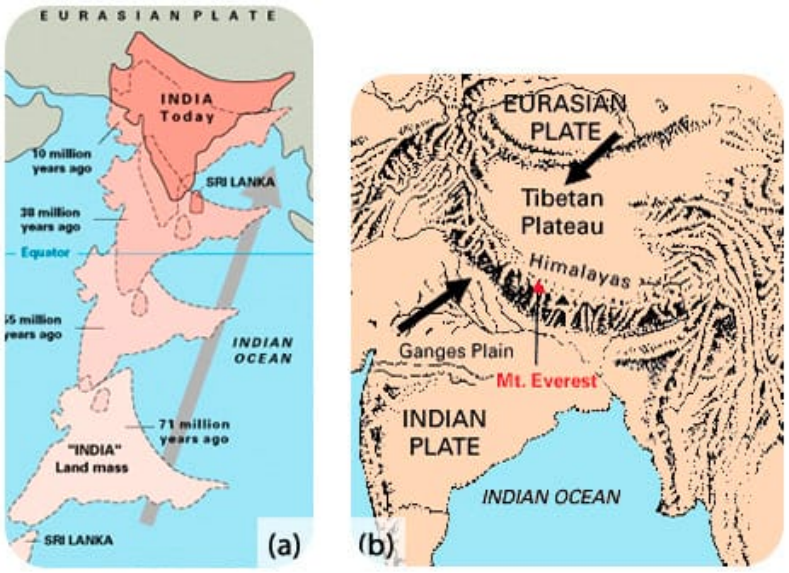


Figure 17. (a) The world's highest mountain range, the Himalayas, is growing from the collision between the Indian and the Eurasian plates. (b) The crumpling of the Indian and Eurasian plates of continental crust creates the Himalayas.

Subduction of oceanic lithosphere at convergent plate boundaries also builds mountain ranges (figure 18).



Figure 18. The Andes Mountains are a chain of continental arc volcanoes that build up as the Nazca Plate subducts beneath the South American Plate.

When tensional stresses pull crust apart, it breaks into blocks that slide up and drop down along normal faults. The result is alternating mountains and valleys, known as a basin-and-range (figure 19).

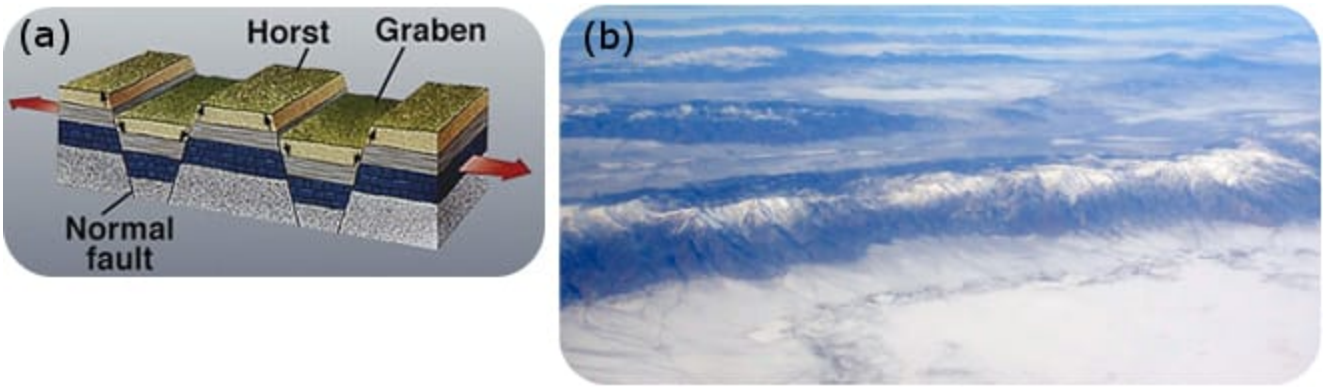


Figure 19. (a) In basin-and-range, some blocks are uplifted to form ranges, known as horsts, and some are down-dropped to form basins, known as grabens. (b) Mountains in Nevada are of classic basin-and-range form.

This is a very quick animation of movement of blocks in a basin-and-range setting.

Lesson Summary

- Stress is the force applied to a rock and may cause deformation. The three main types of stress are typical of the three types of plate boundaries: compression at convergent boundaries, tension at divergent boundaries, and shear at transform boundaries.
- Where rocks deform plastically, they tend to fold. Brittle deformation brings about fractures and faults.
- The two main types of faults are dip-slip (the fault plane is inclined to the horizontal) and strike-slip (the fault plane is perpendicular to the horizontal).
- The world's largest mountains grow at convergent plate boundaries, primarily by thrust faulting and folding.

Contributors and Attributions

CC licensed content, Shared previously

- 7.1: Stress in Earthu2019s Crust. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/7.1/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.7.1: Stress In Earth's Crust is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.7.2: Strain

As we've just learned, the earth's crust is constantly subjected to forces that push, pull, or twist it. These forces are called stress. In response to stress, the rocks of the earth undergo **strain**, also known as deformation.

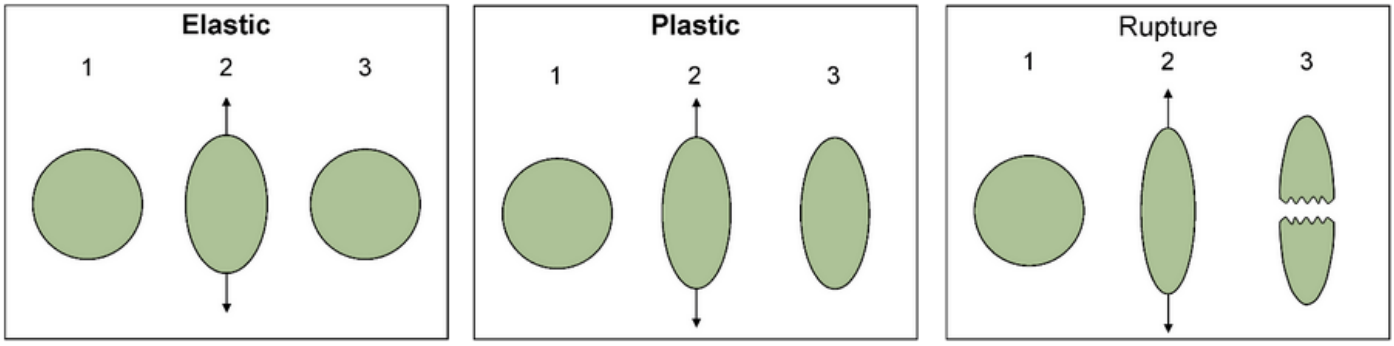
Strain is any change in volume or shape.There are four general types of stress. One type of stress is uniform, which means the force applies equally on all sides of a body of rock. The other three types of stress, tension, compression and shear, are non-uniform, or directed, stresses.All rocks in the earth experience a uniform stress at all times. This uniform stress is called lithostatic pressure and it comes from the weight of rock above a given point in the earth. Lithostatic pressure is also called hydrostatic pressure. (Included in lithostatic pressure are the weight of the atmosphere and, if beneath an ocean or lake, the weight of the column of water above that point in the earth. However, compared to the pressure caused by the weight of rocks above, the amount of pressure due to the weight of water and air above a rock is negligible, except at the earth's surface.) The only way for lithostatic pressure on a rock to change is for the rock's depth within the earth to change.Because lithostatic pressure is a uniform stress, a change in lithostatic pressure does not cause fracturing and slippage along faults. Nevertheless, it may be the cause of certain types of earthquakes. In subducting tectonic plates, the increased pressure of greater depth within the earth may cause the minerals in the plate to metamorphose spontaneously into a new set of denser minerals that are stable at the higher pressure. This is thought to be the likely cause of certain types of deep earthquakes in subduction zones, including the deepest earthquakes ever recorded.

Rocks are also subjected to the three types of directed (non-uniform) stress – tension, compression, and shear.

- **Tension** is a directed (non-uniform) stress that pulls rock apart in opposite directions. The tensional (also called extensional) forces pull away from each other.
- **Compression** is a directed (non-uniform) stress that pushes rocks together. The compressional forces push towards each other.
- **Shear** is a directed (non-uniform) stress that pushes one side of a body of rock in one direction, and the opposite side of the body of rock in the opposite direction. The shear forces are pushing in opposite ways.

In response to stress, rock may undergo three different types of strain – elastic strain, ductile strain, or fracture.

- **Elastic strain** is reversible. Rock that has undergone only elastic strain will go back to its original shape if the stress is released.
- **Ductile strain** is irreversible. A rock that has undergone ductile strain will remain deformed even if the stress stops. Another term for ductile strain is plastic deformation.
- **Fracture** is also called rupture. A rock that has ruptured has abruptly broken into distinct pieces. If the pieces are offset—shifted in opposite directions from each other—the fracture is a fault.



Ductile and Brittle Strain

Earth's rocks are composed of a variety of minerals and exist in a variety of conditions. In different situations, rocks may act either as ductile materials that are able to undergo an extensive amount of ductile strain in response to stress, or as brittle materials, which will only undergo a little or no ductile strain before they fracture. The factors that determine whether a rock is ductile or brittle include:

- Composition—Some minerals, such as quartz, tend to be brittle and are thus more likely to break under stress. Other minerals, such as calcite, clay, and mica, tend to be ductile and can undergo much plastic deformation. In addition, the presence of water in rock tends to make it more ductile and less brittle.
- Temperature—Rocks become softer (more ductile) at higher temperature. Rocks at mantle and core temperatures are ductile and will not fracture under the stresses that occur deep within the earth. The crust, and to some extent the lithosphere, are cold enough to fracture if the stress is high enough.
- Lithostatic pressure—The deeper in the earth a rock is, the higher the lithostatic pressure it is subjected to. High lithostatic pressure reduces the possibility of fracture because the high pressure closes fractures before they can form or spread. The high lithostatic pressures of the earth's sub-lithospheric mantle and solid inner core, along with the high temperatures, are why there are no earthquakes deep in the earth.
- Strain rate—The faster a rock is being strained, the greater its chance of fracturing. Even brittle rocks and minerals, such as quartz, or a layer of cold basalt at the earth's surface, can undergo ductile deformation if the strain rate is slow enough.

Most earthquakes occur in the earth's crust. A smaller number of earthquakes occur in the uppermost mantle (to about 700 km deep) where subduction is taking place. Rocks in the deeper parts of the earth do not undergo fracturing and do not produce earthquakes because the temperatures and pressures there are high enough to make all strain ductile. No earthquakes originate from below the the earth's upper mantle.

Stress and Fault Types

The following correlations can be made between types of stress in the earth, and the type of fault that is likely to result:

- **Tension leads to normal faults.**
- **Compression leads to reverse or thrust faults.**
- **Horizontal shear leads to strike-slip faults.**

Correlations between type of stress and type of fault can have exceptions. For example, zones of horizontal stress will likely have strike-slip faults as the predominant fault type. However there may be active normal and thrust faults in such zones as well, particularly where there are bends or gaps in the major strike-slip faults.

To give another example, in a region of compression stress in the crust, where sheets of rock are stacked on active thrust faults, strike-slip faults commonly connect some of the thrust faults together.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Earthquakes. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/earthquakes.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.7.2: Strain is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

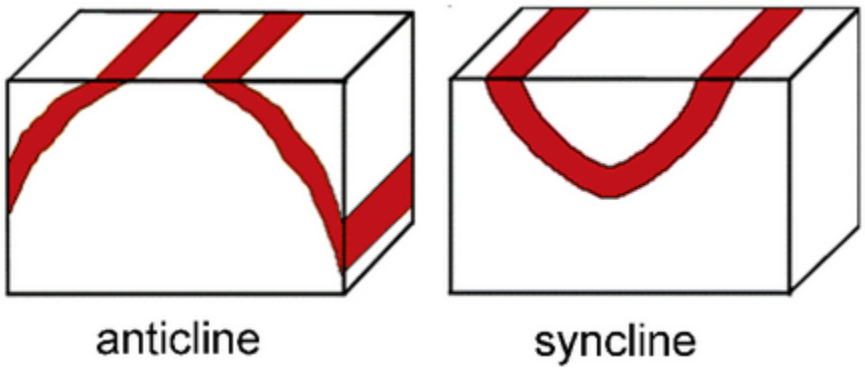
4.7.3: Folds

Ductile rocks behave plastically and become folded in response to stress. Even in the shallow crust where rocks are cool and relatively brittle, folding can occur if the stress is slow and steady and gives the rock enough time to gradually bend. If the stress is applied too quickly, rocks in the shallow crust will behave as brittle solids and break. Deeper in the crust, where the rocks are more ductile, folding happens more readily, even when the stress and strain occurs rapidly.

Anticlines and Synclines

The most basic types of folds are anticlines and synclines. Imagine a rug, the sides of which have been pushed toward each other forming ridges and valleys – the ridges are “up” folds and the valleys are “down” folds. In terms of geologic structures, the up folds are called anticlines and the down folds are called synclines.

In block diagrams like those shown below, the top of the block is the horizontal surface of the earth, the map view. The other two visible sides of the box are cross-sections, vertical slices through the crust. The colored layers represent stratified geologic formations that were originally horizontal, such as sedimentary beds or lava flows. Use the block diagrams to visualize the three-dimensional shapes of the geologic structures. Keep in mind that erosion has stripped away the upper parts of these structures so that map view reveals the interior of these structures.



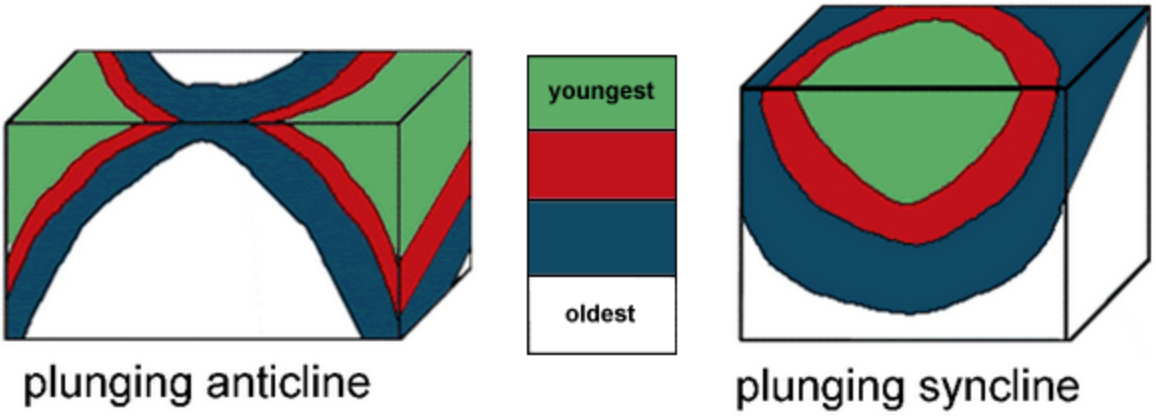
In map view, an anticline appears as parallel beds of the same rock type that dip away from the center of the fold. In an anticline, the oldest beds, the ones that were originally underneath the other beds, are at the center, along the axis of the fold. The axis is an imaginary line that marks the center of the fold on the map.

In map view, a syncline appears as a set of parallel beds that dip toward the center. In a syncline the youngest beds, the ones that were originally on top of the rest of the beds, are at the center, along the axis of the fold.

Anticlines and synclines form in sections of the crust that are undergoing compression, places where the crust is being pushed together.

Plunging Anticlines and Synclines

A plunging anticline or a plunging syncline is one that has its axis tilted from the horizontal so that the fold is plunging into the earth along its length. Plunge direction is the direction in which the axis of the fold tilts down into the earth.



In map view, a plunging anticline makes a U-shaped or V-shaped pattern that points, or closes, in the direction of plunge. A cross-section at a right angle to the axis of a plunging anticline looks the same as an anticline.

In map view, a plunging syncline makes a U-shaped or V-shaped pattern that opens in the direction of plunge.

Anticlines

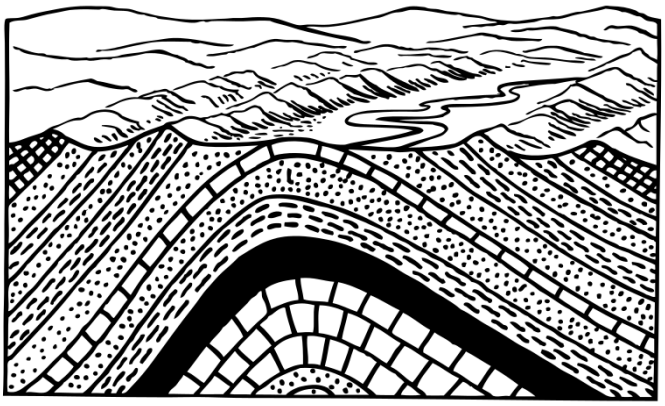


Figure 1. Diagram of an anticline.

In structural geology, an **anticline** is a fold that is convex up and has its oldest beds at its core. The term is not to be confused with *antiform*, which is a purely descriptive term for any fold that is convex up. Therefore if age relationships between various strata are unknown, the term antiform should be used.

On a geologic map, anticlines are usually recognized by a sequence of rock layers that are progressively older toward the center of the fold because the uplifted core of the fold is preferentially eroded to a deeper stratigraphic level relative to the topographically lower flanks. The strata dip away from the center, or *crest*, of the fold.



Figure 2. Anticline with syncline visible at far right. Note the man

standing before the formation, for scale.

If an anticline plunges (i.e., is inclined to the Earth's surface), the surface strata will form *V*s that point in the direction of plunge. Anticlines are often flanked by synclines although faulting can complicate and obscure the relationship between the two. Folds often form during crustal deformation as the result of shortening that accompanies orogenic mountain building. In many cases anticlines are formed by movement on non-planar faults during both shortening and extension, such as *ramp anticline*s and rollover anticlines.

Terminology

Any fold whose form is **convex** upward is an antiform. Antiforms containing progressively younger rocks from their core outwards are anticlines.

An anticline or antiform has a *crest*, which is the highest point on a given stratum along the top of the fold. A *hinge* in an anticline is the locus of maximum curvature or bending in a given stratum in the fold. An *axis* is an imaginary line connecting the hinges in the different strata in a two-dimensional cross-section through the anticline. Connecting the hinges or points of maximum curvature in the different layers in three dimensions produces an *axial plane* or *axial surface*. In a symmetrical anticline, a surface trace of the axial plane coincides with the crest. With an asymmetrical anticline, the surface trace of the axial plane or axis will be offset from the crest toward the steeper flank of the fold. An*overturned anticline* is an asymmetrical anticline with a flank or limb that has been tilted beyond perpendicular so that the beds in that limb are upside-down.

A structure that plunges in all directions to form a circular or elongate structure is a dome. Domes are generally formed from one main deformation event, or via diapirism from underlying magmatic intrusions or movement of upwardly mobile, mechanically ductile material such as rock salt (salt dome) and shale (shale diapir). The Richat Structure of the Sahara is considered a dome that has been laid bare by erosion.

An anticline which plunges at both ends is termed a *doubly plunging anticline*, and may be formed from multiple deformations, or superposition of two sets of folds, or be related to the geometry of the underlying detachment fault and the varying amount of displacement along the surface of that detachment fault. The highest point on a doubly plunging anticline (or any geologic structure for that matter) is called the "culmination."

An elongate dome which developed as the sediments were being deposited is referred to as a *pericline*.

An *anticlinorium* is a series of anticlinal folds on a regional-scale anticline. Examples include the Late Jurassic to Early Cretaceous Purcell Anticlinorium in British Columbia and the Blue Ridge anticlinorium of northern Virginia and Maryland in the Appalachians, or the Nittany Valley in central Pennsylvania.

Economic significance



Figure 3. Structural trap: anticlinal fold

Doubly plunging or faulted anticlines, culminations, and structural domes are favored locations for oil and natural gas drilling; the low density of petroleum causes it to buoyantly migrate upward to the highest parts of the fold, until stopped by a low-permeability barrier such as an impermeable stratum or fault zone. Examples of low-permeability seals that contain the hydrocarbons, oil and gas, in the ground include shale, limestone, sandstone, and even salt domes. The actual type of stratum does not matter as long

as it has low permeability.

Periclines are important focal points for pooling of hot, metal-laden formational brines, which can form manto ore deposits, Irish-type lead-zinc deposits and uranium deposits, amongst others.

Culminations in folded strata which are cut by shears and faults are favoured loci for deposition of saddle-reef style lode gold deposits.

Synclines


The Sideling Hill syncline as exposed in the Interstate 68 roadcut to the west of Hancock, Maryland, USA

In structural geology, a **syncline** is a fold with younger layers closer to the center of the structure. A **synclinorium** (plural **synclinoriums** or **synclinoria**) is a large syncline with superimposed smaller folds. Synclines are typically a downward fold, termed a synformal syncline (i.e. a trough); but synclines that point upwards, or perched, can be found when strata have been overturned and folded (an antiformal syncline).

Characteristics

On a geologic map, synclines are recognized by a sequence of rock layers that grow progressively younger, followed by the youngest layer at the fold's center or *hinge*, and by a reverse sequence of the same rock layers on the opposite side of the hinge. If the fold pattern is circular or elongate circular the structure is a basin. Folds typically form during crustal deformation as the result of compression that accompanies orogenic mountain building.

Notable Examples

- Powder River Basin, Wyoming, USA.
- Sideling Hill roadcut along Interstate 68 in western Maryland, USA, where the Rockwell Formation and overlying Purslane Sandstone are exposed.
- Western Lake Superior, which occupies a basin created by the Midcontinent Rift System
- Saou, a commune in the Drôme department in southeastern France
- The Catlins, an area in the southeastern corner of the South Island of New Zealand

Basins and Domes

A basin is a bowl-like depression in the strata (layers of rock). A basin is similar to a syncline, but instead of an axis it has a single point at the center. The strata all dip toward the center point and the youngest rock is at the center. In map view, the strata form concentric circles – a bull's eye pattern – around the center point.

A dome is an bulge in strata. A dome is similar to an anticline, but instead of an axis it has a single point at the center. The strata all dip away from the center point and the oldest rock is at the center. In map view, the strata form concentric circles – a bull's eye pattern – around the center point.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Geologic Structures. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/structures.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*
- Anticline. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Anticline. **License**: *CC BY-SA: Attribution-ShareAlike*
- Syncline. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Syncline. **License**: *CC BY-SA: Attribution-ShareAlike*
- Sideling Hill cut. **Authored by**: Acroterion. **Located at**: https://commons.wikimedia.org/wiki/File:Sideling_Hill_cut_MD1.jpg. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.7.3: Folds is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

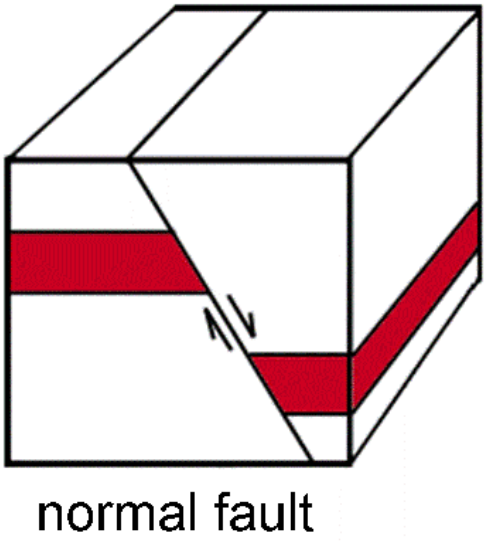
4.7.4: Faults

A fault is a planar surface within the earth, along which rocks have broken and slid. Faults are caused by elastic strain that culminates in brittle failure. The rocks on either side of a fault have shifted in opposite directions, called the offset directions. If a fault is not vertical, there are rocks above the fault and rocks beneath the fault.

- The rocks above a fault are called the hanging wall.
- The rocks beneath a fault are called the footwall.

Normal and Detachment Faults

In a normal fault, the hanging wall has moved down relative to the footwall.

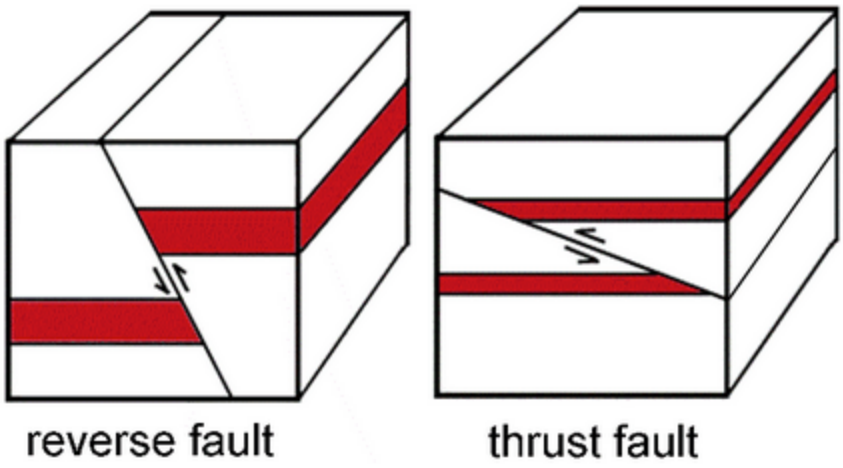


A detachment fault is a particular kind of normal fault that generally dips at a low angle. It separates rocks that were deep in the crust and ductile (granite and gneiss) from rocks of the upper crust (sedimentary or volcanic) that were brittle. Detachment faults occur along the boundaries of metamorphic core complexes (see below).

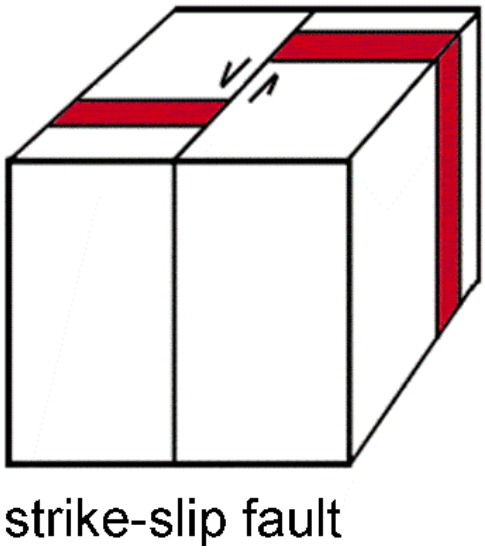
Normal and detachment faults form in sections of the crust that are undergoing tension, places where the crust is being stretched apart. A divergent plate boundary is a zone of large normal faults. Normal faults also occur in other zones of crustal tension, such as in the Basin and Range landscape region of the western United States.

Reverse and Thrust Faults

In a reverse or thrust fault, the hanging wall has moved up relative to the footwall. The difference between a reverse fault and a thrust fault is that a reverse fault has a steeper dip, more than 30°.



Reverse and thrust faults form in sections of the crust that are undergoing compression. A convergent plate boundary is a zone of major reverse and thrust faults. In fact, subduction zones are sometimes referred to as mega-thrust faults. Reverse and thrust faults also occur in other settings where the crust is being compressed, such as the Transverse Mountain Ranges, just north of Los Angeles.



Strike-Slip Faults

Strike-slip faults are steep or vertical faults along which the rocks on either side have moved horizontally in opposite directions. A transform plate boundary is a zone of large strike-slip faults. The San Andreas fault is an example of a major strike-slip fault at a transform boundary. Strike-slip faults also occur in other settings.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Geologic Structures. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/structures.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.7.4: Faults is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.7.5: Mountains

How do plate motions create mountains?



Plate tectonic processes create some of the world's most beautiful places. The North Cascades Mountains in Washington State are a continental volcanic arc. The mountains currently host some glaciers and there are many features left by the more abundant ice age glaciers. Changes in altitude make the range a habitable place for many living organisms.

Converging Plates

Converging plates create the world's largest mountain ranges. Each combination of plate types—continent-continent, continent-ocean, and ocean-ocean—creates mountains.

Converging Continental Plates

Two converging continental plates smash upwards to create gigantic mountain ranges (Figure 1). Stresses from this uplift cause folds, reverse faults, and thrust faults, which allow the crust to rise upwards. As was stated previously there is currently no mountain range of this type in the western U.S., but we can find one where India is pushing into Eurasia.

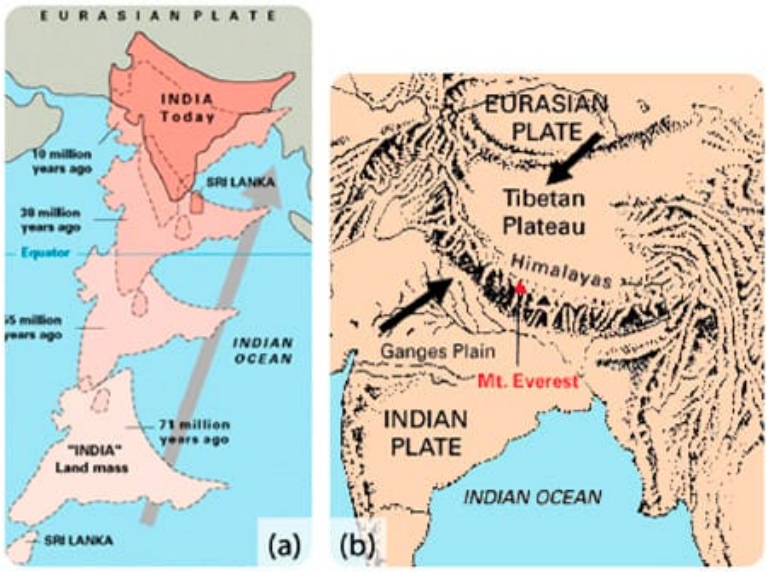


Figure 1. (a) The world's highest mountain range, the Himalayas, is growing from the collision between the Indian and the Eurasian plates. (b) The crumpling of the Indian and Eurasian plates of continental crust creates the Himalayas.

Subducting Oceanic Plates



Figure 2. The Andes Mountains are a chain of continental arc volcanoes that build up as the Nazca Plate subducts beneath the South American Plate.

Subduction of oceanic lithosphere at convergent plate boundaries also builds mountain ranges. This happens on continental crust, as in the Andes Mountains (Figure 2), or on oceanic crust, as with the Aleutian Islands, which we visited earlier. The Cascades Mountains of the western U.S. are also created this way.

Diverging Plates

Amazingly, even divergence can create mountain ranges. When tensional stresses pull crust apart, it breaks into blocks that slide up and drop down along normal faults. The result is alternating mountains and valleys, known as a basin-and-range (Figure 3). In basin-and-range, some blocks are uplifted to form ranges, known as horsts, and some are down-dropped to form basins, known as grabens.

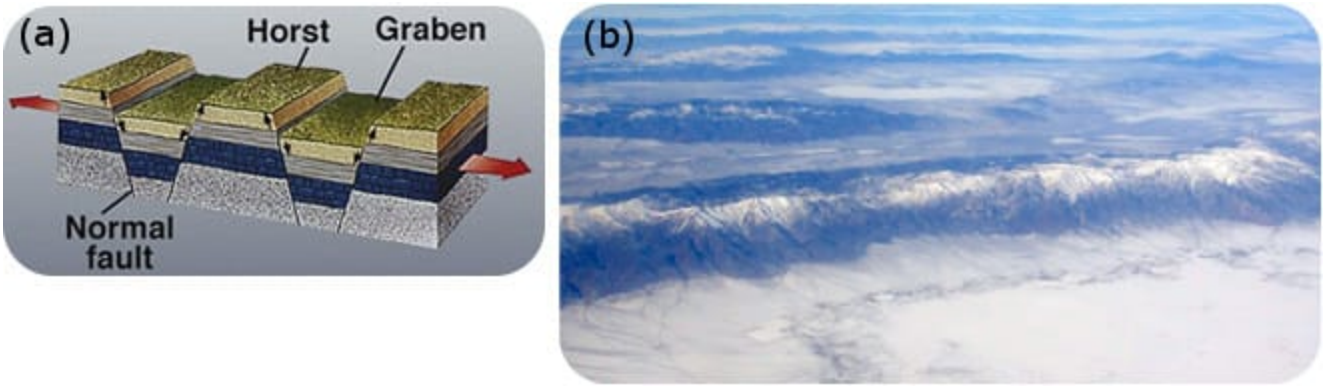


Figure 3. (a) Horsts and grabens. (b) Mountains in Nevada are of classic basin-and-range form. Watch this quick animation of movement of blocks in a basin-and-range setting.

Summary

- Converging or diverging plates cause mountains to grow.
- Subduction of oceanic crust beneath a continental or oceanic plate creates a volcanic arc.
- Tensional forces bring about block faulting, which creates a basin-and-range topography.

Contributors and Attributions

CC licensed content, Shared previously

- Mountain Building. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Mountain-Building/lesson/Mountain-Building-HS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.7.5: Mountains is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.8: EARTHQUAKES

Learning Objectives

Interpret types, causes, effects, and scientific measurements of earthquakes and seismic waves.

- Investigate the causes and common locations of earthquakes.
- Compare and contrast properties and mechanics of different types of waves.
- Define and compare magnitude versus intensity.
- Understand the hazards of earthquakes and recognize our limited ability to predict seismic activity.

4.8.1: The Nature of Earthquakes

4.8.2: Causes of Earthquakes

4.8.3: Common Locations of Earthquakes

4.8.4: The Ring of Fire

4.8.5: Body Waves

4.8.6: Surface Waves

4.8.7: Measuring Earthquakes

4.8.8: Magnitude versus Intensity

4.8.9: Earthquake Hazards

4.8.10: Earthquake Prediction

4.8.11: Seismicity and Earthquake Prediction

This page titled Chapter 4.8: Earthquakes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.1: The Nature of Earthquakes

Seismology

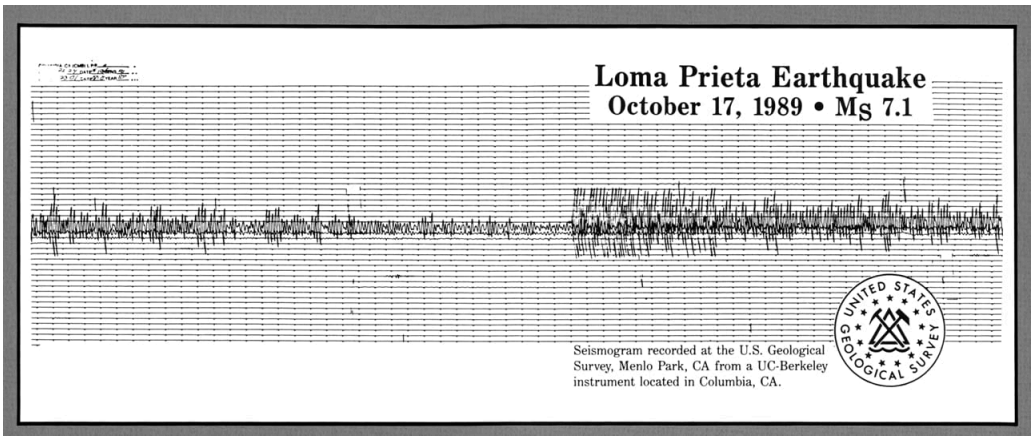
Seismology is the study of seismic waves. Seismology is also the study of earthquakes, mainly through the waves they produce. By measuring and analyzing seismic waves, seismologists can derive such information as:

- The epicenter of an earthquake
- The depth of an earthquake focus
- The magnitude (power) of an earthquake
- The type of fault movement that produced an earthquake
- Whether an earthquake beneath the ocean is likely to have generated a tsunami (a set of giant ocean waves)

In addition to information about earthquakes and faults, seismology gives us knowledge of the layers of the earth. Much of what we know about the crust, lithosphere, asthenosphere, mantle, and core comes from seismology. See the Earth's interior Basics page.

Seismology also gives us information about underground nuclear testing that takes place anywhere on earth, allows possible oil reservoirs to be located within the earth's crust, and helps us predict when a volcano is about to erupt.

Seismographs and seismometers are the instruments used to measure seismic waves. The traditional analog seismograph utilizes a pen (stylus) embedded in a heavy weight, which is suspended on springs. When the earth moves during an earthquake, a piece of paper rolling beneath the stylus moves with the earth, but the stylus, with its weight suspended on springs, remains stationary, drawing lines on the sheet of paper that show the seismic motions of the earth. The USGS photo below shows a seismogram from a seismograph located in Columbia, California that recorded the 1989 Loma Prieta Earthquake.



With modern technology, seismographs with pens and rolling sheets of paper are being replaced by seismometers with electronic sensors and computer screens. Seismographs and seismometers both produce a seismogram, which is a graphic record of the seismic waves, viewed either on paper or on a computer monitor.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Earthquakes. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/earthquakes.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.8.1: The Nature of Earthquakes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.2: Causes of Earthquakes

The following video explains the cause of earthquakes.

A link to an interactive elements can be found at the bottom of this page.

Overview of Elastic Rebound Theory

In an earthquake, the initial point where the rocks rupture in the crust is called the **focus**. The epicenter is the point on the land surface that is directly above the focus. In about 75% of earthquakes, the focus is in the top 10 to 15 kilometers (6 to 9 miles) of the crust. Shallow earthquakes cause the most damage because the focus is near where people live. However, it is the epicenter of an earthquake that is reported by scientists and the media (figure 1).

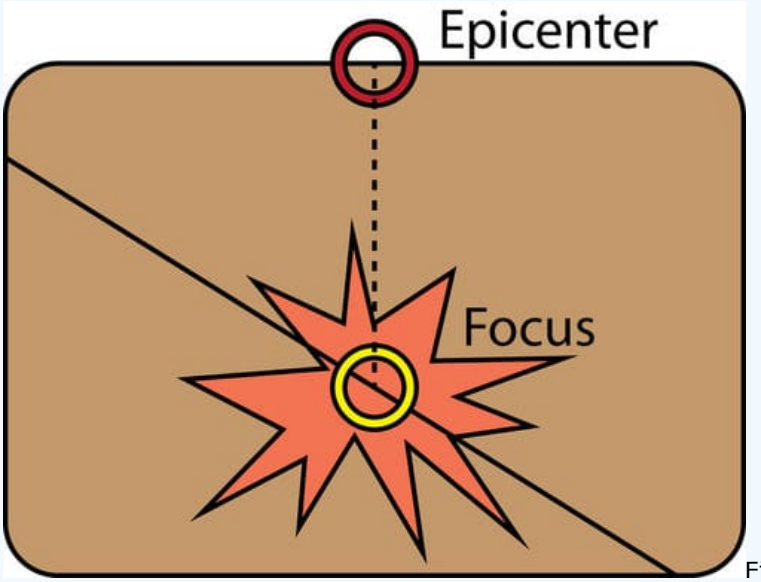


Figure 1. In the vertical cross section of crust, there are two features labeled—the focus and the epicenter, which is directly above the focus.

Watch this animation summarizing elastic rebound theory.

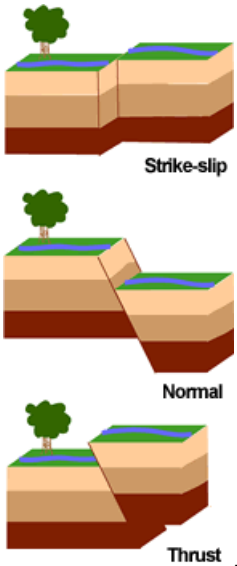


Figure 2. Fault types

Tectonic earthquakes occur anywhere in the earth where there is sufficient stored elastic strain energy to drive fracture propagation along a **fault plane**. The sides of a fault move past each other smoothly and aseismically only if there are no irregularities or asperities along the fault surface that increase the frictional resistance. Most fault surfaces do have such asperities and this leads to a form of stick-slip behavior. Once the fault has locked, continued relative motion between the plates leads to increasing stress and therefore, stored strain energy in the volume around the fault surface. This continues until the stress has risen sufficiently to break through the asperity, suddenly allowing sliding over the locked portion of the fault, releasing the stored energy.[1]

This energy is released as a combination of radiated elastic strain seismic waves, frictional heating of the fault surface, and cracking of the rock, thus causing an earthquake. This process of gradual build-up of strain and stress punctuated by occasional sudden earthquake failure is referred to as the elastic-rebound theory. It is estimated that only 10 percent or less of an earthquake's total energy is radiated as seismic energy. Most of the earthquake's energy is used to power the earthquake fracture growth or is converted into heat generated by friction. Therefore, earthquakes lower the Earth's available elastic potential energy and raise its temperature, though these changes are negligible compared to the conductive and convective flow of heat out from the Earth's deep interior.[2]

Earthquake Fault Types

There are three main types of fault, all of which may cause an interplate earthquake: normal, reverse (thrust) and strike-slip. Normal and reverse faulting are examples of dip-slip, where the displacement along the fault is in the direction of dip and movement on them involves a vertical component. Normal faults occur mainly in areas where the crust is being extended such as a divergent boundary. Reverse faults occur in areas where the crust is being shortened such as at a convergent boundary. Strike-slip faults are steep structures where the two sides of the fault slip horizontally past each other; transform boundaries are a particular type of strike-slip fault. Many earthquakes are caused by movement on faults that have components of both dip-slip and strike-slip; this is known as oblique slip.

Reverse faults, particularly those along convergent plate boundaries are associated with the most powerful earthquakes, megathrust earthquakes, including almost all of those of magnitude 8 or more. Strike-slip faults, particularly continental transforms, can produce major earthquakes up to about magnitude 8. Earthquakes associated with normal faults are generally less than magnitude 7. For every unit increase in magnitude, there is a roughly thirtyfold increase in the energy released. For instance, an earthquake of magnitude 6.0 releases approximately 30 times more energy than a 5.0 magnitude earthquake and a 7.0 magnitude earthquake releases 900 times (30 × 30) more energy than a 5.0 magnitude of earthquake. An 8.6 magnitude earthquake releases the same amount of energy as 10,000 atomic bombs like those used in World War II.


Figure 3. Aerial photo of the San Andreas Fault in the Carrizo Plain, northwest of Los Angeles

This is so because the energy released in an earthquake, and thus its magnitude, is proportional to the area of the fault that ruptures2.0.CO;2." id="return-footnote-276-3" href="#footnote-276-3" aria-label="Footnote 3">[3] and the stress drop. Therefore, the longer the length and the wider the width of the faulted area, the larger the resulting magnitude. The topmost, brittle part of

the Earth's crust, and the cool slabs of the tectonic plates that are descending down into the hot mantle, are the only parts of our planet which can store elastic energy and release it in fault ruptures. Rocks hotter than about 300 degrees Celsius flow in response to stress; they do not rupture in earthquakes.[4] The maximum observed lengths of ruptures and mapped faults (which may break in a single rupture) are approximately 1000 km. Examples are the earthquakes in Chile, 1960; Alaska, 1957; Sumatra, 2004, all in subduction zones. The longest earthquake ruptures on strike-slip faults, like the San Andreas Fault (1857, 1906), the North Anatolian Fault in Turkey (1939) and the Denali Fault in Alaska (2002), are about half to one third as long as the lengths along subducting plate margins, and those along normal faults are even shorter.

The most important parameter controlling the maximum earthquake magnitude on a fault is however not the maximum available length, but the available width because the latter varies by a factor of 20. Along converging plate margins, the dip angle of the rupture plane is very shallow, typically about 10 degrees.[5] Thus the width of the plane within the top brittle crust of the Earth can become 50 to 100 km (Japan, 2011; Alaska, 1964), making the most powerful earthquakes possible.

Strike-slip faults tend to be oriented near vertically, resulting in an approximate width of 10 km within the brittle crust,[6] thus earthquakes with magnitudes much larger than 8 are not possible. Maximum magnitudes along many normal faults are even more limited because many of them are located along spreading centers, as in Iceland, where the thickness of the brittle layer is only about 6 km.[7]

In addition, there exists a hierarchy of stress level in the three fault types. Thrust faults are generated by the highest, strike slip by intermediate, and normal faults by the lowest stress levels.[8] This can easily be understood by considering the direction of the greatest principal stress, the direction of the force that "pushes" the rock mass during the faulting. In the case of normal faults, the rock mass is pushed down in a vertical direction, thus the pushing force (**greatest** principal stress) equals the weight of the rock mass itself. In the case of thrusting, the rock mass "escapes" in the direction of the **least** principal stress, namely upward, lifting the rock mass up, thus the overburden equals the least principal stress. Strike-slip faulting is intermediate between the other two types described above. This difference in stress regime in the three faulting environments can contribute to differences in stress drop during faulting, which contributes to differences in the radiated energy, regardless of fault dimensions.

Earthquakes away from Plate Boundaries

Where plate boundaries occur within the continental lithosphere, deformation is spread out over a much larger area than the plate boundary itself. In the case of the San Andreas fault continental transform, many earthquakes occur away from the plate boundary and are related to strains developed within the broader zone of deformation caused by major irregularities in the fault trace (e.g., the "Big bend" region). The Northridge earthquake was associated with movement on a blind thrust within such a zone. Another example is the strongly oblique convergent plate boundary between the Arabian and Eurasian plates where it runs through the northwestern part of the Zagros Mountains. The deformation associated with this plate boundary is partitioned into nearly pure thrust sense movements perpendicular to the boundary over a wide zone to the southwest and nearly pure strike-slip motion along the Main Recent Fault close to the actual plate boundary itself. This is demonstrated by earthquake focal mechanisms.[9]

All tectonic plates have internal stress fields caused by their interactions with neighboring plates and sedimentary loading or unloading (e.g. deglaciation).[10] These stresses may be sufficient to cause failure along existing fault planes, giving rise to intraplate earthquakes.[11]

Shallow-Focus and Deep-Focus Earthquakes



Figure 4. Collapsed Gran Hotel building in the San Salvador metropolis, after the shallow 1986 San Salvador earthquake.

The majority of tectonic earthquakes originate at the ring of fire in depths not exceeding tens of kilometers. Earthquakes occurring at a depth of less than 70 km are classified as **shallow-focus** earthquakes, while those with a focal-depth between 70 and 300 km are commonly termed **mid-focus** or **intermediate-depth** earthquakes. In subduction zones, where older and colder oceanic crust descends beneath another tectonic plate, **deep-focus** earthquakes may occur at much greater depths (ranging from 300 up to 700 kilometers).[12]

These seismically active areas of subduction are known as Wadati–Benioff zones. Deep-focus earthquakes occur at a depth where the subducted lithosphere should no longer be brittle, due to the high temperature and pressure. A possible mechanism for the generation of deep-focus earthquakes is faulting caused by olivine undergoing a phase transition into a spinel structure.[13]

Earthquakes and Volcanic Activity

Earthquakes often occur in volcanic regions and are caused there, both by tectonic faults and the movement of magma in volcanoes. Such earthquakes can serve as an early warning of volcanic eruptions, as during the 1980 eruption of Mount St. Helens.[14] Earthquake swarms can serve as markers for the location of the flowing magma throughout the volcanoes. These swarms can be recorded by seismometers and tiltmeters (a device that measures ground slope) and used as sensors to predict imminent or upcoming eruptions.[15]

Rupture Dynamics

A tectonic earthquake begins by an initial rupture at a point on the fault surface, a process known as nucleation. The scale of the nucleation zone is uncertain, with some evidence, such as the rupture dimensions of the smallest earthquakes, suggesting that it is smaller than 100 m while other evidence, such as a slow component revealed by low-frequency spectra of some earthquakes, suggest that it is larger. The possibility that the nucleation involves some sort of preparation process is supported by the observation that about 40% of earthquakes are preceded by foreshocks. Once the rupture has initiated, it begins to propagate along the fault surface. The mechanics of this process are poorly understood, partly because it is difficult to recreate the high sliding velocities in a laboratory. Also the effects of strong ground motion make it very difficult to record information close to a nucleation zone.[16]

Rupture propagation is generally modeled using a fracture mechanics approach, likening the rupture to a propagating mixed mode shear crack. The rupture velocity is a function of the fracture energy in the volume around the crack tip, increasing with decreasing fracture energy. The velocity of rupture propagation is orders of magnitude faster than the displacement velocity across the fault. Earthquake ruptures typically propagate at velocities that are in the range 70–90% of the S-wave velocity, and

this is independent of earthquake size. A small subset of earthquake ruptures appear to have propagated at speeds greater than the S-wave velocity. These supershear earthquakes have all been observed during large strike-slip events. The unusually wide zone of coseismic damage caused by the 2001 Kunlun earthquake has been attributed to the effects of the sonic boom developed in such earthquakes. Some earthquake ruptures travel at unusually low velocities and are referred to as slow earthquakes. A particularly dangerous form of slow earthquake is the tsunami earthquake, observed where the relatively low felt intensities, caused by the slow propagation speed of some great earthquakes, fail to alert the population of the neighboring coast, as in the 1896 Sanriku earthquake.[17]

Earthquake Clusters

Most earthquakes form part of a sequence, related to each other in terms of location and time.[18] Most earthquake clusters consist of small tremors that cause little to no damage, but there is a theory that earthquakes can recur in a regular pattern.[19]

Aftershocks

An aftershock is an earthquake that occurs after a previous earthquake, the mainshock. An aftershock is in the same region of the main shock but always of a smaller magnitude. If an aftershock is larger than the main shock, the aftershock is redesignated as the main shock and the original main shock is redesignated as a foreshock. Aftershocks are formed as the crust around the displaced fault plane adjusts to the effects of the main shock.[20]

Earthquake Swarms

Earthquake swarms are sequences of earthquakes striking in a specific area within a short period of time. They are different from earthquakes followed by a series of aftershocks by the fact that no single earthquake in the sequence is obviously the main shock, therefore none have notable higher magnitudes than the other. An example of an earthquake swarm is the 2004 activity at Yellowstone National Park.[21] In August 2012, a swarm of earthquakes shook Southern California's Imperial Valley, showing the most recorded activity in the area since the 1970s.[22]

Sometimes a series of earthquakes occur in what has been called an earthquake storm, where the earthquakes strike a fault in clusters, each triggered by the shaking or stress redistribution of the previous earthquakes. Similar to aftershocks but on adjacent segments of fault, these storms occur over the course of years, and with some of the later earthquakes as damaging as the early ones. Such a pattern was observed in the sequence of about a dozen earthquakes that struck the North Anatolian Fault in Turkey in the 20th century and has been inferred for older anomalous clusters of large earthquakes in the Middle East.[23]

---

1. Ohnaka, M. (2013). *The Physics of Rock Failure and Earthquakes*. Cambridge University Press. p. 148. ↩
2. Spence, William; S. A. Sipkin; G. L. Choy (1989). "Measuring the Size of an Earthquake." United States Geological Survey. Retrieved 2006-11-03. ↩
3. Wyss, M. (1979). "Estimating expectable maximum magnitude of earthquakes from fault dimensions". *Geology***7** (7): 336–340 doi:10.1130/0091-7613(1979)7<336:EMEMOE>2.0.CO;2. ↩
4. Sibson R. H. (1982) "Fault Zone Models, Heat Flow, and the Depth Distribution of Earthquakes in the Continental Crust of the United States," *Bulletin of the Seismological Society of America*, Vol 72, No. 1, pp. 151–163. See also Sibson, R. H. (2002) "Geology of the crustal earthquake source," *International handbook of earthquake and engineering seismology*, Volume 1, Part 1, page 455, eds. W.H.K. Lee, H. Kanamori, P.C. Jennings, and C. Kisslinger, Academic Press. ↩
5. "Global Centroid Moment Tensor Catalog." Globalcmt.org. Retrieved 2011-07-24. ↩
6. "Instrumental California Earthquake Catalog." WGCEP. Retrieved 2011-07-24. ↩
7. Hjaltadóttir S., 2010, "Use of relatively located microearthquakes to map fault patterns and estimate the thickness of the brittle crust in Southwest Iceland." See also "Reports and publications | Seismicity | Icelandic Meteorological office." En.vedur.is. Retrieved 2011-07-24. ↩
8. Schorlemmer, D.; Wiemer, S.; Wyss, M. (2005). "Variations in earthquake-size distribution across different stress regimes." *Nature* **437** (7058): 539–542. doi:10.1038/nature04094. ↩
9. Talebian, M; Jackson, J (2004). "A reappraisal of earthquake focal mechanisms and active shortening in the Zagros mountains of Iran." *Geophysical Journal International***156** (3): 506–526. doi:10.1111/j.1365-246X.2004.02092.x. ↩
10. Nettles, M.; Ekström, G. (May 2010). "Glacial Earthquakes in Greenland and Antarctica." *Annual Review of Earth and Planetary Sciences***38** (1): 467–491. doi:10.1146/annurev-earth-040809-152414. ↩

11. Noson, Qamar, and Thorsen (1988). *Washington State Earthquake Hazards: Washington State Department of Natural Resources*. Washington Division of Geology and Earth Resources Information Circular 85. ↩
12. "M7.5 Northern Peru Earthquake of 26 September 2005" (PDF). National Earthquake Information Center. 17 October 2005. Retrieved2008-08-01. ↩
13. Greene II, H. W.; Burnley, P. C. (October 26, 1989). "A new self-organizing mechanism for deep-focus earthquakes." *Nature***341**(6244): 733–737. doi: 10.1038/341733a0. ↩
14. Foxworthy and Hill (1982). *Volcanic Eruptions of 1980 at Mount St. Helens, The First 100 Days: USGS Professional Paper 1249*. ↩
15. Watson, John; Watson, Kathie (January 7, 1998). "Volcanoes and Earthquakes." United States Geological Survey. Retrieved May 9,2009. ↩
16. National Research Council (U.S.). Committee on the Science of Earthquakes (2003). "5. Earthquake Physics and Fault-System Science." *Living on an Active Earth: Perspectives on Earthquake Science*. Washington D.C.: National Academies Press. p. 418. Retrieved 8 July 2010. ↩
17. *Ibid.* ↩
18. "What are Aftershocks, Foreshocks, and Earthquake Clusters?" ↩
19. "Repeating Earthquakes." United States Geological Survey. January 29, 2009. Retrieved May 11, 2009. ↩
20. "What are Aftershocks, Foreshocks, and Earthquake Clusters?" ↩
21. "Earthquake Swarms at Yellowstone." United States Geological Survey. Retrieved 2008-09-15. ↩
22. Duke, Alan. "Quake 'swarm' shakes Southern California." CNN. Retrieved 27 August 2012. ↩
23. Amos Nur; Cline, Eric H. (2000). "Poseidon's Horses: Plate Tectonics and Earthquake Storms in the Late Bronze Age Aegean and Eastern Mediterranean" (PDF). *Journal of Archaeological Science***27** (1): 43–63. See also "Earthquake Storms." *Horizon*. 1 April 2003. Retrieved 2007-05-02. ↩

Contributors and Attributions

CC licensed content, Shared previously

- 7.2: The Nature of Earthquakes. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/7.2/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Naturally occurring earthquakes. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Earthquake#Naturally_occurring_earthquakes. **License**: *CC BY-SA: Attribution-ShareAlike*

All rights reserved content

- Earthquakes 101. **Authored by**: National Geographic. **Located at**: https://youtu.be/VSgB1IWr6O4. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

This page titled 4.8.2: Causes of Earthquakes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.3: Common Locations of Earthquakes

Earthquakes and Plate Boundaries

Most, but not all, earthquakes occur at or near plate boundaries. A great deal of stress is concentrated and a great deal of strain, much of it in the form of rupture of the earth, takes place at locations where two plates diverge, transform, or converge relative to each other.

Tension is the dominant stress at divergent plate boundaries. Normal faults and rift valleys as the predominant earthquake-related structures at divergent plate boundaries. Earthquakes at divergent plate boundaries are usually relatively shallow, and, though they can be damaging, the most powerful earthquakes at divergent plate boundaries are not nearly as powerful as the most powerful earthquakes at convergent plate boundaries.

Transform plate boundaries are zones dominated by horizontal shear, with strike-slip faults the most characteristic fault type. Most transform plate boundaries cut through relatively thin oceanic crust, part of the structure of the ocean floor, and produce relatively shallow earthquakes that are only rarely of major magnitude. However, where transform plate boundaries and their strike-slip faults cut through the thicker crust of islands or the even thicker crust of continents, more stress may need to build up before the thicker masses of rock will rupture, and so the magnitudes of earthquakes can be higher than in transform plate boundary zones confined to thin oceanic crust. This is evident in such places as the San Andreas fault zone of California, where a transform fault cuts through continental crust and earthquakes there sometimes exceed 7.0 in magnitude.

Convergent plate boundaries are dominated by compression. The major faults found in convergent plate boundaries are usually reverse or thrust faults, including a master thrust fault at the boundary between the two plates and typically several more major thrust faults running roughly parallel to the plate boundary. The most powerful earthquakes that have been measured are subduction earthquakes, up to greater than 9.0 in magnitude. All subduction zones in the world are at risk of subduction earthquakes with magnitudes up to or even greater than 9.0 in extreme cases, and are likely to produce tsunamis. This includes the Cascadia subduction zone of northern California and coastal Oregon and Washington, the Aleutian subduction zone of southern Alaska, the Kamchatka subduction zone of Pacific Russia, the Acapulco subduction zone of southern Pacific Mexico, the Central American subduction zone, the Andean subduction zone, the West Indian or Caribbean subduction zone, and subduction zones of Indonesia, Japan, the Phillipines, and several more subduction zones in the western and southwestern Pacific Ocean.

Intraplate Earthquakes

Some earthquakes take place far away from plate boundaries. Earthquakes can occur wherever there is sufficient stress in the earth's crust to drive rocks to rupture.

For example, Hawaii is thousands of km (thousands of miles) from any plate boundary, but the volcanoes that compose the islands have built up so rapidly that they are still undergoing gravitational stabilization. Sectors of the Hawaiian islands occasionally slump along normal faults, producing intraplate earthquakes. Most of the earthquakes occur on the big island of Hawaii, which is composed of the youngest, most recently built volcanoes. The geologic record shows that parts of the older islands have undergone major collapses in the last few million years, with sections of the islands sliding out to the seafloor in landslides floored on shallow normal faults.

Another example is the Basin and Range region of the western United States, including Nevada and eastern Utah, where the crust is subjected to tension. Earthquakes occur there on normal faults, far inland from the plate boundaries on the West Coast. The tension in the crust of the Basin and Range province may be partly due to a mid-ocean ridge system that subducted beneath California and is now located beneath the Basin and Range, causing tension in the lithosphere.

The region around Yellowstone National Park also undergoes occasional major earthquakes on normal faults. Earthquakes in that area may be due to the Yellowstone hot spot causing differential thermal expansion of the lithosphere in a broad zone round the hot spot center.

Several East Coast cities, including Boston, New York, and Charleston in South Carolina, have experienced damaging earthquakes in the last two centuries. The faults beneath these cities may date back to the rifting of Pangea and the opening up of the Atlantic

Ocean beginning around 200 million years ago.

In the area of the town of New Madrid, along the Mississippi River in southeastern Missouri and western Tennessee, great earthquakes occurred in 1811-1812. Minor to moderate earthquakes continue to occur there, keeping active the possibility of damaging earthquakes occurring there again in the future. The fault system beneath that area may date from times of continental collision and continental rifting in the distant geologic past, and recent stress in the crust around New Madrid may be from the massive build-up of sediment in the Mississippi River delta region, which spreads out to the south of that area.

Earthquakes and Volcanoes

The connections between earthquakes and volcanoes are not always obvious. However, when magma is moving up beneath a volcano, and when a volcano is erupting, it produces earthquakes. Volcanic earthquakes are distinct from the more common type of earthquakes that occur by elastic rebound along faults.

Seismologists can use the patterns and signals of earthquakes coming from beneath volcanoes to predict that the volcano is about to erupt, and can use seismic waves to see that a volcano is undergoing an eruption even if the volcano is at a remote location, hidden in darkness, or hidden in storm clouds.

Volcanic vents, and volcanoes in general, are commonly located along faults, or at the intersection of several faults. Major faults that already exist in the crust may be natural paths to channel rising magma. However, on major volcanic edifices, shallower faults are a product of the development of the volcano. There are feedback effects between the upward pressure of magma buoyancy in the crust, the growth of faults in volcanic zones, and the venting of volcanoes, which is not yet completely understood.

As was noted at the beginning of this section, not quite all earthquakes are due to the slippage of solid blocks of rock along faults. When a volcano undergoes a powerful pyroclastic eruption – in other words, when a volcano explodes – it causes the earth to shake. Earthquakes caused by an explosive volcanic eruptions produce a different seismic signal than earthquakes caused by slippage along faults.

Another example of earthquakes that are caused at least in part by magma movement, rather than by slippage of entirely solid rock along faults, is earthquakes set off by the movement of magma upward beneath a volcano, or up to higher levels in the crust whether or not there is a volcano on top. Such upward movement of magma within the crust is sometimes called magma injection. Seismologists are still researching the interactions between movement of magma in the crust, and related slippage along faults that may be caused by the pressure and movement of the magma.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Earthquakes. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/earthquakes.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.8.3: Common Locations of Earthquakes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.4: The Ring of Fire

The **Ring of Fire** is an area where a large number of earthquakes and volcanic eruptions occur in the basin of the Pacific Ocean. In a 40,000 km (25,000 mi) horseshoe shape, it is associated with a nearly continuous series of oceanic trenches, volcanic arcs, and volcanic belts and/or plate movements. It has 452 volcanoes and is home to over 75% of the world's active and dormant volcanoes. It is sometimes called the **circum-Pacific belt**.

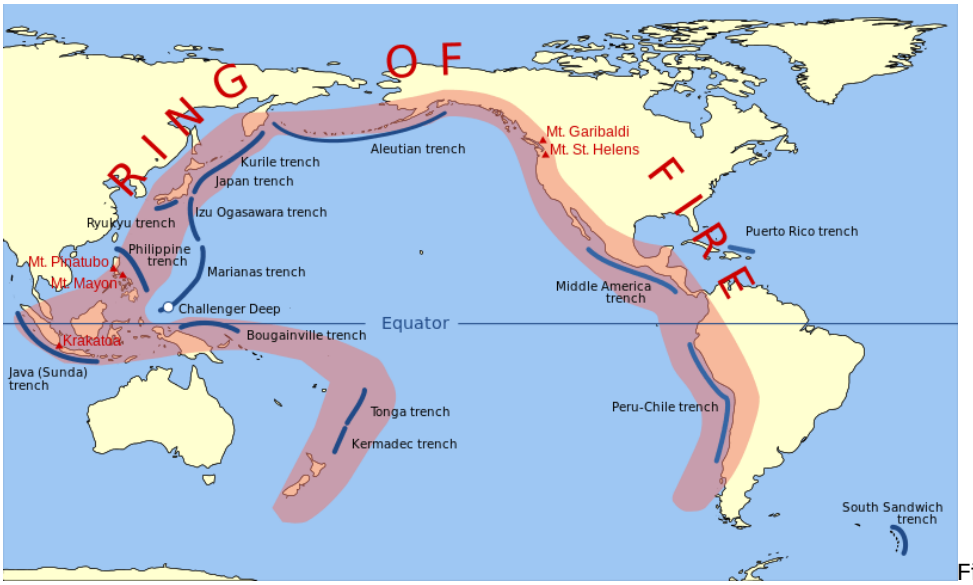


Figure 1. The Pacific Ring of Fire

About 90% of the world's earthquakes and 81% of the world's largest earthquakes occur along the Ring of Fire. The next most seismically active region (5–6% of earthquakes and 17% of the world's largest earthquakes) is the Alpide belt, which extends from Java to Sumatra through the Himalayas, theMediterranean, and out into the Atlantic. The Mid-Atlantic Ridge is the third most prominent earthquake belt.

The Ring of Fire is a direct result of plate tectonics and the movement and collisions of lithospheric plates. The eastern section of the ring is the result of the Nazca Plate and the Cocos Plate being subducted beneath the westward moving South American Plate. The Cocos Plate is being subducted beneath theCaribbean Plate, in Central America. A portion of the Pacific Plate along with the small Juan de Fuca Plate are being subducted beneath the North American Plate. Along the northern portion, the northwestward-moving Pacific plate is being subducted beneath the Aleutian Islands arc. Farther west, the Pacific plate is being subducted along the Kamchatka Peninsula arcs on south past Japan. The southern portion is more complex, with a number of smaller tectonic plates in collision with the Pacific plate from the Mariana Islands, the Philippines,Bougainville, Tonga, and New Zealand; this portion excludes Australia, since it lies in the center of its tectonic plate. Indonesia lies between the *Ring of Fire* along the northeastern islands adjacent to and including New Guinea and the *Alpide belt* along the south and west from Sumatra, Java, Bali, Flores, and Timor. The famous and very active San Andreas Fault zone of California is a transform faultwhich offsets a portion of the East Pacific Rise under southwestern United Statesand Mexico. The motion of the fault generates numerous small earthquakes, at multiple times a day, most of which are too small to be felt. The active Queen Charlotte Fault on the west coast of the Haida Gwaii, British Columbia, Canada, has generated three large earthquakes during the 20th century: a magnitude 7 event in 1929; a magnitude 8.1 in 1949 (Canada's largest recorded earthquake); and a magnitude 7.4 in 1970.

Contributors and Attributions

CC licensed content, Shared previously

- Ring of Fire. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Ring_of_Fire. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.8.4: The Ring of Fire is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.5: Body Waves

Body waves travel through the interior of the earth. There are two types of body waves: P-waves and S-waves.

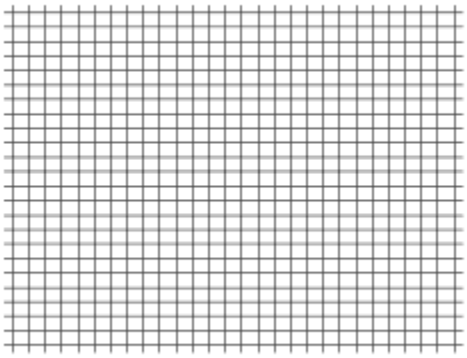
P-Waves

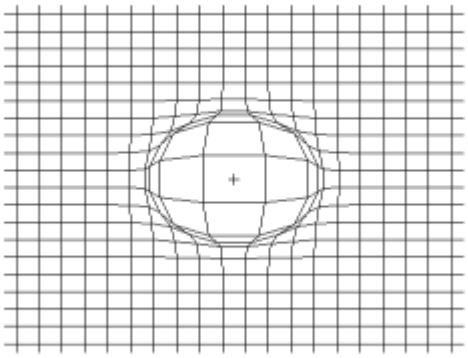
The P in P-waves stands for primary, because these are the fastest seismic waves and are the first to be detected once an earthquake has occurred. P-waves travel through the earth's interior many times faster than the speed of a jet airplane, taking only a few minutes to travel across the earth.

P-waves are predominantly compressional waves. As a P-wave passes, material compresses in the same direction the wave is moving, and then extends back to its original thickness once the wave has passed. The speed at which P-waves travel through material is determined by:

- rigidity—how strongly the material resists being bent sideways and is able to straighten itself out once the shearing force has passed – the more rigid the material, the faster the P-waves
- compressibility—how much the material can be compressed into a smaller volume and then recover its previous volume once the compressing force has passed; the more compressible the material, the faster the P-waves
- density—how much mass the material contains in a unit of volume; the greater the density of the material, the slower the P-waves

The animations below show P-waves propagating across a plane (left) and from a point source (right).





P-waves travel through liquids and gases as well as through solids. Although liquids and gases have zero rigidity, they have compressibility, which enables them to transmit P-waves. Sound waves are P-waves moving through the air.

Because the earth's mantle becomes more rigid and compressible as the depth below the asthenosphere increases, P-waves travel faster as they go deeper in the mantle. The density of the mantle also increases with depth below the asthenosphere. The higher density reduces the speed of seismic waves. However, the effects of increased rigidity and compressibility in the deep mantle are much greater than the effect of the increased density.

P-waves travel through materials with rigidity and/or compressiblity, and density	
greater rigidity	faster P-waves
greater compressibility	faster P-waves
greater density	slower P-waves

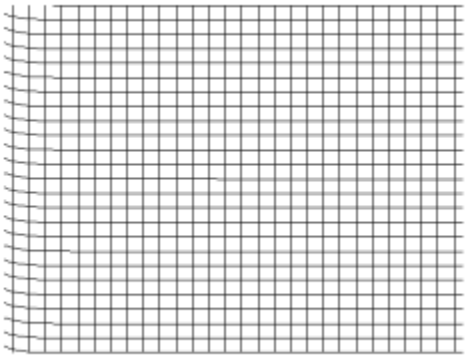
S-Waves

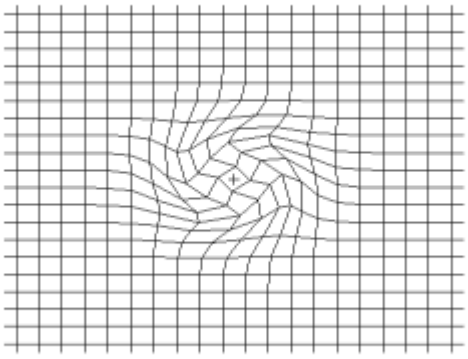
The S in S-waves stands for secondary, because they are the second-fastest seismic waves and the second type to be detected once an earthquake has occurred. Although S-waves are slower than P-waves, they still travel fast, over half the speed of P-waves, moving at thousands of kilometers per hour through the earth's crust and mantle.

S-waves are shear waves (though that is not what the S stands for). They move by material flexing or deforming sideways (shearing) from the direction of wave travel, and then returning to the original shape once the wave passes. The speed at which S-waves travel through material is determined only by:

- rigidity — how strongly the material resists being bent sideways and is able to straighten itself out once the shearing force has passed – the more rigid the material, the faster the S-waves
- density — how much mass the material contains in a unit of volume – the greater the density of the material, the slower the S-waves

The animations below show S-waves propogating across a plane (left) and from a point source (right):





S-waves can travel only through solids, because only solids have rigidity. S-waves cannot travel through liquids or gases.

Because the earth's mantle becomes more rigid as its depth below the asthenosphere increases, S-waves travel faster as they go deeper in the mantle. The density of the mantle also increases at greater depth, which has the effect of reducing the speed of seismic waves, but the increase in rigidity is much greater than the increase in density, so S-waves speed up as they get deeper in the mantle, in spite of the increased density.

S-waves travel through materials with rigidity and density	
greater rigidity	faster S-waves
greater density	slower S-waves

Contributors and Attributions

CC licensed content, Shared previously

- Seismic waves. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/earthquakes.html#waves. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*
- Onde compression impulsion 1d (p-wave). **Authored by**: Christophe Dang Ngoc Chan. **Located at**: https://en.Wikipedia.org/wiki/File:Onde_compression_impulsion_1d_30_petit.gif. **License**: *CC BY-SA: Attribution-ShareAlike*
- Ondes compression 2d 20 (p-wave). **Authored by**: Christophe Dang Ngoc Chan. **Located at**: https://en.Wikipedia.org/wiki/File:Ondes_compression_2d_20_petit.gif. **License**: *CC BY-SA: Attribution-ShareAlike*
- Onde cisaillement impulsion 1d 30 (s-wave). **Authored by**: Christophe Dang Ngoc Chan. **Located at**: https://en.Wikipedia.org/

wiki/File:Onde_cisaillement_impulsion_1d_30_petit.gif. **License**: *CC BY-SA: Attribution-ShareAlike*
- Ondes cisaillement 2d 20 (s-wave). **Authored by**: Christophe Dang Ngoc Chan. **Located at**: https://en.Wikipedia.org/wiki/File:Ondes_cisaillement_2d_20_petit.gif. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.8.5: Body Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.6: Surface Waves

There are two types of surface waves, Rayleigh waves and Love waves. Rayleigh waves are named after Lord Rayleigh (John Strutt), an English aristocrat who, in his work as a scientist and mathematician, developed a detailed mathematical accounting of the type of surface wave named after him. Rayleigh waves are set off by the combined effect of P- and S-waves on the earth's surface. Rayleigh waves are sometimes called rolling waves. In Rayleigh waves the surface of the earth rises up and sinks down in crests and troughs, similar to waves on the surface of water. People who are outdoors during a major earthquake commonly see Rayleigh waves moving across the surface of the earth, and can feel the ground rising and falling as the waves pass beneath them.

Love waves, sometimes called L-waves, are named after Augustus Love, an English mathematician and physicist who first modeled them mathematically. Love waves involve the surface shearing sideways and then returning to its original form as each wave passes.

All surface waves travel slower than body waves and Rayleigh waves are slower than Love waves.

Contributors and Attributions

CC licensed content, Shared previously

- Seismic waves. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/earthquakes.html#waves. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.8.6: Surface Waves is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.7: Measuring Earthquakes

People have always tried to quantify the size of and damage done by earthquakes. Since early in the 20th century, there have been three methods. What are the strengths and weaknesses of each?

- Mercalli Intensity Scale. Earthquakes are described in terms of what nearby residents felt and the damage that was done to nearby structures.
- Richter magnitude scale. Developed in 1935 by Charles Richter, this scale uses a seismometer to measure the magnitude of the largest jolt of energy released by an earthquake.
- Moment magnitude scale. Measures the total energy released by an earthquake. Moment magnitude is calculated from the area of the fault that is ruptured and the distance the ground moved along the fault.

The Richter scale and the moment magnitude scale are logarithmic.

- The amplitude of the largest wave increases ten times from one integer to the next.
- An increase in one integer means that thirty times more energy was released.
- These two scales often give very similar measurements.

How does the amplitude of the largest seismic wave of a magnitude 5 earthquake compare with the largest wave of a magnitude 4 earthquake? How does it compare with a magnitude 3 quake? The amplitude of the largest seismic wave of a magnitude 5 quake is 10 times that of a magnitude 4 quake and 100 times that of a magnitude 3 quake.

How does an increase in two integers on the moment magnitude scale compare in terms of the amount of energy released? Two integers equals a 900-fold increase in released energy.

Which scale do you think is best? With the Richter scale, a single sharp jolt measures higher than a very long intense earthquake that releases more energy. The moment magnitude scale more accurately reflects the energy released and the damage caused. Most seismologists now use the moment magnitude scale.

The way scientists measure earthquake intensity and the two most common scales, Richter and moment magnitude, are described along with a discussion of the 1906 San Francisco earthquake in *Measuring Earthquake*s video:

A link to an interactive elements can be found at the bottom of this page.

Contributors and Attributions

CC licensed content, Shared previously

- 7.3: Measuring and Predicting Earthquakes. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/7.3/. **License**: *CC BY-NC: Attribution-NonCommercial*

All rights reserved content

- Measuring Earthquakes. **Authored by**: NASASciFiles's channel. **Located at**: https://youtu.be/wtlu_aDteCA. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

This page titled 4.8.7: Measuring Earthquakes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.8: Magnitude versus Intensity

Magnitude and Intensity measure different characteristics of earthquakes. Magnitude measures the energy released at the source of the earthquake. **Magnitude** is determined from measurements on seismographs. **Intensity** measures the strength of shaking produced by the earthquake at a certain location. Intensity is determined from effects on people, human structures, and the natural environment.

Calculating Earthquake Magnitude

The magnitude of an earthquake is a number that allows earthquakes to be compared with each other in terms of their relative power. For several decades, earthquake magnitudes were calculated based on a method first developed by Charles Richter, a seismologist based in California. Richter used seismograms of earthquakes that occurred in the San Andreas fault zone to calibrate his magnitude scale.

Two measurements are factored together to determine the Richter magnitude of an earthquake: the amplitude of the largest waves recorded on a seismogram of the earthquake, and the distance to the epicenter of the earthquake. The maximum amplitude seismic wave – the height of the tallest one – is measured in mm on a seismogram. The distance to the epicenter must also be taken into account because the greater the distance from the earthquake, the smaller the waves get. The effect of distance is factored out of the calculation. There is no upper limit defined for the Richter scale, but after a century of seismograph measurements, it appears that rocks in the earth release their stress before building up enough energy to reach magnitude 10.

The Richter scale was found to not transfer very well from the San Andreas fault zone, a transform plate boundary, to the much more powerful earthquakes that occur at convergent plate boundaries, particularly subduction zone earthquakes. Therefore, the Richter scale has been replaced by the moment magnitude scale, symbolized as $\mathbf{M_w}$.

The moment magnitude scale is broadly similar to the Richter scale, but it takes more factors into account, including the total area of the fault that moves during the earthquake, and how much it moves. This produces a magnitude number that is a better indicator of the total amount of energy released by the earthquake. Because the moment magnitude scale has replaced the Richter scale, we will assume from here on that we are referring to moment magnitude, not Richter magnitude, when we speak of earthquake magnitude.

The magnitude scale portrays energy logarithmically to approximately base 32. For example, a magnitude 6.0 earthquake releases about 32 times as much energy as a magnitude 5.0 earthquake. A magnitude 7.0 releases about 32 × 32 = 1024 times as much energy as a magnitude 5.0 earthquake. A magnitude 9.0 earthquake, which rarely occurs, releases over a million times as much energy as a magnitude 5.0 earthquake.

Ranking Earthquake Intensity

Earthquake intensity is very different from earthquake magnitude. Earthquake intensity is a ranking based on the observed effects of an earthquake in each particular place. Therefore, each earthquake produces a range of intensity values, ranging from highest in the epicenter area to zero at a distance from the epicenter. The most commonly used earthquake intensity scale is the Modified Mercalli earthquake intensity scale. Refer to the Modified Mercalli Intensity Scale page on the US Geological Survey Earthquake Hazards Program website for an abbreviated version.

The table below shows approximately how many earthquakes occur each year in each magnitude range and what the intensity might be at the epicenter for each magnitude range.

Magnitude	Average number per year	Modified Mercalli Intensity	Description
0 – 1.9	>1 million	—	micro – not felt
2.0 – 2.9	>1 million	I	minor – rarely felt

3.0 – 3.9	about 100,000	II – III	minor – noticed by a few people
4.0 – 4.9	about 10,000	IV – V	light – felt by many people, minor damage possible
5.0 – 5.9	about 1,000	VI – VII	moderate – felt by most people, possible broken plaster and chimneys
6.0 – 6.9	about 130	VII – IX	strong – damage variable depending on building construction and substrate
7.0 – 7.9	about 15	IX – X	major – extensive damage, some buildings destroyed
8.0 – 8.9	about 1	X – XII	great – extensive damage over broad areas, many buildings destroyed
9.0 and above	< 1	XI – XII	great – extensive damage over broad areas, most buildings destroyed

Magnitude / Intensity Comparison

The following table gives intensities that are typically observed at locations near the epicenter of earthquakes of different magnitudes.

Magnitude	**Typical Maximum Modified Mercalli Intensity**
1.0 – 3.0	**I**
3.0 – 3.9	**II – III**
4.0 – 4.9	**IV – V**
5.0 – 5.9	**VI – VII**
6.0 – 6.9	**VII – IX**
7.0 and higher	**VIII** or higher

Abbreviated Modified Mercalli Intensity Scale

I. Not felt except by a very few under especially favorable conditions.
II. Felt only by a few persons at rest, especially on upper floors of buildings.
III. Felt quite noticeably by persons indoors, especially on upper floors of buildings. Many people do not recognize it as an earthquake. Standing motor cars may rock slightly. Vibrations similar to the passing of a truck. Duration estimated.
IV. Felt indoors by many, outdoors by few during the day. At night, some awakened. Dishes, windows, doors disturbed; walls make cracking sound. Sensation like heavy truck striking building. Standing motor cars rocked noticeably.
V. Felt by nearly everyone; many awakened. Some dishes, windows broken. Unstable objects overturned. Pendulum clocks may stop.
VI. Felt by all, many frightened. Some heavy furniture moved; a few instances of fallen plaster. Damage slight.
VII. Damage negligible in buildings of good design and construction; slight to moderate in well-built ordinary structures; considerable damage in poorly built or badly designed structures; some chimneys broken.
VIII. Damage slight in specially designed structures; considerable damage in ordinary substantial buildings with partial collapse. Damage great in poorly built structures. Fall of chimneys, factory stacks, columns, monuments, walls. Heavy furniture overturned.
IX. Damage considerable in specially designed structures; well-designed frame structures thrown out of plumb. Damage great in substantial buildings, with partial collapse. Buildings shifted off foundations.
X. Some well-built wooden structures destroyed; most masonry and frame structures destroyed with foundations. Rails bent.
XI. Few, if any (masonry) structures remain standing. Bridges destroyed. Rails bent greatly.
XII. Damage total. Lines of sight and level are distorted. Objects thrown into the air.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Earthquakes. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/earthquakes.html#seismlgy. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

Public domain content

- Magnitude / Intensity Comparison. **Provided by**: USGS. **Located at**: http://earthquake.usgs.gov/learn/topics/mag_vs_int.php. **License**: *Public Domain: No Known Copyright*

This page titled 4.8.8: Magnitude versus Intensity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.9: Earthquake Hazards

Earthquakes can be hazardous to humans and property in a variety of ways. Earthquake hazards arise from a combination of factors such as the size of the earthquake, distance to the epicenter, the underlying material and geologic structures, and building construction.

Potential Hazards

Ground Shaking

Ground shaking is caused by seismic waves During a significant earthquake, a particular location, and any building at that location, will be shaken by body waves (P- and S-) and surface waves (Rayleigh and Love. Each type of wave will have a different frequency (different number of waves passing by each second), and can shear or move a building in a different way, sometimes simultaneously. Adjacent neighborhoods or towns may experience very different intensities from the same earthquake, based on how far they are from the epicenter and what sort of rocks or sediments are in the ground beneath each area. Places underlain by thick deposits of unconsolidated sediments will experience a higher amplitude of shaking, at the same distance from the same earthquake, than places underlain by solid bedrock all the way to the surface. If unconsolidated sediments are fine-grained and wet, they may undergo liquefaction, increasing the damage to buildings and infrastructure and therefore increasing the intensity of the earthquake there. If unconsolidated sediments are overlain by a layer of artificial fill, the area is likely to experience more intense shaking and undergo more ground settling, and liquefaction if wet, than places that have not had a layer of artificial fill added.

Mexico City, one of the most populated cities in the world, is in a basin in the mountains of Mexico. Much of the city is built on artificial fill on top of fine-grained sediments from an extensive lake and wetlands that were drained and filled in as the city grew. As a result of how seismic waves are amplified in soft sediments, the shaking of the ground in Mexico City during an earthquake is greater than it is in areas outside the basin, which have bedrock close to the surface. During the 1985 earthquake, which originated offshore of Acapulco on the Pacific coast, 300 km (200 miles) away, many buildings in Mexico City collapsed and more than 20,000 people died.

The Marina District in northern San Francisco is built on artificial fill on wet bayshore sediments. Rubble and debris from buildings that collapsed or burned in that district as a result of the 1906 earthquake was used as artificial fill beneath structures built during the reconstruction of that area. In 1989, the Loma Prieta earthquake caused collapse of several buildings in the Marina District, and several people died there. Another example of a structure on wet ground covered by artificial fill that collapsed during the Loma Prieta earthquake is the Cypress viaduct in Oakland, CA. The USGS photo below shows the failed support columns.

Permanent Ground Displacement

During large earthquakes, the ground may permanently shift to a new position up, down, or sideways (up to 10 or more m, 30 or 40 ft, in extreme cases). This change in the location of the ground, which also tilts the ground, may cause disruption of roads and utilities and, in coastal cities, submergence or emergence of harbor facilities. Even ground shifts of less than a meter (a foot or two) can cause serious disruption to infrastructure.

Earth Rupture

During most earthquakes, some rupturing of the earth's surface takes place along the fault trace. This produces a fault scarp, which may have up to several m (up to 10 ft or more) of vertical displacement. This can disrupt roads and utilities, and any buildings on a fault that ruptures may undergo extensive damage.

Rupture of the earth during an earthquake may also occur on secondary faults. Earthquake-induced rupture of the earth's surface may also take place in weak zones of surface sediment that fracture and spread. If enough spreading of a ruptured surface layer takes place, it can be classified as a landslide. The NOAA photo below shows substantial damage in the Turnagain-By-The-Sea subdivision caused by the 1964 Great Alaskan Earthquake.

Landslides and Avalanches

On steep slopes and in mountainous areas, large earthquakes can set off many landslides, rockfalls, or avalanches. These can damage buildings, towns, or roads in the path of the landslides.

Liquefaction

If fine- or medium-grained, unconsolidated sediments are saturated with groundwater, the shaking that occurs during an earthquake may cause the sediment grains to lose contact with each other and become suspended in the water, temporarily turning what was solid ground into liquid ground. Building and other structures may sink, tilt, or slide a short distant in liquefied ground, causing serious damage.

Fires

Fires are a secondary rather than a primary effect of earthquakes. Broken electrical wires and natural gas pipes commonly set off fires during earthquakes. To compound the problem, the water supply may also be disrupted by earthquake damage, making it impossible to put the fire out with water from fire hydrants. The fire that broke out as a result of the great San Francisco earthquake of 1906 burned much of the city to the ground, causing more extensive damage to buildings than the shaking of the ground did during the earthquake.

Tsunamis

A tsunami is a set of waves in the ocean (or a large lake) with an extremely long wavelength, typically over 100 km long. Tsunamis move at many 100s of km per hour in deep water. Tsunamis can be set off by violent volcanic eruptions originating just below sea level, by giant landslides that either occur underwater or tumble into the sea from coastal mountains, by large meteorite impacts, and, most commonly, by earthquakes that greatly shake the ocean floor, which commonly happens at subduction zones.

The amplitude, or crest height, of an individual tsunami wave may be only about 1 m (roughly 3 feet high) in the open ocean. It is common for a tsunami to pass ships at sea without being noticed. However, as the wave approaches shore where the bottom grows shallower, the crest builds up to a height of up to several tens of meters (over 30 feet in some cases). The wave crest may wash ashore for several minutes before subsiding. Even a tsunami wave of no higher than 3 m (10 feet) coming ashore can cause extensive damage in harbors and to shores, as the long-wavelength wave keeps pouring in for several minutes.

Tsunamis consist of more than one wave, so a second wave crest may climb ashore several minutes later. Some tsunamis lead with the wave trough, so the first thing noticed as that type of tsunami approaches a shore is a dramatic drawback, or retreat, of the sea, like the tide suddenly going out. Such a drawback will inevitably be followed by a rising tsunami wave crest.

The tsunamis from the Sumatra earthquake of 2004 killed over 100,000 people in coastal areas of the Indian Ocean, some on shores several thousand km (several thousand miles) away from the epicenter of the earthquake. The great southeast Alaska earthquake of 1964 generated a tsunami that killed 16 people on the coast of Oregon and northern California, over 1,000 km (600 miles) away. Tsunamis have been known to cross the entire Pacific Ocean and cause fatalities a third of the way around the world. Hawaii, in the middle of the Pacific Ocean, has been damaged by tsunamis, originating from subduction earthquakes on the Pacific Rim, several times in the last few centuries. Japan, with its complex set of subduction zones and its eastern shoreline open to the Pacific Ocean, has experienced over 100 tsunamis in its recorded history, most recently as a result of a great subduction earthquake off the northern island in 2011 which resulted in several coastal towns being destroyed and thousands of people dying. The last great tsunami generated by the Cascadia subduction zone along the coast of the Pacific Northwest flooded coastal areas of Washington, Oregon northern California, and Vancouver Island in Canada. It occurred in 1701, long before the region was as populated as it is now. Subduction continues in the Cascadia subduction zone and more tsunamis can be expected to be generated by earthquakes there in the future.

Mitigating Earthquake Damage

Much can be done to reduce the risk of fatalities during earthquakes, and to reduce the damage to buildings and infrastructure; in other words, to mitigate the effects of earthquakes.

In many cases, it is collapsed buildings that cause the most harm during an earthquake. Buildings should be constructed in ways

that make them unlikely to collapse during an earthquake. The strategies that engineers have developed include having sufficient flexibility in the structure to absorb shaking during an earthquake. Bricks, mortar, and concrete are rigid and brittle. However, bricks and mortar, and concrete, can be reinforced with steel to make them better able to survive an earthquake. Wood and steel are more flexible than bricks, mortar, and concrete, and lend themselves to the type of building that, properly designed and built according to code, is likely to survive an earthquake without collapsing.

The way a building is attached to its foundation, and how the foundation is anchored in the earth, are important considerations in earthquake design. Many houses built in the the early and mid-1900s in California were not attached to their foundations, based on the assumption that the weight of a house would keep it on its foundation. It turned out to be a bad assumption. Earthquakes caused houses to slide off their foundations. Many home-owners in the state have taken steps to make sure that their houses are now attached to their foundations; if a home-owner buys earthquake insurance, it is usually required by the insurance company that they do so. A large building or skyscraper built in an earthquake-prone area will normally have a great deal of flexibility built into it, including some sort of elastic strain absorption mechanism focused on points where the building attaches to its foundation.

Infrastructure—roads, bridges, utilities—can be built with margins of safety for the event of an earthquake. This includes gas pipelines designed to slide back and forth on their supports and having built-in shutoff valves that may be activated by automatic sensors, electric lines and grids with similar flexibility and shut-off capabilities, and roads, overpasses, and bridges built to withstand shaking during an earthquake.

Development and enforcement of building codes aimed at reducing risk from earthquakes often requires resources that are not available in impoverished regions. This leads to a higher likelihood that buildings will collapse from the same size earthquake in some areas of the world than in other areas.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Earthquakes. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/earthquakes.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.8.9: Earthquake Hazards is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.10: Earthquake Prediction

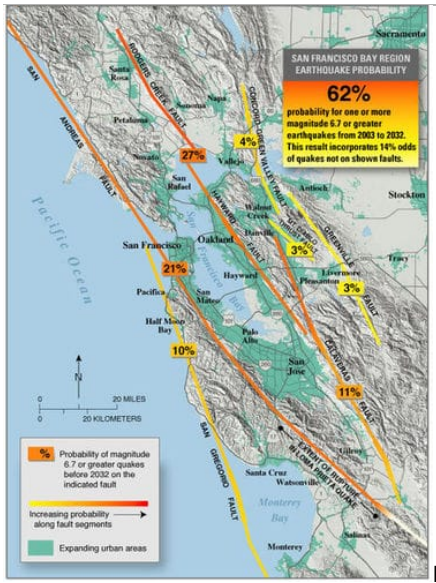


Figure 1. The probabilities of earthquakes striking along various faults in the San Francisco area between 2003 (when the work was done) and 2032.

Scientists are a long way from being able to predict earthquakes. A good prediction must be accurate as to where an earthquake will occur, when it will occur, and at what magnitude it will be so that people can evacuate. An unnecessary evacuation is expensive and causes people not to believe authorities the next time an evacuation is ordered.

Where an earthquake will occur is the easiest feature to predict. Scientists know that earthquakes take place at plate boundaries and tend to happen where they've occurred before (figure 1). Earthquake-prone communities should always be prepared for an earthquake. These communities can implement building codes to make structures earthquake safe.

When an earthquake will occur is much more difficult to predict. Since stress on a fault builds up at the same rate over time, earthquakes should occur at regular intervals (figure 2). But so far scientists cannot predict when quakes will occur even to within a few years.

Signs sometimes come before a large earthquake. Small quakes, called foreshocks, sometimes occur a few seconds to a few weeks before a major quake. However, many earthquakes do not have foreshocks and small earthquakes are not necessarily followed by a large earthquake. Often, the rocks around a fault will dilate as microfractures form. Ground tilting, caused by the buildup of stress in the rocks, may precede a large earthquake, but not always. Water levels in wells fluctuate as water moves into or out of fractures before an earthquake. This is also an uncertain predictor of large earthquakes. The relative arrival times of P-waves and S-waves also decreases just before an earthquake occurs.

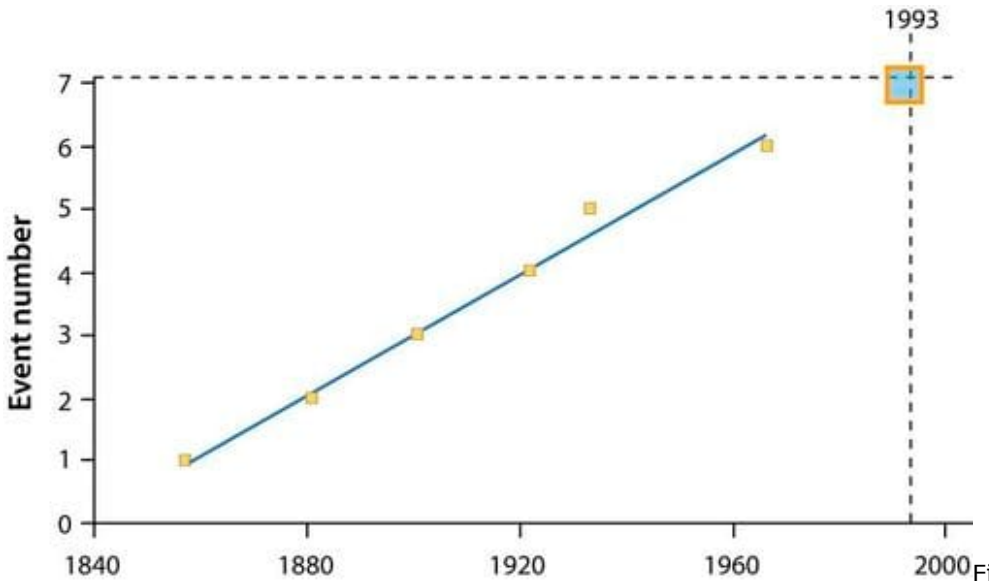


Figure 2. Around Parkfield, California, an earthquake of magnitude 6.0 or higher occurs about every 22 years. So seismologists predicted that one would strike in 1993, but that quake came in 2004—11 years late.

Folklore tells of animals behaving erratically just before an earthquake. Mostly these anecdotes are told after the earthquake. If indeed animals sense danger from earthquakes or tsunami, scientists do not know what it is they could be sensing, but they would like to find out.

Contributors and Attributions

CC licensed content, Shared previously

- 7.3: Measuring and Predicting Earthquakes. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/7.3/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.8.10: Earthquake Prediction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.8.11: Seismicity and Earthquake Prediction

Seismicity is the study of how often earthquakes occur in a particular area, which types of earthquakes occur there, and why.

In the United States, the areas that most frequently experience earthquakes are the coast of California, Oregon, and Washington, the southern coast and Aleutian Islands of Alaska, Hawaii, and the mountain west from the Rocky Mountains to the Pacific coast. The central and eastern United States rarely experience significant earthquakes.

Earthquake epicenters compiled on a map show that, on a global basis, most earthquakes occur around the rim of the Pacific Ocean, in the mountains of southern Asia from China to the Middle East, and in the Mediterranean Sea area. Earthquake epicenters also trace the mid-ocean ridges across the floors of the oceans.

Because nearly all earthquakes occur on faults, determining seismic risks on a finer scale largely consists of identifying, mapping, and studying active faults in each state or region. However, many active faults are hidden, either because any scarps they formed at the surface have been eroded or covered by sediments, soil, and vegetation, or because they are blind faults. A hidden fault is often not identified and located until one or more significant earthquakes has occurred on it and the seismic waves have been studied to determine its location and type of fault motion.

Information used to determine the seismicity of an area includes:

- frequency of earthquakes in the past, as deduced from:
  - historic records
  - geologic studies that examine evidence of the prehistoric earthquake record
- location of known active faults
- seismologic data collected on recent earthquakes that have occurred in the area
- tectonic setting of the area in terms of proximity to plate boundaries, and information about the plate boundary if one is nearby
- stress and strain being experienced by the crust in that area based on measurements from GPS equipment and from stress and strain measurements conducted in boreholes
- underground geologic layers and structures in that area based on cross-sections from geologic mapping, data from drilling, and remote imaging of deeper layers of the crust and mantle

Based on this information, the seismic risk of a particular area can be quantified statistically. For example, the odds of a major earthquake happening in the next century, or in the next 10 years, can be estimated for a specific seismic zone.

However, no scientific method has yet been developed that can predict precisely when the next earthquake in a specific region will happen, where it will happen, or what its magnitude will be. Scientists have looked into using such possible pre-earthquake indicators as ground tilting, changes in well water levels, changes in radon gas in groundwater near fault zones, changes in electrical conductivity in the earth around faults, changes or patterns in seismic activity that can be measured by seismometers even though it is not felt by humans, and strange animal behavior which, according to numerous, largely unconfirmed anecdotes, takes place before an earthquake. But, so far, none of these types of data have been found to lead to reliable earthquake predictions.

People have also looked into correlations between earthquakes and phases of the Moon, earthquakes and the time of day (such as dawn when the Sun is first shining on the ground), and so on. No connections have been found between earthquakes and these other types of phenomena.

As the study of seismicity stands now, we can identify which areas on earth will undergo major earthquakes in the coming decades and centuries, we can delineate which areas on earth are at risk for the most powerful types of earthquakes, and map the coastal areas that are most at risk of being inundated by a tsunami, but we cannot pinpoint in advance the date or location of the next major earthquake.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Earthquakes. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College.

Located at: http://commons.wvc.edu/rdawes/G101OCL/Basics/earthquakes.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.8.11: Seismicity and Earthquake Prediction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.9: VOLCANOES

Learning Objectives

Explain the origin of volcanic features and landforms; as well as the causes and hazards associated with volcanic activity.

- Describe and compare different volcano types.
- Describe the processes that form volcanoes.
- Classify and identify igneous features.
- Describe major volcanic hazards and their potential effects.
- Identify the various methods used to monitor volcanic activity.

4.9.1: Types of Volcanoes

4.9.2: Types of Eruptions

4.9.3: Supervolcanoes

4.9.4: Volcanoes at Plate Boundaries

4.9.5: Volcanoes Hotspots

4.9.6: Pyroclastic Deposits

4.9.7: Dikes and Sills

4.9.8: Volcanic Landforms

4.9.9: Igneous Features

4.9.10: Volcanic Gases, Pyroclastic Flow, and Tephra

4.9.11: Lahars, Landslides, and Lavaflow

4.9.12: Signs of an Eruption

4.9.13: Tools to Monitor Volcanoes

This page titled Chapter 4.9: Volcanoes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.1: Types of Volcanoes

A volcano is a vent through which molten rock and gas escape from a magma chamber. Volcanoes differ in many features such as height, shape, and slope steepness. Some volcanoes are tall cones and others are just cracks in the ground (figure1). As you might expect, the shape of a volcano is related to the composition of its magma.



Figure 1. Mount St. Helens was a beautiful, classic, cone-shaped volcano. The volcano's 1980 eruption blew more than 400 meters (1,300 feet) off the top of the mountain.

Composite Volcanoes

Composite volcanoes are made of felsic to intermediate rock. The viscosity of the lava means that eruptions at these volcanoes are often explosive (figure 2).



Figure 2. Mt. Fuji, the highest mountain in Japan, is a dormant composite volcano.

The viscous lava cannot travel far down the sides of the volcano before it solidifies, which creates the steep slopes of a composite volcano. Viscosity also causes some eruptions to explode as ash and small rocks. The volcano is constructed layer by layer, as ash and lava solidify, one upon the other (figure 3). The result is the classic cone shape of composite volcanoes.

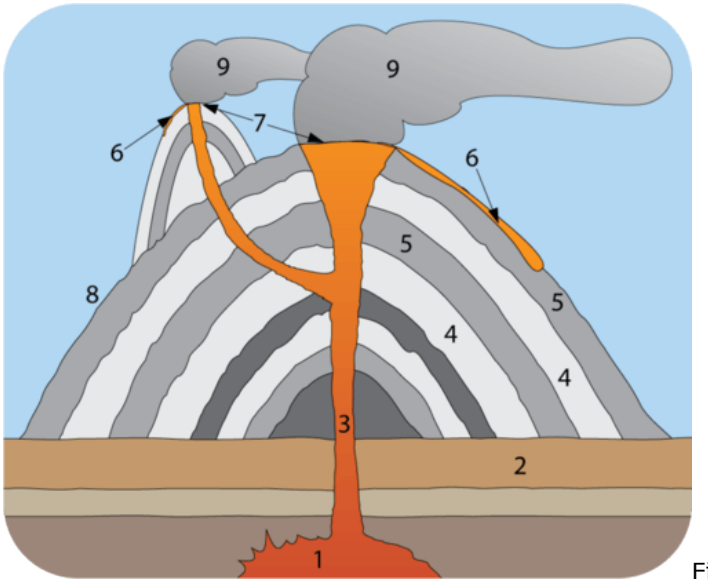


Figure 3. A cross section of a composite volcano reveals alternating layers of rock and ash: (1) magma chamber, (2) bedrock, (3) pipe, (4) ash layers, (5) lava layers, (6) lava flow, (7) vent, (8) lava, (9) ash cloud. Frequently there is a large crater at the top from the last eruption.

Shield Volcanoes

Shield volcanoes get their name from their shape. Although shield volcanoes are not steep, they may be very large. Shield volcanoes are common at spreading centers or intraplate hot spots (figure 4).


Figure 4. Mauna Loa Volcano in Hawaii (in the background) is the largest shield volcano on Earth with a diameter of more than 112 kilometers (70 miles). The volcano forms a significant part of the island of Hawaii.

The lava that creates shield volcanoes is fluid and flows easily. The spreading lava creates the shield shape. Shield volcanoes are built by many layers over time and the layers are usually of very similar composition. The low viscosity also means that shield eruptions are non-explosive.

This *Volcanoes 101* video from National Geographic discusses where volcanoes are found and what their properties come from:

A link to an interactive elements can be found at the bottom of this page.

Cinder Cones



Figure 5. In 1943, a Mexican farmer first witnessed a cinder cone erupting in his field. In a year, Paricutín was 336 meters high. By 1952, it reached 424 meters and then stopped erupting.

Cinder cones are the most common type of volcano. A cinder cone has a cone shape, but is much smaller than a composite volcano. Cinder cones rarely reach 300 meters in height but they have steep sides. Cinder cones grow rapidly, usually from a single eruption cycle (figure 5). Cinder cones are composed of small fragments of rock, such as pumice, piled on top of one another. The rock shoots up in the air and doesn't fall far from the vent. The exact composition of a cinder cone depends on the composition of the lava ejected from the volcano. Cinder cones usually have a crater at the summit.

Cinder cones are often found near larger volcanoes (figure 6).

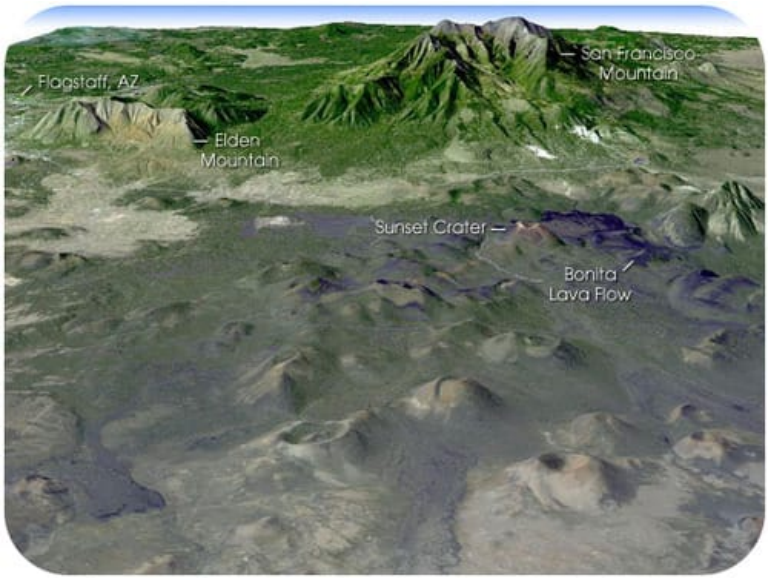


Figure 6. This Landsat image shows the topography of San Francisco Mountain, an extinct volcano, with many cinder cones near it in northern Arizona. Sunset crater is a cinder cone that erupted about 1,000 years ago.

Lesson Summary

- Composite, shield, cinder cones, and supervolcanoes are the main types of volcanoes.
- Composite volcanoes are tall, steep cones that produce explosive eruptions.
- Shield volcanoes form very large, gently sloped mounds from effusive eruptions.
- Cinder cones are the smallest volcanoes and result from accumulation of many small fragments of ejected material.
- An explosive eruption may create a caldera, a large hole into which the mountain collapses.

Contributors and Attributions

CC licensed content, Shared previously

- 8.3: Types of Volcanoes. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/8.3/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.9.1: Types of Volcanoes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.2: Types of Eruptions



Figure 1. Some of the eruptive structures formed during volcanic activity: a Plinian eruption column, Hawaiian pahoehoe flows, and a lava arc from a Strombolian eruption.

Several **types ofvolcanic eruptions**—during which lava, tephra (ash, lapilli, volcanic bombs and blocks), and assorted gases are expelled from a volcanic vent or fissure—have been distinguished by volcanologists. These are often named after famous volcanoes where that type of behavior has been observed. Some volcanoes may exhibit only one characteristic type of eruption during a period of activity, while others may display an entire sequence of types all in one eruptive series.

There are three different types of eruptions. The most well-observed are magmatic eruptions, which involve the decompression of gas within magma that propels it forward. Phreatomagmatic eruptions are another type of volcanic eruption, driven by the compression of gas within magma, the direct opposite of the process powering magmatic activity. The third eruptive type is the phreatic eruption, which is driven by the superheating of steam via contact with magma; these eruptive types often exhibit no magmatic release, instead causing the granulation of existing rock.

Within these wide-defining eruptive types are several subtypes. The weakest are Hawaiian and submarine, then Strombolian, followed by Vulcanian and Surtseyan. The stronger eruptive types are Pelean eruptions, followed by Plinian eruptions; the strongest eruptions are called “Ultra-Plinian.” Subglacial and phreatic eruptions are defined by their eruptive mechanism, and vary in strength. An important measure of eruptive strength is Volcanic Explosivity Index (VEI), an order of magnitude scale ranging from 0 to 8 that often correlates to eruptive types.

Eruption Mechanisms

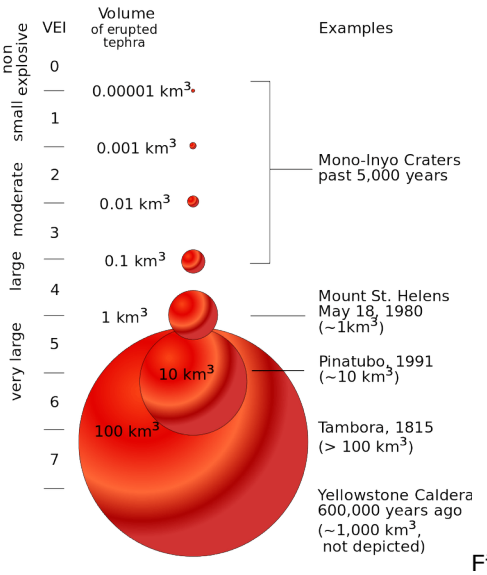


Figure 2. Diagram showing the scale of VEI correlation with total ejecta volume.

Volcanic eruptions arise through three main mechanisms:[1]

- Gas release under decompression causing magmatic eruptions
- Thermal contraction from chilling on contact with water causing phreatomagmatic eruptions
- Ejection of entrained particles during steam eruptions causing phreatic eruptions

There are two types of eruptions in terms of activity, explosive eruptions and effusive eruptions. Explosive eruptions are characterized by gas-driven explosions that propels magma and tephra.[2] Effusive eruptions, meanwhile, are characterized by the outpouring of lava without significant explosive eruption.[3]

Volcanic eruptions vary widely in strength. On the one extreme there are effusive Hawaiian eruptions, which are characterized by lava fountains and fluid lava flows, which are typically not very dangerous. On the other extreme, Plinian eruptions are large, violent, and highly dangerous explosive events. Volcanoes are not bound to one eruptive style, and frequently display many different types, both passive and explosive, even the span of a single eruptive cycle.[4] Volcanoes do not always erupt vertically from a single crater near their peak, either. Some volcanoes exhibit lateral and fissure eruptions. Notably, many Hawaiian eruptions start from rift zones,[5] and some of the strongest Surtseyan eruptions develop along fracture zones.[6] Scientists believed that pulses of magma mixed together in the chamber before climbing upward—a process estimated to take several thousands of years. But Columbia University volcanologists found that the eruption of Costa Rica's Irazú Volcano in 1963 was likely triggered by magma that took a nonstop route from the mantle over just a few months.[7]

Volcano Explosivity Index

The volcanic explosivity index (commonly shortened to VEI) is a scale, from 0 to 8, for measuring the strength of eruptions. It is used by the Smithsonian Institution's Global Volcanism Program in assessing the impact of historic and prehistoric lava flows. It operates in a way similar to the Richter scale for earthquakes, in that each interval in value represents a tenfold increasing in magnitude (it is logarithmic).[8] The vast majority of volcanic eruptions are of VEIs between 0 and 2.[9]

Volcanic eruptions by VEI index[10]

VEI	Plume height	Eruptive volume*	Eruption type	Frequency**	Example
0	<100 m (330 ft)	1,000 m^3 (35,300 cu ft)	Hawaiian	Continuous	Kilauea
1	100–1,000 m (300–3,300 ft)	10,000 m^3 (353,000 cu ft)	Hawaiian/ Strombolian	Fortnightly	Stromboli

Volcanic eruptions by VEI index[10]					
2	1–5 km (1–3 mi)	1,000,000 m^3 (35,300,000 cu ft)†	Strombolian/ Vulcanian	Monthly	Galeras (1992)
3	3–15 km (2–9 mi)	10,000,000 m^3 (353,000,000 cu ft)	Vulcanian	3 monthly	Nevado del Ruiz (1985)
4	10–25 km (6–16 mi)	100,000,000 m^3 (0.024 cu mi)	Vulcanian/Peléan	18 months	Eyjafjallajökull (2010)
5	>25 km (16 mi)	1 km^3 (0.24 cu mi)	Plinian	10–15 years	Mount St. Helens (1980)
6	>25 km (16 mi)	10 km^3 (2 cu mi)	Plinian/Ultra-Plinian	50–100 years	Krakatoa (1883)
7	>25 km (16 mi)	100 km^3 (20 cu mi)	Ultra-Plinian	500–1000 years	Tambora (1815)
8	>25 km (16 mi)	1,000 km^3 (200 cu mi)	Supervolcanic	50,000+ years[11]	Lake Toba (74 ka)

* This is the minimum eruptive volume necessary for the eruption to be considered within the category.
** Values are a rough estimate. They indicate the frequencies for volcanoes of that magnitude OR HIGHER
† There is a discontinuity between the 1st and 2nd VEI level; instead of increasing by a magnitude of 10, the value increases by a magnitude of 100 (from 10,000 to 1,000,000).

Magmatic Eruptions

Magmatic eruptions produce juvenile clasts during explosive decompression from gas release. They range in intensity from the relatively small lava fountains on Hawaii to catastrophic Ultra-Plinian eruption columns more than 30 km (19 mi) high, bigger than the eruption of Mount Vesuvius in 79 that buried Pompeii.[12]

Hawaiian

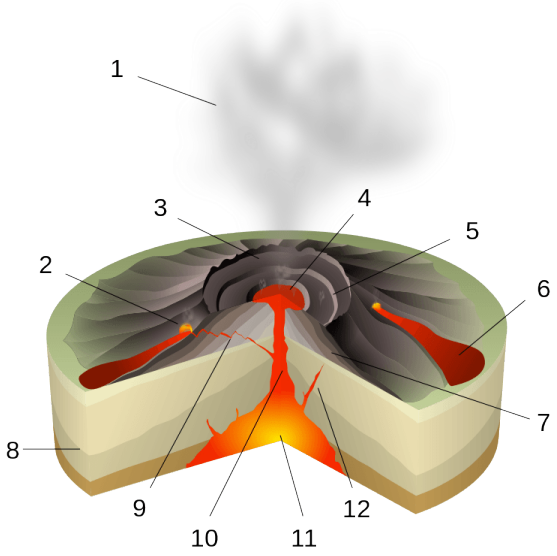
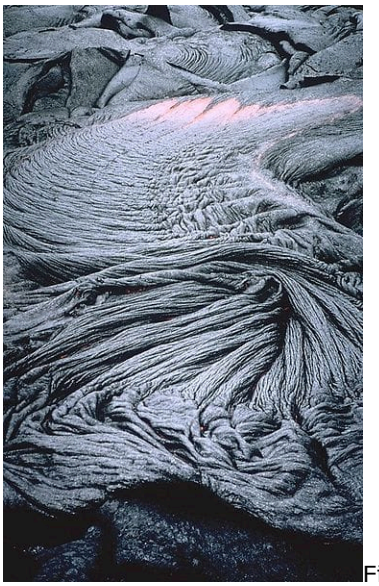


Figure 3. Diagram of a Hawaiian eruption. (key: 1. Ash plume 2. Lava fountain 3. Crater 4. Lava lake 5. Fumaroles 6. Lava flow 7. Layers of lava and ash 8. Stratum 9. Sill 10. Magma conduit 11. Magma chamber 12. Dike) Click for larger version.

Hawaiian eruptions are a type of volcanic eruption, named after the Hawaiian volcanoes with which this eruptive type is hallmark. Hawaiian eruptions are the calmest types of volcanic events, characterized by the effusive eruption of very fluid basalt-type lavas with low gaseous content. The volume of ejected material from Hawaiian eruptions is less than half of that found in other eruptive

types. Steady production of small amounts of lava builds up the large, broad form of a shield volcano. Eruptions are not centralized at the main summit as with other volcanic types, and often occur at vents around the summit and from fissure vents radiating out of the center.[13]

Hawaiian eruptions often begin as a line of vent eruptions along a fissure vent, a so-called “curtain of fire.” These die down as the lava begins to concentrate at a few of the vents. Central-vent eruptions, meanwhile, often take the form of large lava fountains (both continuous and sporadic), which can reach heights of hundreds of meters or more. The particles from lava fountains usually cool in the air before hitting the ground, resulting in the accumulation of cindery scoria fragments; however, when the air is especially thick with clasts, they cannot cool off fast enough due to the surrounding heat, and hit the ground still hot, the accumulation of which forms spatter cones. If eruptive rates are high enough, they may even form splatter-fed lava flows. Hawaiian eruptions are often extremely long lived; Puʻu ʻŌʻō, a cinder cone of Kilauea, has been erupting continuously since 1983. Another Hawaiian volcanic feature is the formation of active lava lakes, self-maintaining pools of raw lava with a thin crust of semi-cooled rock; there are currently only 5 such lakes in the world, and the one at Kīlauea’s Kupaianaha vent is one of them.[14]


Figure 4. Ropey pahoehoe lava from Kilauea, Hawaiʻi.

Flows from Hawaiian eruptions are basaltic, and can be divided into two types by their structural characteristics. Pahoehoe lava is a relatively smooth lava flow that can be billowy or ropey. They can move as one sheet, by the advancement of “toes,” or as a snaking lava column. A’a lava flows are denser and more viscous then pahoehoe, and tend to move slower. Flows can measure 2 to 20 m (7 to 66 ft) thick. A’a flows are so thick that the outside layers cools into a rubble-like mass, insulating the still-hot interior and preventing it from cooling. A’a lava moves in a peculiar way—the front of the flow steepens due to pressure from behind until it breaks off, after which the general mass behind it moves forward. Pahoehoe lava can sometimes become A’a lava due to increasing viscosity or increasing rate of shear, but A’a lava never turns into pahoehoe flow.[15]

Volcanoes known to have Hawaiian activity include:[16]

- Puʻu ʻŌʻō, a parasitic cinder cone located on Kilauea on the island of Hawaiʻi which has been erupting continuously since 1983. The eruptions began with a 6 km (4 mi)-longfissure-based “curtain of fire” on 3 January. These gave way to centralized eruptions on the site of Kilauea’s east rift, eventually building up the still active cone.
- For a list of all of the volcanoes of Hawaii, see List of volcanoes in the Hawaiian–Emperor seamount chain.
- Mount Etna, Italy.
- Mount Mihara in 1986 (see above paragraph)

Strombolian

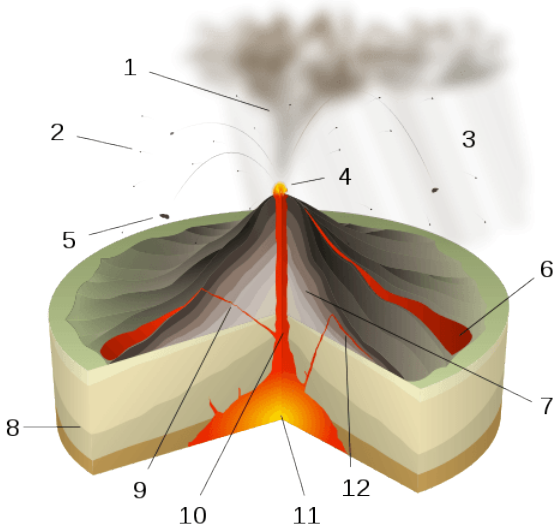


Figure 5. Diagram of a Strombolian eruption. (key: 1. Ash plume 2. Lapilli 3. Volcanic ash rain 4. Lava fountain 5. Volcanic bomb 6. Lava flow 7. Layers of lava and ash 8. Stratum 9. Dike 10. Magma conduit 11. Magma chamber 12. Sill) Click for larger version.

Strombolian eruptions are a type of volcanic eruption, named after the volcano Stromboli, which has been erupting continuously for centuries.[17] Strombolian eruptions are driven by the bursting of gas bubbles within the magma. These gas bubbles within the magma accumulate and coalesce into large bubbles, called gas slugs. These grow large enough to rise through the lava column.[18] Upon reaching the surface, the difference in air pressure causes the bubble to burst with a loud pop,[19] throwing magma in the air in a way similar to a soap bubble. Because of the high gas pressures associated with the lavas, continued activity is generally in the form of episodic explosive eruptions accompanied by the distinctive loud blasts.[20] During eruptions, these blasts occur as often as every few minutes.[21]

The term “Strombolian” has been used indiscriminately to describe a wide variety of volcanic eruptions, varying from small volcanic blasts to large eruptive columns. In reality, true Strombolian eruptions are characterized by short-lived and explosive eruptions of lavas with intermediate viscosity, often ejected high into the air. Columns can measure hundreds of meters in height. The lavas formed by Strombolian eruptions are a form of relatively viscous basaltic lava, and its end product is mostly scoria.[22] The relative passivity of Strombolian eruptions, and its non-damaging nature to its source vent allow Strombolian eruptions to continue unabated for thousands of years, and also makes it one of the least dangerous eruptive types.[23]

Figure 6. An example of the lava arcs formed during Strombolian activity. This image is of Stromboli itself.

Strombolian eruptions eject volcanic bombs and lapilli fragments that travel in parabolic paths before landing around their source vent. The steady accumulation of small fragments builds cinder cones composed completely of basaltic pyroclasts. This form of accumulation tends to result in well-ordered rings of tephra.[24]

Strombolian eruptions are similar to Hawaiian eruptions, but there are differences. Strombolian eruptions are noisier, produce no sustained eruptive columns, do not produce some volcanic products associated with Hawaiian volcanism (specifically Pele’s tears and Pele’s hair), and produce fewer molten lava flows (although the eruptive material does tend to form small rivulets).[25]

Volcanoes known to have Strombolian activity include:

- Parícutin, Mexico, which erupted from a fissure in a cornfield in 1943. Two years into its life, pyroclastic activity began to wane, and the outpouring of lava from its base became its primary mode of activity. Eruptions ceased in 1952, and the final height was 424 m (1,391 ft). This was the first time that scientists are able to observe the complete life cycle of a volcano.[26]
- Mount Etna, Italy, which has displayed Strombolian activity in recent eruptions, for example in 1981, 1999,[27] 2002-2003,

and 2009.[28]

- Mount Erebus in Antarctica, the southernmost active volcano in the world, having been observed erupting since 1972.[29] Eruptive activity at Erebus consists of frequent Strombolian activity.[30]
- Stromboli itself. The namesake of the mild explosive activity that it possesses has been active throughout historical time; essentially continuous Strombolian eruptions, occasionally accompanied by lava flows, have been recorded at Stromboli for more than a millennium.[31]

Vulcanian

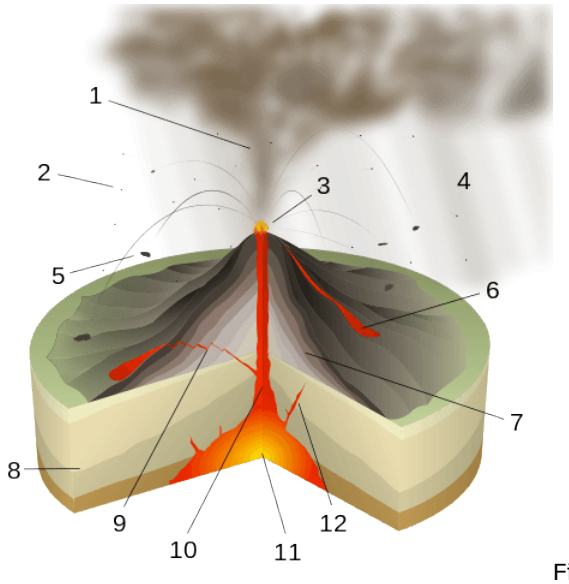


Figure 7. Diagram of a Vulcanian eruption. (key: 1. Ash plume 2. Lapilli 3. Lava fountain 4. Volcanic ash rain 5. Volcanic bomb 6. Lava flow 7. Layers of lava and ash 8. Stratum 9. Sill 10. Magma conduit 11. Magma chamber 12. Dike)

Vulcanian eruptions are a type of volcanic eruption, named after the volcano Vulcano.[32] It was named so following Giuseppe Mercalli's observations of its 1888-1890 eruptions.[33] In Vulcanian eruptions, highly viscous magma within the volcano make it difficult for vesiculate gases to escape. Similar to Strombolian eruptions, this leads to the buildup of high gas pressure, eventually popping the cap holding the magma down and resulting in an explosive eruption. However, unlike Strombolian eruptions, ejected lava fragments are not aerodynamic; this is due to the higher viscosity of Vulcanian magma and the greater incorporation of crystalline material broken off from the former cap. They are also more explosive than their Strombolian counterparts, with eruptive columns often reaching between 5 and 10 km (3 and 6 mi) high. Lastly, Vulcanian deposits are andesitic to dacitic rather than basaltic.[34]



Figure 8. Tavurvur in Papua New Guinea erupting.

Volcanoes that have exhibited Vulcanian activity include:

- Sakurajima, Japan has been the site of Vulcanian activity near-continuously since 1955.[35]
- Tavurvur, Papua New Guinea, one of several volcanoes in the Rabaul Caldera.[36]
- Irazú Volcano in Costa Rica exhibited Vulcanian activity in its 1965 eruption.[37]

Peléan

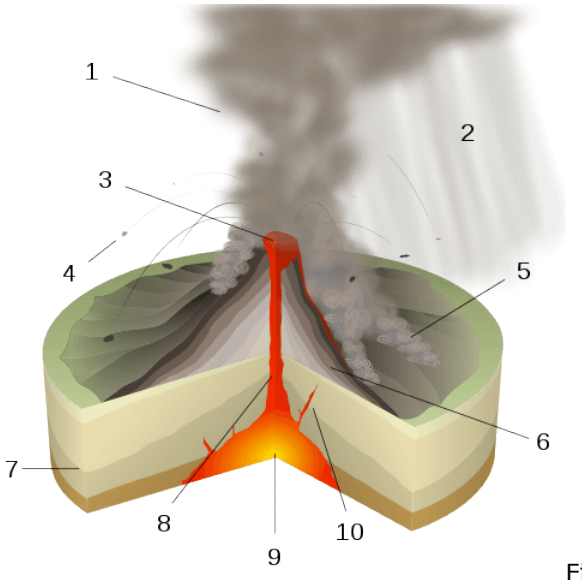


Figure 9. Diagram of Peléan eruption. (key: 1. Ash plume 2. Volcanic ash rain 3. Lava dome 4. Volcanic bomb 5. Pyroclastic flow 6. Layers of lava and ash 7. Stratum 8. Magma conduit 9. Magma chamber 10. Dike)

Peléan eruptions (or nuée ardente) are a type of volcanic eruption, named after the volcano Mount Pelée in Martinique, the site of a massive Peléan eruption in 1902 that is one of the worst natural disasters in history. In Peléan eruptions, a large amount of gas, dust, ash, and lava fragments are blown out the volcano's central crater,[38] driven by the collapse of rhyolite, dacite, and andesite lava dome collapses that often create large eruptive columns. An early sign of a coming eruption is the growth of a so-called Peléan or lava spine, a bulge in the volcano's summit preempting its total collapse.[39] The material collapses upon itself, forming a fast-moving pyroclastic flow[40] (known as a block-and-ash flow)[41] that moves down the side of the mountain at tremendous

speeds, often over 150 km (93 mi) per hour. These massive landslides make Peléan eruptions one of the most dangerous in the world, capable of tearing through populated areas and causing massive loss of life. The 1902 eruption of Mount Pelée caused tremendous destruction, killing more than 30,000 people and competely destroying the town of St. Pierre, the worst volcanic event in the 20th century.[42]

Peléan eruptions are characterized most prominently by the incandescent pyroclastic flows that they drive. The mechanics of a Peléan eruption are very similar to that of a Vulcanian eruption, except that in Peléan eruptions the volcano's structure is able to withstand more pressure, hence the eruption occurs as one large explosion rather than several smaller ones.[43]

Volcanoes known to have Peléan activity include:

- Mount Pelée, Martinique. The 1902 eruption of Mount Pelée completely devastated the island, destroying the town of St. Pierre and leaving only 3 survivors.[44] The eruption was directly preceded by lava dome growth.[45]
- Mayon Volcano, the Philippines most active volcano. It has been the site of many different types of eruptions, Peléan included. Approximately 40 ravines radiate from the summit and provide pathways for frequent pyroclastic flows and mudslides to the lowlands below. Mayon's most violent eruption occurred in 1814 and was responsible for over 1200 deaths.[46]
- The 1951 Peléan eruption of Mount Lamington. Prior to this eruption the peak had not even been recognized as a volcano. Over 3,000 people were killed, and it has become a benchmark for studying large Peléan eruptions.[47]

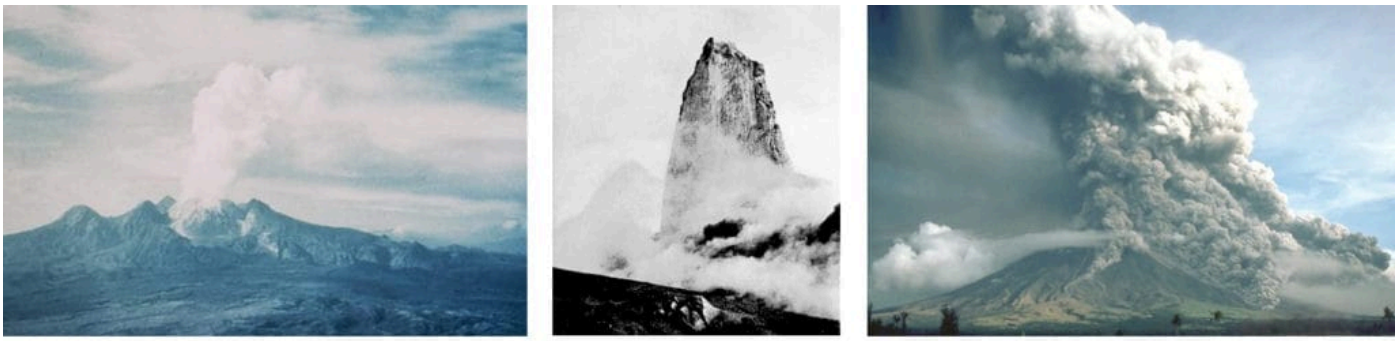


Figure 10. (a) Mount Lamington following the devastating 1951 eruption. (b) The lava spine that developed after the 1902 eruption of Mount Pelée. (c) Pyroclastic flows at Mayon Volcano, Philippines, 1984.

Plinian

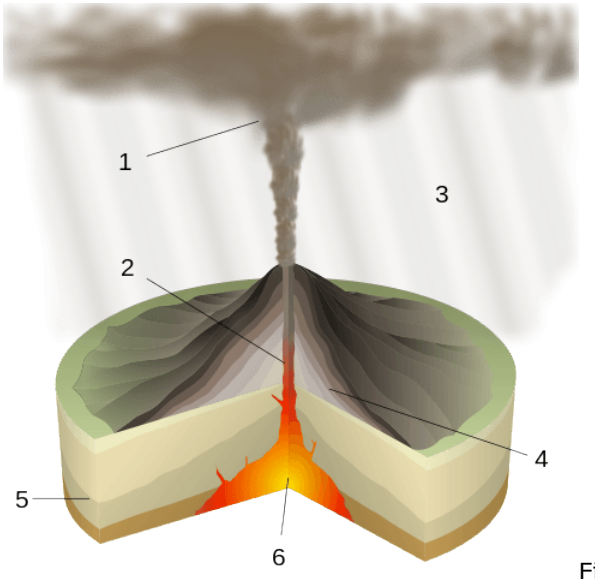


Figure 11. Diagram of a Plinian eruption. (key: 1. Ash plume 2. Magma conduit 3. Volcanic ash rain 4. Layers of lava and ash 5. Stratum 6. Magma chamber)

Plinian eruptions (or Vesuvian) are a type of volcanic eruption, named for the historical eruption of Mount Vesuvius in 79 of Mount Vesuvius that buried the Roman towns of Pompeii and Herculaneum and, specifically, for its chronicler Pliny the Younger.[48] The process powering Plinian eruptions starts in the magma chamber, where dissolved volatile gases are stored in the magma. The gases vesiculate and accumulate as they rise through the magma conduit. These bubbles agglutinate and once they reach a certain size (about 75% of the total volume of the magma conduit) they explode. The narrow confines of the conduit force the gases and associated magma up, forming an eruptive column. Eruption velocity is controlled by the gas contents of the column, and low-strength surface rocks commonly crack under the pressure of the eruption, forming a flared outgoing structure that pushes the gases even faster.[49]

These massive eruptive columns are the distinctive feature of a Plinian eruption, and reach up 2 to 45 km (1 to 28 mi) into the atmosphere. The densest part of the plume, directly above the volcano, is driven internally by gas expansion. As it reaches higher into the air the plume expands and becomes less dense, convection and thermal expansion of volcanic ash drive it even further up into the stratosphere. At the top of the plume, powerful prevailing winds drive the plume in a direction away from the volcano.[50]


Figure 12. 21 April 1990 eruptive column from Redoubt Volcano, as viewed to the west from the Kenai Peninsula.

These highly explosive eruptions are associated with volatile-rich dacitic to rhyolitic lavas, and occur most typically at stratovolcanoes. Eruptions can last anywhere from hours to days, with longer eruptions being associated with more felsic volcanoes. Although they are associated with felsic magma, Plinian eruptions can just as well occur at basaltic volcanoes, given that the magma chamber differentiates and has a structure rich in silicon dioxide.[51]

Plinian eruptions are similar to both Vulcanian and Strombolian eruptions, except that rather than creating discrete explosive events, Plinian eruptions form sustained eruptive columns. They are also similar to Hawaiian lava fountains in that both eruptive types produce sustained eruption columns maintained by the growth of bubbles that move up at about the same speed as the magma surrounding them.[52]


Figure 13. Lahar flows from the 1985 eruption of Nevado del Ruiz, which totally destroyed the town of Armero in Colombia.

Major Plinian eruptive events include:

- The AD 79 eruption of Mount Vesuvius buried the Roman towns of Pompeii and Herculaneum under a layer of ash and tephra. It is the model Plinian eruption. Mount Vesuvius has erupted several times since then. Its last eruption was in 1944 and caused problems for the allied armies as they advanced through Italy.[53] It was the report by Pliny that Younger that lead scientists to refer to vesuvian eruptions as “Plinian.”
- The 1980 eruption of Mount St. Helens in Washington, which ripped apart the volcano’s summit, was a Plinian eruption of Volcanic Explosivity Index (**VEI**) 5.[54]
- The strongest types of eruptions, with a VEI of 8, are so-called “Ultra-Plinian” eruptions, such as the most recent one at Lake Toba 74 thousand years ago, which put out 2800 times the material erupted by Mount St. Helens in 1980.[55]
- Hekla in Iceland, an example of basaltic Plinian volcanism being its 1947-48 eruption. The past 800 years have been a pattern of violent initial eruptions of pumice followed by prolonged extrusion of basaltic lava from the lower part of the volcano.[56]
- Pinatubo in the Philippines on 15 June 1991, which produced 5 km^3 (1 cu mi) of dacitic magma, a 40 km (25 mi) high eruption column, and released 17 megatons of sulfur dioxide.[57]

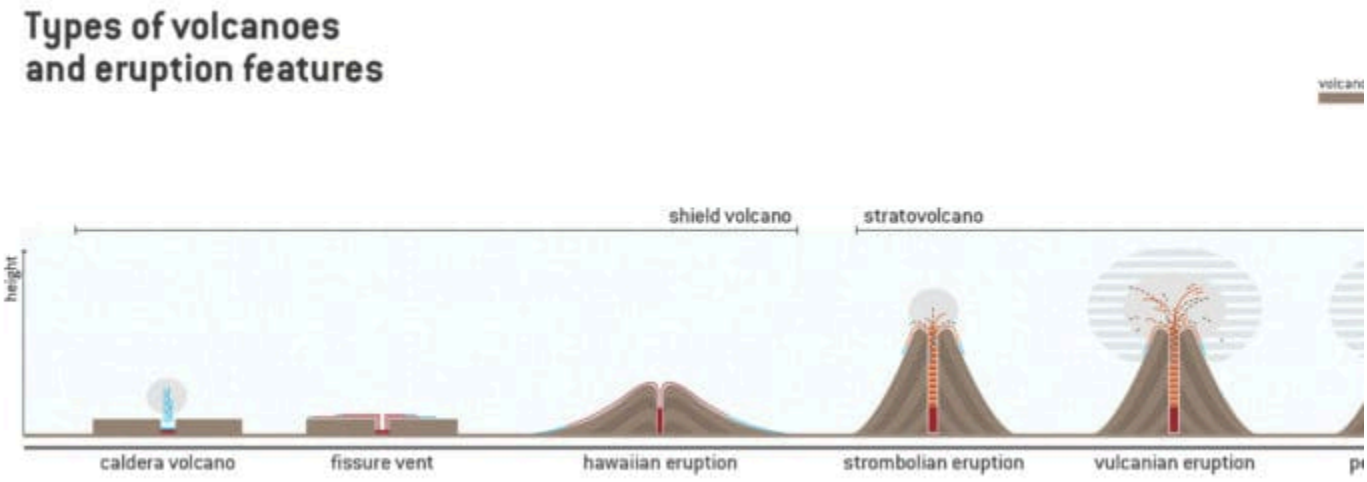


Figure 14. The image correlates types of volcanoes with their respective eruption, highlighting the differences. Click to view a larg

Phreatomagmatic Eruptions

Phreatomagmatic eruptions are eruptions that arise from interactions between water and magma. They are driven from thermal contraction (as opposed to magmatic eruptions, which are driven by thermal expansion) of magma when it comes in contact with water. This temperature difference between the two causes violent water-lava interactions that make up the eruption. The products of phreatomagmatic eruptions are believed to be more regular in shape and finer grained than the products of magmatic eruptions because of the differences in eruptive mechanisms.[58]

There is debate about the exact nature of phreatomagmatic eruptions, and some scientists believe that fuel-coolant reactions may be more critical to the explosive nature than thermal contraction.[59] Fuel coolant reactions may fragment the volcanic material by propagating stress waves, widening cracks and increasing surface area that ultimetly lead to rapid cooling and explosive contraction-driven eruptions.[60]

Surtseyan

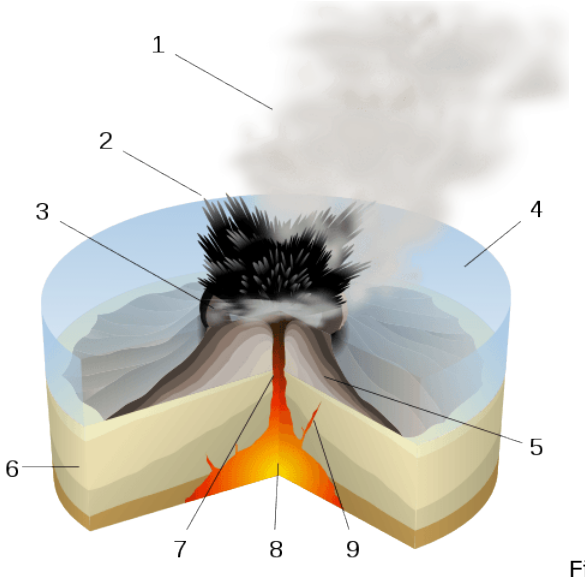


Figure 15. Diagram of a Surtseyan eruption. (key: 1. Water vapor cloud 2. Compressed ash 3. Crater 4. Water 5. Layers of lava and ash 6. Stratum 7. Magma conduit 8. Magma chamber 9. Dike)

A Surtseyan eruption (or hydrovolcanic) is a type of volcanic eruption caused by shallow-water interactions between water and lava, named so after its most famous example, the eruption and formation of the island of Surtsey off the coast of Iceland in 1963. Surtseyan eruptions are the "wet" equivalent of ground-based Strombolian eruptions, but because of where they are taking place they are much more explosive. This is because as water is heated by lava, it flashes in steam and expands violently, fragmenting the magma it is in contact with into fine-grained ash. Surtseyan eruptions are the hallmark of shallow-water volcanic oceanic islands, however they are not specifically confined to them. Surtseyan eruptions can happen on land as well, and are caused by rising magma that comes into contact with an aquifer (water-bearing rock formation) at shallow levels under the volcano.[61] The products of Surtseyan eruptions are generally oxidized palagonite basalts (though andesitic eruptions do occur, albeit rarely), and like Strombolian eruptions Surtseyan eruptions are generally continuous or otherwise rhythmic.[62]

Volcanoes known to have Surtseyan activity include:[63]

- Surtsey, Iceland. The volcano built itself up from depth and emerged above the Atlantic Ocean off the coast of Iceland in 1963. Initial hydrovolcanics were highly explosive, but as the volcano grew out rising lava started to interact less with the water and more with the air, until finally Surtseyan activity waned and became more Strombolian in character.
- Ukinrek Maars in Alaska, 1977, and Capelinhos in the Azores, 1957, both examples of above-water Surtseyan activity.
- Mount Tarawera in New Zealand erupted along a rift zone in 1886, killing 150 people.



Figure 16. (a) Surtsey, erupting 13 days after breaching the water. A tuff ring surrounds the vent. (b) The fissure formed by the 1886 eruption of Mount Tarawera, an example of a fracture zone eruption.

Submarine

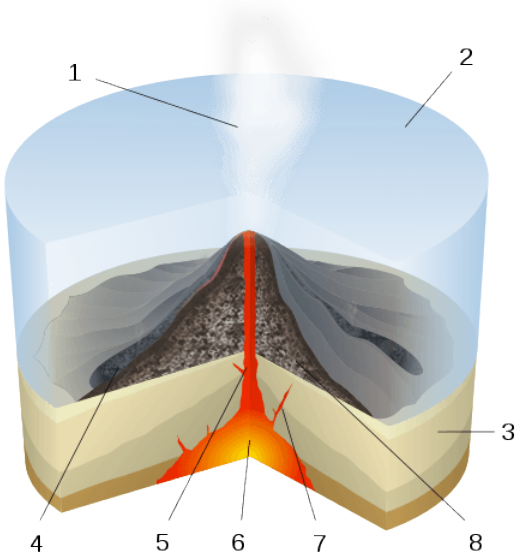


Figure 17. Diagram of a Submarine eruption. (key: 1. Water vapor cloud 2. Water 3. Stratum 4. Lava flow 5. Magma conduit 6. Magma chamber 7. Dike 8. Pillow lava)

Submarine eruptions are a type of volcanic eruption that occurs underwater. An estimated 75% of the total volcanic eruptive volume is generated by submarine eruptions near mid ocean ridges alone, however because of the problems associated with detecting deep sea volcanics, they remained virtually unknown until advances in the 1990s made it possible to observe them.[64]

Submarine eruptions may produce seamounts which may break the surface to form volcanic islands and island chains.

Submarine volcanism is driven by various processes. Volcanoes near plate boundaries and mid-ocean ridges are built by the decompression melting of mantle rock that rises on an upwelling portion of a convection cell to the crustal surface. Eruptions associated with subducting zones, meanwhile, are driven by subducting plates that add volatiles to the rising plate, lowering its melting point. Each process generates different rock; mid-ocean ridge volcanics are primarily basaltic, whereas subduction flows are mostly calc-alkaline, and more explosive and viscous.[65]

Subglacial

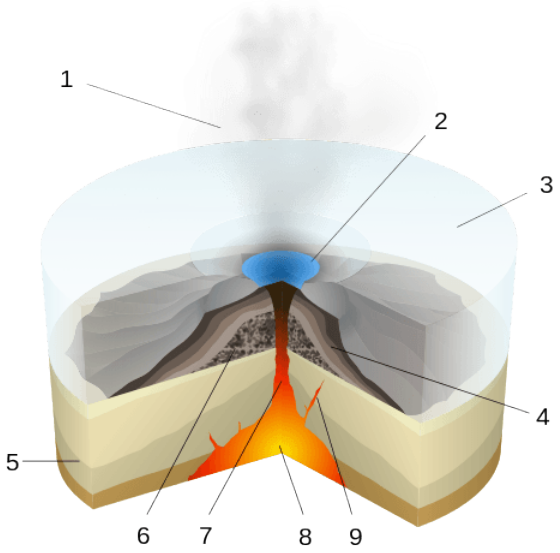


Figure 18. A diagram of a Subglacial eruption. (key: 1. Water vapor cloud 2. Crater lake 3. Ice 4. Layers of lava and ash 5. Stratum 6. Pillow lava 7. Magma conduit 8. Magma chamber 9. Dike)

Subglacial eruptions are a type of volcanic eruption characterized by interactions between lava and ice, often under a glacier.

The nature of glaciovolcanism dictates that it occurs at areas of high latitude and high altitude.[66] It has been suggested that subglacial volcanoes that are not actively erupting often dump heat into the ice covering them, producing meltwater.[67] This meltwater mix means that subglacial eruptions often generate dangerous jökulhlaups (floods) and lahars.[68]

The study of glaciovolcanism is still a relatively new field. Early accounts described the unusual flat-topped steep-sided volcanoes (called tuyas) in Iceland that were suggested to have formed from eruptions below ice. The first English-language paper on the subject was published in 1947 by William Henry Mathews, describing the Tuya Butte field in northwest British Columbia, Canada. The eruptive process that builds these structures, originally inferred in the paper,[69] begins with volcanic growth below the glacier. At first the eruptions resemble those that occur in the deep sea, forming piles of pillow lava at the base of the volcanic structure. Some of the lava shatters when it comes in contact with the cold ice, forming a glassy breccia called hyaloclastite. After a while the ice finally melts into a lake, and the more explosive eruptions of Surtseyan activity begins, building up flanks made up of mostly hyaloclastite. Eventually the lake boils off from continued volcanism, and the lava flows become more effusive and thicken as the lava cools much more slowly, often forming columnar jointing. Well-preserved tuyas show all of these stages, for example Hjorleifshofdi in Iceland.[70]

Glaciovolcanic products have been identified in Iceland, the Canadian province of British Columbia, the U.S. states of Hawaii and Alaska, the Cascade Range of western North America, South America and even on the planet Mars.[71] Volcanoes known to have subglacial activity include:

- Mauna Kea in tropical Hawaii. There is evidence of past subglacial eruptive activity on the volcano in the form of a subglacial deposit on its summit. The eruptions originated about 10,000 years ago, during the last ice age, when the summit of Mauna Kea was covered in ice.[72]
- In 2008, the British Antarctic Survey reported a volcanic eruption under the Antarctica ice sheet 2,200 years ago. It is believed to be that this was the biggest eruption in Antarctica in the last 10,000 years. Volcanic ash deposits from the volcano were identified through an airborne radar survey, buried under later snowfalls in the Hudson Mountains, close to Pine Island Glacier.[73]
- Iceland, well known for both glaciers and volcanoes, is often a site of subglacial eruptions. An example an eruption under the Vatnajökull ice cap in 1996, which occurred under an estimated 2,500 ft (762 m) of ice.[74]
- As part of the search for life on Mars, scientists have suggested that there may be subglacial volcanoes on the red planet. Several potential sites of such volcanism have been reviewed, and compared extensively with similar features in Iceland:[75]
    - Viable microbial communities have been found living in deep (–2800 m) geothermal groundwater at 349 K and pressures >300 bar. Furthermore, microbes have been postulated to exist in basaltic rocks in rinds of altered volcanic glass. All of these conditions could exist in polar regions of Mars today where subglacial volcanism has occurred.



Figure 19. Herðubreið, a tuya in Iceland.

Phreatic eruptions

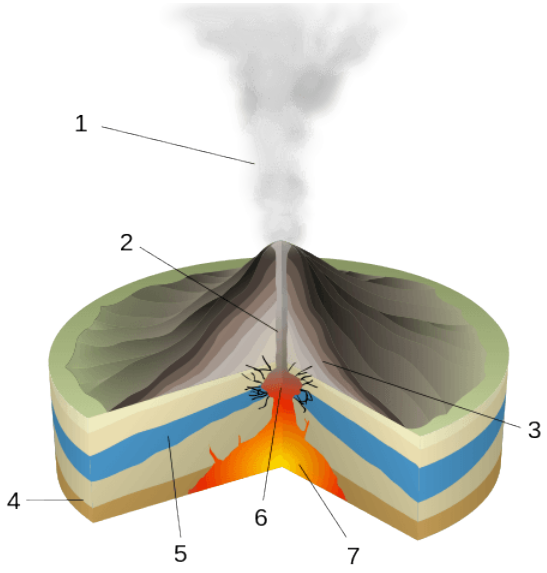


Figure 20. Diagram of a phreatic eruption. (key: 1. Water vapor cloud 2. Magma conduit 3. Layers of lava and ash 4. Stratum 5. Water table 6. Explosion 7. Magma chamber)

Phreatic eruptions (or steam-blast eruptions) are a type of eruption driven by the expansion of steam. When cold ground or surface water come into contact with hot rock or magma it superheats and explodes, fracturing the surrounding rock[76] and thrusting out a mixture of steam, water, ash, volcanic bombs, and volcanic blocks.[77]The distinguishing feature of phreatic explosions is that they only blast out fragments of pre-existing solid rock from the volcanic conduit; no new magma is erupted.[78]Because they are driven by the cracking of rock strata under pressure, phreatic activity does not always result in an eruption; if the rock face is strong enough to withstand the explosive force, outright eruptions may not occur, although cracks in the rock will probably develop and weaken it, furthering future eruptions.[79]

Volcanoes known to exhibit phreatic activity include:

- Mount St. Helens, which exhibited phreatic activity just prior to its catastrophic 1980 eruption (which was itself Plinian).[80]
- Taal Volcano, Philippines, 1965.[81]
- La Soufrière of Guadeloupe (Lesser Antilles), 1975-1976 activity.[82]
- Soufrière Hills volcano on Montserrat, West Indies, 1995–2012.
- Poás Volcano, has frequent geyser like phreatic eruptions from its crater lake.
- Mount Bulusan, well known for its sudden phreatic eruptions.
- Mount Ontake, all historical eruptions of this volcano have been phreatic including the deadly 2014 eruption.

---

1. Heiken, G. and Wohletz, K. *Volcanic Ash*. University of California Press. p. 246. ↩
2. *Ibid*. ↩
3. "VHP Photo Glossary: Effusive Eruption." USGS. 29 December 2009. Retrieved 3 August 2010. ↩
4. "Volcanoes of Canada: Volcanic eruptions." *Geological Survey of Canada*. Natural Resources Canada. 2 April 2009. Retrieved 3 August 2010. ↩
5. "How Volcanoes Work: Hawaiian Eruptions." San Diego State University. Retrieved 2 August 2010. ↩
6. "How Volcanoes Work: Hydrovolcic Eruptions." San Diego State University. Retrieved 4 August 2010. ↩
7. Ruprecht P, Plank T. Feeding andesitic eruptions with a high-speed connection from the mantle. *Nature*. 2013; 500(7460):68-72. ↩
8. "How Volcanoes Work: Eruption Variability." San Diego State University. Retrieved 3 August 2010. ↩
9. "Volcanoes of Canada: Volcanic eruptions." *Geological Survey of Canada*. Natural Resources Canada. 2 April 2009. Retrieved 3 August 2010. ↩
10. "How Volcanoes Work: Eruption Variability." San Diego State University. Retrieved 3 August 2010. ↩
11. Dosseto, A., Turner, S. P. and Van-Orman, J. A. (editors) (2011). *Timescales of Magmatic Processes: From Core to Atmosphere*. Wiley-Blackwell. See also Rothery, David A. (2010). *Volcanoes, Earthquakes and Tsunamis*. Teach Yourself. ↩

12. Heiken, G. and Wohletz, K. *Volcanic Ash*. University of California Press. p. 246. ↩
13. "How Volcanoes Work: Hawaiian Eruptions." San Diego State University. Retrieved 2 August 2010. ↩
14. *Ibid*. ↩
15. "How Volcanoes Work: Basaltic Lava." San Diego State University. Retrieved 2 August 2010. ↩
16. "How Volcanoes Work: Hawaiian Eruptions." San Diego State University. Retrieved 2 August 2010. ↩
17. "How Volcanoes Work: Strombolian Eruptions." San Diego State University. Retrieved 29 July 2010. ↩
18. Mike Burton, Patrick Allard, Filippo Muré, Alessandro La Spina (2007). "Magmatic Gas Composition Reveals the Source Depth of Slug-Driven Strombolian Explosive Activity." *Science* (American Association for the Advancement of Science) **317** (5835): 227–230. doi:10.1126/science.1141900. Retrieved 30 July 2010. ↩
19. "How Volcanoes Work: Strombolian Eruptions." San Diego State University. Retrieved 29 July 2010. ↩
20. *Ibid*. ↩
21. Cain, Fraser (22 April 2010). "Strombolian Eruption." *Universe Today*. Retrieved 30 July 2010. ↩
22. "How Volcanoes Work: Strombolian Eruptions." San Diego State University. Retrieved 29 July 2010. ↩
23. Cain, Fraser (22 April 2010). "Strombolian Eruption." *Universe Today*. Retrieved 30 July 2010. ↩
24. "How Volcanoes Work: Strombolian Eruptions." San Diego State University. Retrieved 29 July 2010. ↩
25. "How Volcanoes Work: Strombolian Eruptions." San Diego State University. Retrieved 29 July 2010. See also Cain, Fraser (22 April 2010). "Strombolian Eruption." *Universe Today*. Retrieved 30 July 2010. ↩
26. "How Volcanoes Work: Strombolian Eruptions." San Diego State University. Retrieved 29 July 2010. ↩
27. Seach, John. "Mt Etna Volcano Eruptions—John Seach." *Old eruptions*. Volcanolive. Retrieved 30 July 2010. ↩
28. Seach, John. "Mt Etna Volcano Eruptions—John Seach." *Recent eruptions*. Volcanolive. Retrieved 30 July 2010. ↩
29. "Erebus." *Global Volcanism Program*. Smithsonian National Museum of Natural History. Retrieved 31 July 2010. ↩
30. Kyle, P. R. (Ed.), Volcanological and Environmental Studies of Mount Erebus, Antarctica, Antarctic Research Series, American Geophysical Union, Washington DC, 1994. ↩
31. "Stromboli." *Global Volcanism Program*. Smithsonian National Museum of Natural History. Retrieved 31 July 2010. ↩
32. "How Volcanoes Work: Vulcanian Eruptions." San Diego State University. Retrieved 1 August 2010. ↩
33. Cain, Fraser. "Vulcanian Eruptions." Universe Today. Retrieved 1 August 2010. ↩
34. "How Volcanoes Work: Vulcanian Eruptions." San Diego State University. Retrieved 1 August 2010. ↩
35. "How Volcanoes Work: Sakurajima Volcano." San Diego State University. Retrieved 1 August 2010. ↩
36. "How Volcanoes Work: Vulcanian Eruptions." San Diego State University. Retrieved 1 August 2010. ↩
37. "VHP Photo Glossary: Vulcanian eruption." USGS. Retrieved 1 August 2010. ↩
38. Cain, Fraser. "Pelean Eruption." Universe Today. Retrieved 2 August 2010. ↩
39. Donald Hyndman and David Hyndman (April 2008). *Natural Hazards and Disasters*. Cengage Learning. pp. 134–135.↩
40. Cain, Fraser. "Pelean Eruption." Universe Today. Retrieved 2 August 2010. ↩
41. Nelson, Stephan A. (30 September 2007). "Volcanoes, Magma, and Volcanic Eruptions." Tulane University. Retrieved 2 August 2010. ↩
42. Cain, Fraser. "Pelean Eruption." Universe Today. Retrieved 2 August 2010. ↩
43. Richard V. Fisher and Grant Heiken (1982). "Mt. Pelée, Martinique: May 8 and 20 pyroclastic flows and surges." *Journal of Volcanology and Geothermal Research***13** (3–4): 339–371. doi:10.1016/0377-0273(82)90056-7↩
44. "How Volcanoes Work: Mount Pelée Eruption (1902)."San Diego State University. Retrieved 1 August 2010. ↩
45. "How Volcanoes Work: Vulcanian Eruptions." San Diego State University. Retrieved 1 August 2010. ↩
46. "Mayon." *Global Volcanism Program*. Smithsonian National Museum of Natural History. Retrieved 2 August 2010. ↩
47. "Lamington: Photo Gallery." *Global Volcanism Program*. Smithsonian National Museum of Natural History. Retrieved2 August 2010. ↩
48. "How Volcanoes Work: Plinian Eruptions." San Diego State University. Retrieved 3 August 2010. ↩
49. "How Volcanoes Work: Eruption Model." San Diego State University. Retrieved 3 August 2010. ↩
50. *Ibid*. ↩
51. "How Volcanoes Work: Plinian Eruptions." San Diego State University. Retrieved 3 August 2010. ↩
52. *Ibid*. ↩
53. *Ibid*. ↩
54. "Volcanoes of Canada: Volcanic eruptions." *Geological Survey of Canada*. Natural Resources Canada. 2 April 2009. Retrieved 3 August 2010. ↩
55. "How Volcanoes Work: Eruption Variability." San Diego State University. Retrieved 3 August 2010. See also "How Volcanoes Work: Calderas." San Diego State University. Retrieved 3 August 2010. ↩
56. "How Volcanoes Work: Plinian Eruptions." San Diego State University. Retrieved 3 August 2010. ↩
57. Stephen Self, Jing-Xia Zhao, Rick E. Holasek, Ronnie C. Torres, and Alan J. King. "The Atmospheric Impact of the 1991 Mount Pinatubo Eruption." USGS. Retrieved 3 August 2010. ↩
58. Heiken, G. and Wohletz, K. *Volcanic Ash*. University of California Press. p. 246. See alsoA.B. Starostin, A.A. Barmin, and O.E. Melnik (May 2005). "A transient model for explosive and phreatomagmatic eruptions." *Journal of Volcanology and*

Geothermal Research. Volcanic Eruption Mechanisms—Insights from intercomparison of models of conduit processes **143** (1–3): 133–151. doi: 10.1016/j.jvolgeores.2004.09.014. Retrieved 4 August 2010. ↩
59. A.B. Starostin, A.A. Barmin, and O.E. Melnik (May 2005). "A transient model for explosive and phreatomagmatic eruptions." *Journal of Volcanology and Geothermal Research*. Volcanic Eruption Mechanisms—Insights from intercomparison of models of conduit processes **143** (1–3): 133–151. doi: 10.1016/j.jvolgeores.2004.09.014. Retrieved 4 August 2010. ↩
60. Heiken, G. and Wohletz, K. *Volcanic Ash*. University of California Press. p. 246. ↩
61. "How Volcanoes Work: Hydrovolcic Eruptions." San Diego State University. Retrieved 4 August 2010. ↩
62. "X. Classification of Volcanic Eruptions: Surtseyan Eruptions." *Lecture Notes*. University of Alabama. Retrieved5 August 2010. ↩
63. "How Volcanoes Work: Hydrovolcic Eruptions." San Diego State University. Retrieved 4 August 2010. ↩
64. Chadwick, Bill (10 January 2006). "Recent Submarine Volcanic Eruptions." *Vents Program*. NOAA. Retrieved 5 August 2010. ↩
65. Hubert Straudigal and David A Clauge. "The Geological History of Deep-Sea Volcanoes: Biosphere, Hydrosphere, and Lithosphere Interactions" (PDF). *Oceanography*. Seamounts Special Issue (Oceanography Society) **32** (1). Retrieved 4 August 2010. ↩
66. "Glaciovolcanism—University of British Columbia." University of British Columbia. Retrieved 5 August 2010. ↩
67. Black, Richard (20 January 2008). "Ancient Antarctic eruption noted." BBC News. Retrieved 5 August 2010. ↩
68. "Glaciovolcanism—University of British Columbia." University of British Columbia. Retrieved 5 August 2010. ↩
69. *Ibid*. ↩
70. Alden, Andrew. "Tuya or Subglacial Volcano, Iceland." about.com. Retrieved 5 August 2010. ↩
71. *Ibid*. ↩
72. "Kinds of Volcanic Eruptions." *Volcano World*. Oregon State University. Retrieved 5 August 2010. ↩
73. Black, Richard (20 January 2008). "Ancient Antarctic eruption noted." BBC News. Retrieved 5 August 2010. ↩
74. "Iceland's subglacial eruption." *Hawaiian Volcano Observatory*. USGS. 11 October 1996. Retrieved 5 August 2010. ↩
75. "Subglacial Volcanoes On Mars." Space Daily. 27 June 2001. Retrieved 5 August 2010. ↩
76. Leonid N. Germanovich and Robert P. Lowell (1995). "The mechanism of phreatic eruptions." *Journal of Geophysical Research*. Solid Earth (American Geophysical Union) **100** (B5): 8417–8434. doi:10.1029/94JB03096. Retrieved 7 August 2010. ↩
77. "VHP Photo Glossary: Phreatic eruption." USGS. 17 July 2008. Retrieved 6 August 2010. ↩
78. Watson, John (5 February 1997). "Types of volcanic eruptions." USGS. Retrieved 7 August 2010. ↩
79. Leonid N. Germanovich and Robert P. Lowell (1995). "The mechanism of phreatic eruptions." *Journal of Geophysical Research*. Solid Earth (American Geophysical Union) **100** (B5): 8417–8434. doi:10.1029/94JB03096. Retrieved 7 August 2010. ↩
80. "VHP Photo Glossary: Phreatic eruption." USGS. 17 July 2008. Retrieved 6 August 2010. ↩
81. Watson, John (5 February 1997). "Types of volcanic eruptions." USGS. Retrieved 7 August 2010. ↩
82. *Ibid*. ↩

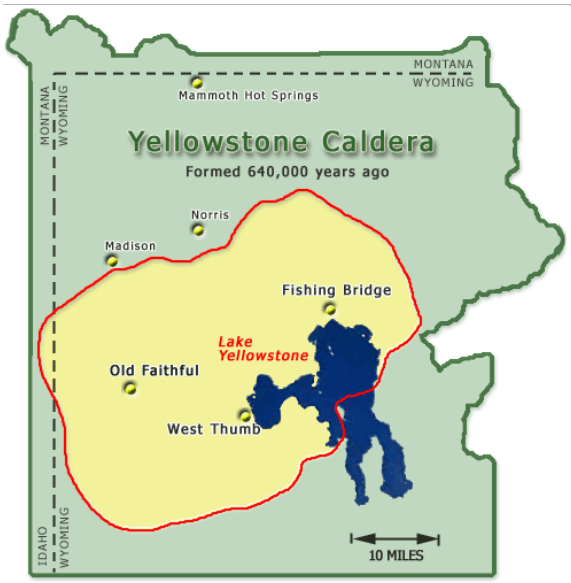
Contributors and Attributions

CC licensed content, Shared previously

- Types of volcanic eruptions. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Types_of_volcanic_eruptions. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.9.2: Types of Eruptions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.3: Supervolcanoes



What would cause such a giant caldera?

You can stand on the rim and view the enormous Yellowstone Caldera, but it's hard to visualize a volcano or a set of eruptions that enormous. Supervolcanoes are a fairly new idea in volcanology. Although their eruptions are unbelievably massive, they are exceedingly rare. The power of Yellowstone, even 640,000 years after the most recent eruption, is seen in its fantastic geysers.

Supervolcano eruptions are extremely rare in Earth's history. It's a good thing because they are unimaginably large. A supervolcano must erupt more than 1,000 cubic km (240 cubic miles) of material, compared with 1.2 km^3 for Mount St. Helens or 25 km^3 for Mount Pinatubo, a large eruption in the Philippines in 1991. Not surprisingly, supervolcanoes are the most dangerous type of volcano.

Supervolcano Eruptions

The exact cause of supervolcano eruptions is still debated. However, scientists think that a very large magma chamber erupts entirely in one catastrophic explosion. This creates a huge hole or caldera into which the surface collapses (Figure 1).



Figure 1. The caldera at Santorini in Greece is so large that it can only be seen by satellite.

Yellowstone Caldera

The largest supervolcano in North America is beneath Yellowstone National Park in Wyoming. Yellowstone sits above a hotspot that has erupted catastrophically three times: 2.1 million, 1.3 million, and 640,000 years ago. Yellowstone has produced many smaller (but still enormous) eruptions more recently (Figure 2). Fortunately, current activity at Yellowstone is limited to the region's famous geysers.

The Old Faithful web cam shows periodic eruptions of Yellowstone's famous geyser in real time.

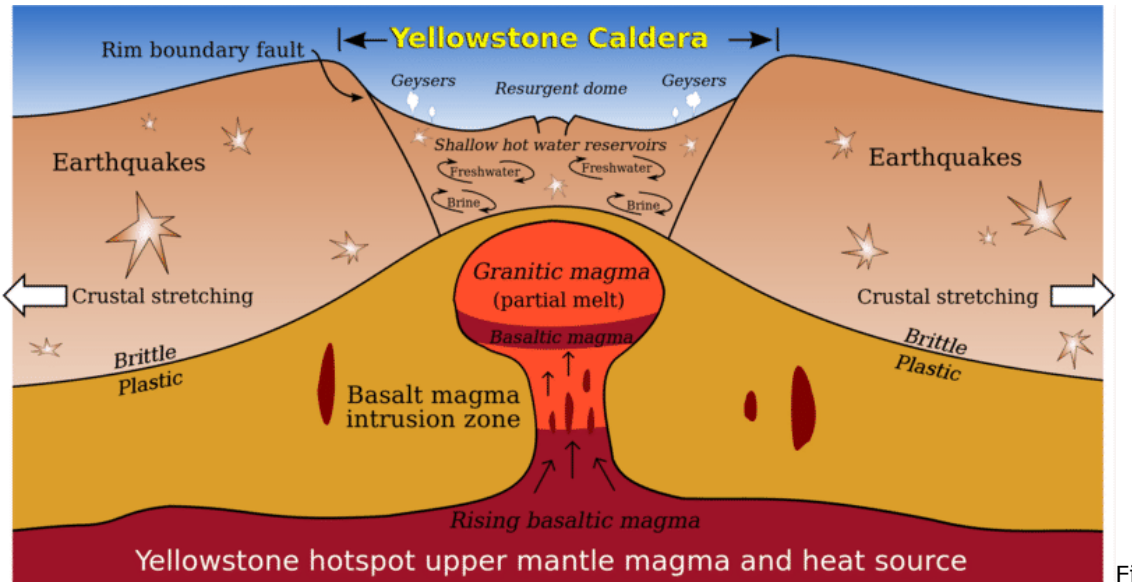


Figure 2. The Yellowstone hotspot has produced enormous felsic eruptions. The Yellowstone caldera collapsed in the most recent super eruption.

Supervolcano Eruptions and Life on Earth

A supervolcano could change life on Earth as we know it. Ash could block sunlight so much that photosynthesis would be reduced and global temperatures would plummet. Volcanic eruptions could have contributed to some of the mass extinctions in our planet's history. No one knows when the next super eruption will be.

Interesting volcano videos are seen on National Geographic Videos: Environment Video, Natural Disasters, Earthquakes. One interesting one is “Mammoth Mountain,” which explores Hot Creek and the volcanic area it is a part of in California.

Summary

- Supervolcano eruptions are rare but massive and deadly.
- Yellowstone Caldera is a supervolcano that has erupted catastropically three times.
- Supervolcano eruptions can change the course of life on Earth.

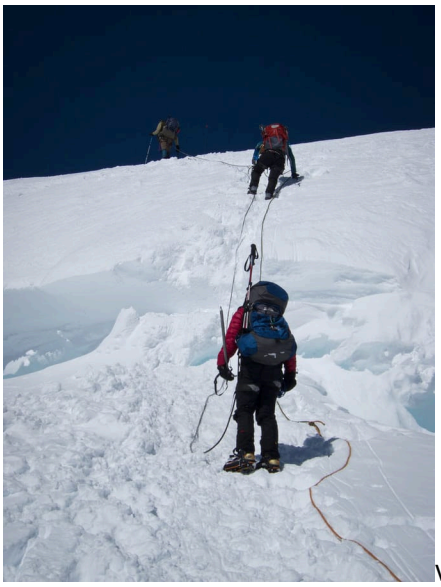
Contributors and Attributions

CC licensed content, Shared previously

- Supervolcanoes. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Supervolcanoes/lesson/Supervolcanoes-HS-ES/?referrer=featured_content. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.9.3: Supervolcanoes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.4: Volcanoes at Plate Boundaries


Volcanoes are fun (and difficult) to climb. Climbing in the Cascades ranges in difficulty from a non-technical hike, like on South Sister, to a technical climb on Mount Baker in which an ice axe, crampons, and experience are needed.

Convergent Plate Boundaries

Converging plates can be oceanic, continental, or one of each. If both are continental they will smash together and form a mountain range. If at least one is oceanic, it will subduct. A subducting plate creates volcanoes. Locations with converging in which at least one plate is oceanic at the boundary have volcanoes.

Melting

Melting at convergent plate boundaries has many causes. The subducting plate heats up as it sinks into the mantle. Also, water is mixed in with the sediments lying on top of the subducting plate. As the sediments subduct, the water rises into the overlying mantle material and lowers its melting point. Melting in the mantle above the subducting plate leads to volcanoes within an island or continental arc.

Why does melting occur at convergent plate boundaries? The subducting plate heats up as it sinks into the mantle. Also, water is mixed in with the sediments lying on top of the subducting plate. This water lowers the melting point of the mantle material, which increases melting. Volcanoes at convergent plate boundaries are found all along the Pacific Ocean basin, primarily at the edges of the Pacific, Cocos, and Nazca plates. Trenches mark subduction zones, although only the Aleutian Trench and the Java Trench appear on the map in figure 1.

Remember your plate tectonics knowledge. Large earthquakes are extremely common along convergent plate boundaries. Since the Pacific Ocean is rimmed by convergent and transform boundaries, about 80% of all earthquakes strike around the Pacific Ocean basin (the ring of fire). Why are 75% of the world's volcanoes found around the Pacific basin? Of course, these volcanoes are caused by the abundance of convergent plate boundaries around the Pacific.



Figure 1. The Cascade Range is formed by volcanoes created from subduction of oceanic crust beneath the North American continent.

Pacific Rim

The Pacific Ring of Fire is where the majority of the volcanic activity on the Earth occurs. A description of the Pacific Ring of Fire along western North America is a description of the plate boundaries.

- Subduction at the Middle American Trench creates volcanoes in Central America.
- The San Andreas Fault is a transform boundary.
- Subduction of the Juan de Fuca plate beneath the North American plate creates the Cascade volcanoes.
- Subduction of the Pacific plate beneath the North American plate in the north creates the Aleutian Islands volcanoes.

This incredible explosive eruption of Mount Vesuvius in Italy in A.D. 79 is an example of a composite volcano that forms as the result of a convergent plate boundary:

A link to an interactive elements can be found at the bottom of this page.

Volcanoes at convergent plate boundaries are found all along the Pacific Ocean basin, primarily at the edges of the Pacific, Cocos, and Nazca plates. Trenches mark subduction zones.

The Cascades are a chain of volcanoes at a convergent boundary where an oceanic plate is subducting beneath a continental plate. Specifically the volcanoes are the result of subduction of the Juan de Fuca, Gorda, and Explorer Plates beneath North America. The volcanoes are located just above where the subducting plate is at the right depth in the mantle for there to be melting (Figure 1).

The Cascades have been active for 27 million years, although the current peaks are no more than 2 million years old. The volcanoes are far enough north and are in a region where storms are common, so many are covered by glaciers.

The Cascades are shown on this interactive map with photos and descriptions of each of the volcanoes.


Figure 3. Mt. Baker, Washington.

Divergent plate boundaries

At divergent plate boundaries hot mantle rock rises into the space where the plates are moving apart. As the hot mantle rock convects upward it rises higher in the mantle. The rock is under lower pressure; this lowers the melting temperature of the rock and so it melts. Lava erupts through long cracks in the ground, or fissures.


Figure 4. A volcanic eruption at Surtsey, a small island near Iceland.

Why does melting occur at divergent plate boundaries? Hot mantle rock rises where the plates are moving apart. This releases pressure on the mantle, which lowers its melting temperature. Lava erupts through long cracks in the ground, or **fissures.**

Volcanoes erupt at mid-ocean ridges, such as the Mid-Atlantic ridge, where seafloor spreading creates new seafloor in the rift valleys. Where a hotspot is located along the ridge, such as at Iceland, volcanoes grow high enough to create islands (figure 4).

Mid-Ocean Ridges



Figure 5. Mount Gahinga in the East African Rift valley.

Volcanoes erupt at mid-ocean ridges, such as the Mid-Atlantic ridge, where seafloor spreading creates new seafloor in the rift valleys. Where a hotspot is located along the ridge, such as at Iceland, volcanoes grow high enough to create islands.

Continental Rifting

Eruptions are found at divergent plate boundaries as continents break apart. The volcanoes in Figure 5 are in the East African Rift between the African and Arabian plates. Remember from the chapter Plate Tectonics that Baja California is being broken apart from mainland Mexico as another example of continental rifting.

Summary

- Melting is common at convergent plate boundaries.
- Convergent plate boundaries line the Pacific Ocean basin so that volcanic arcs line the region.
- Melting at divergent plate boundaries is due to pressure release.
- At mid-ocean ridges seafloor is pulled apart and new seafloor is created.

Explore More

Use this resource to answer the questions that follow.

A link to an interactive elements can be found at the bottom of this page.

(You can stop watching at 11:02.)

1. What percent of volcanoes and earthquakes occur on the Pacific Ring of Fire?
2. How long is the arc of volcanoes along the Pacific Rim?
3. How has Augustine built up so high? Does it have high or low silica?
4. What type of volcanoes are found along the ring of fire? What happens to the gas in the magma? what kills so many people?
5. What does water do in hot rock below the surface?
6. What does carbon-12 indicate?
7. What process brings the sediments and water into the mantle?

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Provided by**: Lumen Learning. **License**: *CC BY-NC: Attribution-NonCommercial*

CC licensed content, Shared previously

- Volcanoes at Plate Boundaries. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Volcanoes-at-Plate-Boundaries/lesson/Volcanoes-at-Plate-Boundaries-HS-ES/?referrer=featured_content#x-ck12-SFMtRVMtMDgtMDQtR2FoaW5nYU11aGFidXJh. **License**: *CC BY-NC: Attribution-NonCommercial*
- Volcanoes. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/8.0/. **License**: *CC BY-NC: Attribution-NonCommercial*

All rights reserved content

- Ring of Fire- Abbreviated. **Authored by**: Frank Gregorio. **Located at**: https://youtu.be/uw7Uq137YJQ. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

This page titled 4.9.4: Volcanoes at Plate Boundaries is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.5: Volcanoes Hotspots

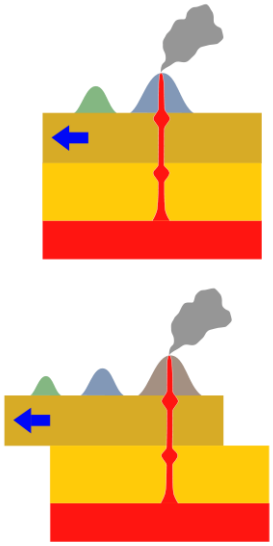


Figure 1. Diagram showing a cross section though the Earth's lithosphere (in yellow) with magma rising from the mantle (in red)

In geology, the places known as **hotspots** or **hot spots** are volcanic regions thought to be fed by underlying mantle that is anomalously hot compared with the surrounding mantle. They may be on, near to, or far from tectonic plate boundaries. Currently, there are two hypotheses that attempt to explain their origins. One suggests that they are due to hot mantle plumes that rise as thermal diapirs from the core–mantle boundary.[1] An alternative hypothesis postulates that it is not high temperature that causes the volcanism, but lithospheric extension that permits the passive rising of melt from shallow depths.[2] This hypothesis considers the term "hotspot" to be a misnomer, asserting that the mantle source beneath them is, in fact, not anomalously hot at all. Well known examples include Hawaii and Yellowstone.

Background

The origins of the concept of hotspots lie in the work of J. Tuzo Wilson, who postulated in 1963 that the Hawaiian Islands result from the slow movement of a tectonic plate across a hot region beneath the surface.[3] It was later postulated that hotspots are fed by narrow streams of hot mantle rising from the Earth's core–mantle boundary in a structure called a mantle plume.[4] Whether or not such mantle plumes exist is currently the subject of a major controversy in Earth science.[5] Estimates for the number of hotspots postulated to be fed by mantle plumes has ranged from about 20 to several thousands, over the years, with most geologists considering a few tens to exist. Hawaii, Réunion, Yellowstone, Galápagos, and Iceland are some of the currently most active volcanic regions to which the hypothesis is applied.

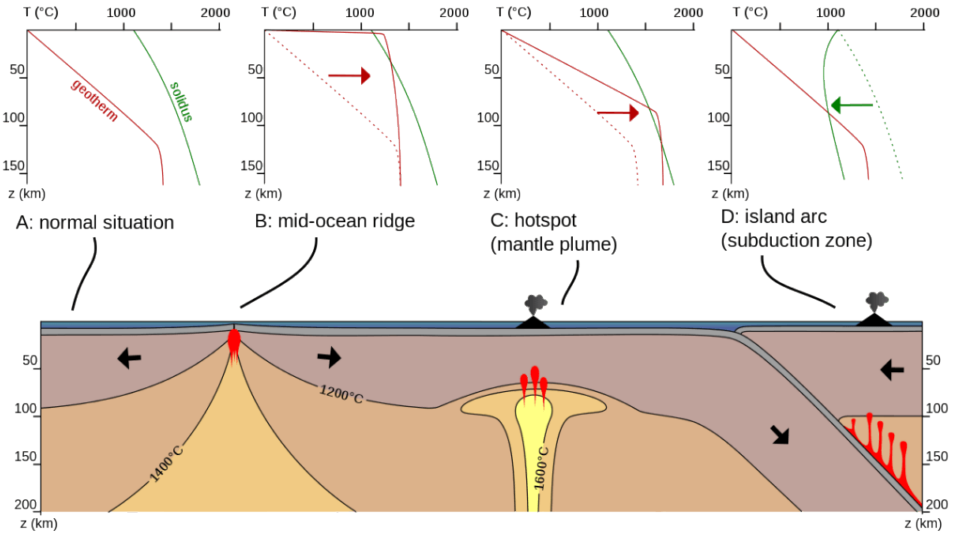


Figure 2. Schematic diagram showing the physical processes inside the Earth that lead to the generation of magma. Partial melting begins above the fusion point.

Most hotspot volcanoes are basaltic (e.g., Hawaii, Tahiti). As a result, they are less explosive than subduction zone volcanoes, in which water is trapped under the overriding plate. Where hotspots occur in continental regions, basaltic magma rises through the continental crust, which melts to form rhyolites. These rhyolites can form violent eruptions. For example, the Yellowstone Caldera was formed by some of the most powerful volcanic explosions in geologic history. However, when the rhyolite is completely erupted, it may be followed by eruptions of basaltic magma rising through the same lithospheric fissures (cracks in the lithosphere). An example of this activity is theIlgachuz Range in British Columbia, which was created by an early complex series of trachyte and rhyolite eruptions, and late extrusion of a sequence of basaltic lava flows.[6]

The hotspot hypothesis is now closely linked to the mantle plume hypothesis.

Comparison with island arc volcanoes

Hotspot volcanoes are considered to have a fundamentally different origin from island arc volcanoes. The latter form over subduction zones, at converging plate boundaries. When one oceanic plate meets another, the denser plate is forced downward into a deep ocean trench. This plate, as it is subducted, releases water into the base of the over-riding plate, and this water mixes with the rock, thus changing its composition causing some rock to melt and rise. It is this that fuels a chain of volcanoes, such as the Aleutian Islands, near Alaska.

Hotspot volcanic chains

The joint mantle plume/hotspot hypothesis envisages the feeder structures to be fixed relative to one another, with the continents and seafloor drifting overhead. The hypothesis thus predicts that time-progressive chains of volcanoes are developed on the surface. Examples are Yellowstone, which lies at the end of a chain of extinct calderas, which become progressively older to the west. Another example is the Hawaiian archipelago, where islands become progressively older and more deeply eroded to the northwest.

Geologists have tried to use hotspot volcanic chains to track the movement of the Earth's tectonic plates. This effort has been vexed by the lack of very long chains, by the fact that many are not time-progressive (e.g. the Galápagos) and by the fact that hotspots do not appear to be fixed relative to one another (e.g. Hawaii and Iceland[7]).

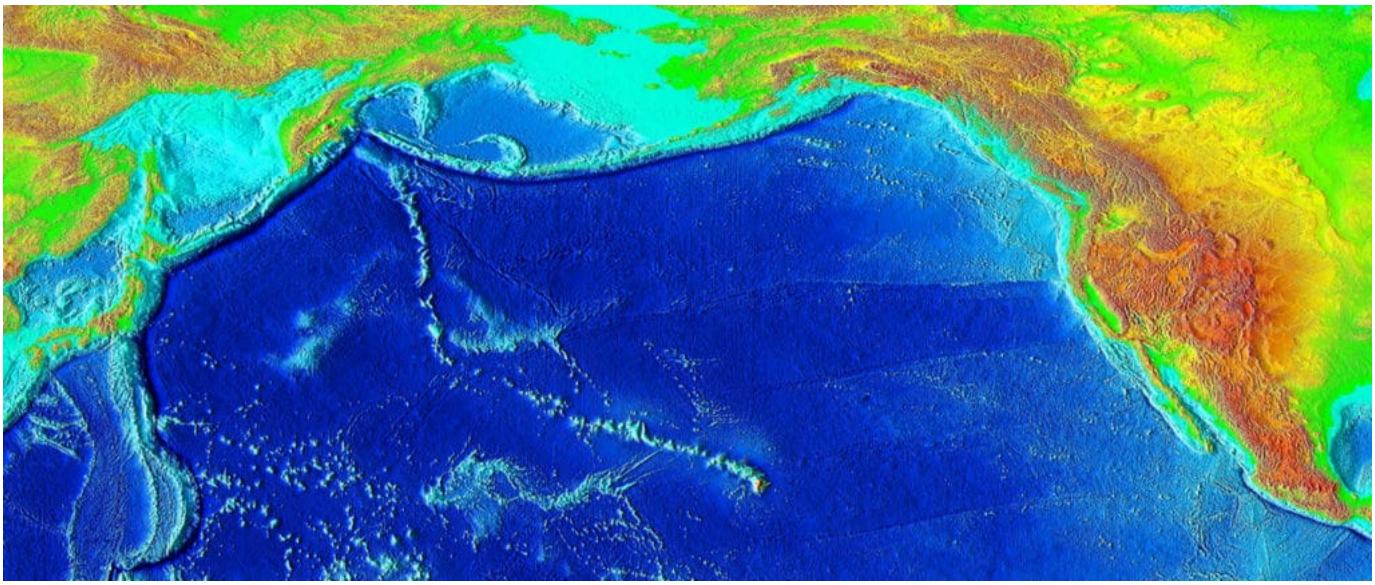


Figure 3. Over millions of years, the Pacific Plate has moved over the Hawaii hotspot, creating a trail of underwater mountains that stretch across the Pacific

1. W. J. Morgan (5 March 1971). "Convection Plumes in the Lower Mantle." *Nature***230** (5288): 42–43. doi: 10.1038/230042a0. ↩
2. "Do plumes exist?". Retrieved 2010-04-25. See also, Foulger, G.R. (2010). *Plates vs. Plumes: A Geological Controversy*. Wiley-Blackwell. ↩
3. Wilson, J. Tuzo (1963). "A possible origin of the Hawaiian Islands" (PDF). *Canadian Journal of Physics***41** (6): 863–870. doi:10.1139/p63-094. ↩
4. "Hotspots: Mantle thermal plumes." United States Geological Survey. 1999-05-05. Retrieved 2008-05-15. ↩
5. Foulger, G.R. (2010). *Plates vs. Plumes: A Geological Controversy*. Wiley-Blackwell. See also,Wright, Laura (November 2000). "Earth's interior: Raising hot spots." *Geotimes*. American Geological Institute. Retrieved 2008-06-15. ↩
6. Holbek, Peter (November 1983). "Report on Preliminary Geology and Geochemistry of the Ilga Claim Group" (PDF). Retrieved 2008-06-15. ↩
7. "What the hell is Hawaii?" Retrieved 2011-01-07. ↩

Contributors and Attributions

CC licensed content, Shared previously

- Hotspot (geology). **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Hotspot_%28geology%29. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.9.5: Volcanoes Hotspots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.6: Pyroclastic Deposits


Figure 1. USGS scientist examines pumice blocks at the edge of a pyroclastic flow from Mount St. Helens

Pyroclastic rocks or **pyroclastics** are clastic rocks composed solely or primarily of volcanic materials. Where the volcanic material has been transported and reworked through mechanical action, such as by wind or water, these rocks are termed ***volcaniclastic***. Commonly associated with unsieved volcanic activity—such as Plinian or krakatoan eruption styles, or phreatomagmatic eruptions—pyroclastic deposits are commonly formed from airborne ash, lapilli and bombs or blocks ejected from the volcano itself, mixed in with shatteredcountry rock.

Pyroclastic rocks may be a range of clast sizes, from the largest agglomerates, to very fine ashes and tuffs. Pyroclasts of different sizes are classified as volcanic bombs, lapilli, and volcanic ash. Ash is considered to be pyroclastic because it is a fine dust made up of volcanic rock. One of the most spectacular forms of pyroclastic deposit are the ignimbrites, deposits formed by the high-temperature gas-and-ash mix of a pyroclastic flow event.

Classification

Types of Pyroclasts[1]			
Clast size	**Pyroclast**	**Mainly unconsolidated: Tephra**	**Mainly consolidated: Pyroclastic rock**
> 64 mm	Block, bomb	Agglomerate	Agglomerate, pyroclastic breccia
< 64 mm	Lapillus	Layer, lapilli tephra	Lapilli tuff, lapillistone
< 2 mm	Coarse ash	Coarse ash	Coarse (ash) tuff
< 0.063 mm	Fine ash	Fine ash	Fine (ash) tuff



Figure 2. Rocks from the Bishop Tuff, uncompressed with pumice on left; compressed with fiamme on right.

Three modes of transport can be distinguished: pyroclastic flow, pyroclastic surge, and pyroclastic fall. During Plinian eruptions, pumice and ash are formed when silicic magma is fragmented in the volcanic conduit, because of decompression and the growth of bubbles. Pyroclasts are then entrained in a buoyant eruption plume which can rise several kilometers into the air and cause aviation hazards. Particles falling from the eruption clouds form layers on the ground (this is pyroclastic fall or tephra). Pyroclastic density currents, which are referred to as “flows” or “surges” depending on particle concentration and the level turbulence, are sometimes called *glowing avalanches*. The deposits of pumice-rich pyroclastic flows can be called ignimbrites.

A pyroclastic eruption entails spitting or “fountaining” lava, where the lava will be thrown into the air along with ash, pyroclastic materials, and other volcanic byproducts. Hawaiian eruptions such as those at Kīlauea can eject clots of magma suspended into gas; this is called a “fire fountain”. The magma clots, if hot enough may coalesce upon landing to form a lava flow.

Pyroclastic deposits consist of pyroclasts which are not cemented together. Pyroclastic rocks (tuff) are pyroclastic deposits which have been lithified.

References

Blatt, Harvey and Robert J. Tracy (1996) *Petrology: Igneous, Sedimentary, and Metamorphic*, W.H.W. Freeman & Company; 2nd ed., pp. 26–29.

---

1. Heiken, G. and Wohletz, K., 1985 *Volcanic Ash*, University of California Press;, pp. 246. ↩

Contributors and Attributions

CC licensed content, Shared previously

- Pyroclastic rock. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Pyroclastic_rock#cite_ref-1. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.9.6: Pyroclastic Deposits is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.7: Dikes and Sills

Dikes

A **dike** is a sheet of rock that formed in a fracture in a pre-existing rock body. Dikes can be either magmatic or sedimentary in origin. Magmatic dikes form when magma intrudes into a crack then crystallizes as a sheet intrusion, either cutting across layers of rock or through an unlayered mass of rock. Clastic dikes are formed when sediment fills a pre-existing crack.[1]



Figure 1. A magmatic dike cross-cutting horizontal layers of sedimentary rock, in Makhtesh Ramon, Israel

Magmatic Dikes

An intrusive dike is an igneous body with a very high aspect ratio, which means that its thickness is usually much smaller than the other two dimensions. Thickness can vary from sub-centimeter scale to many meters, and the lateral dimensions can extend over many kilometres. A dike is an intrusion into an opening cross-cutting fissure, shouldering aside other pre-existing layers or bodies of rock; this implies that a dike is always younger than the rocks that contain it. Dikes are usually high-angle to near-vertical in orientation, but subsequent tectonic deformation may rotate the sequence of strata through which the dike propagates so that the dike becomes horizontal. Near-horizontal, or conformable intrusions, along bedding planes between strata are called intrusive sills.

Sometimes dikes appear in swarms, consisting of several to hundreds of dikes emplaced more or less contemporaneously during a single intrusive event. The world's largest dike swarm is the Mackenzie dike swarm in the Northwest Territories, Canada.[2]

Dikes often form as either radial or concentric swarms around plutonic intrusives, volcanic necks or feeder vents in volcanic cones. The latter are known as ring dikes.

Dikes can vary in texture and their composition can range from diabase or basaltic to granitic or rhyolitic, but on a global perspective the basaltic composition prevails, manifesting ascent of vast volumes of mantle-derived magmas through fractured lithosphere throughout Earth history. Pegmatite dikes comprise extremely coarse crystalline granitic rocks—often associated with late-stage granite intrusions or metamorphic segregations. Aplite dikes are fine-grained or sugary-textured intrusives of granitic composition.



Figure 2. A small dike on the Baranof Cross-Island Trail, Alaska

Sedimentary Dikes



Figure 3. Clastic dike (left of notebook) in the Chinle Formation in Canyonlands National Park, Utah

Sedimentary dikes or clastic dikes are vertical bodies of sedimentary rock that cut off other rock layers. They can form in two ways:

- When a shallow unconsolidated sediment is composed of alternating coarse grained andimpermeable clay layers the fluid pressure inside the coarser layers may reach a critical value due to lithostatic overburden. Driven by the fluid pressure the sediment breaks through overlying layers and forms a dike.
- When a soil is under permafrost conditions the pore water is totally frozen. When cracks are formed in such rocks, they may fill up with sediments that fall in from above. The result is a vertical body of sediment that cuts through horizontal layers: a dike.

Sills

In geology, a **sill** is a tabular sheet intrusion that has intruded between older layers of sedimentary rock, beds of volcanic lava or tuff, or even along the direction of foliation in metamorphic rock. The term sill is synonymous with concordant intrusive sheet.

This means that the sill does not cut across preexisting rocks, in contrast to dikes, discordant intrusive sheets which do cut across older rocks. Sills are fed by dikes, except in unusual locations where they form in nearly vertical beds attached directly to a magma source. The rocks must be brittle and fracture to create the planes along which the magma intrudes the parent rock bodies, whether this occurs along preexisting planes between sedimentary or volcanic beds or weakened planes related to foliation in metamorphic rock. These planes or weakened areas allow the intrusion of a thin sheet-like body of magma paralleling the existing bedding planes, concordant fracture zone, or foliations.

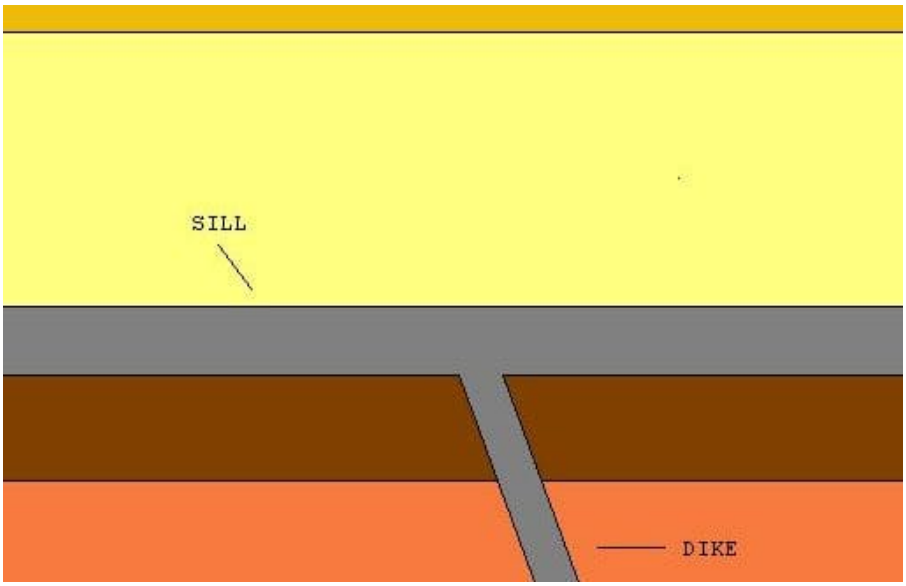


Figure 4. Illustration showing the difference between a dike and a sill.

Sills parallel beds (layers) and foliations in the surrounding country rock. They can be originally emplaced in a horizontal orientation, although tectonic processes may cause subsequent rotation of horizontal sills into near vertical orientations. Sills can be confused with solidified lava flows; however, there are several differences between them. Intruded sills will show partial melting and incorporation of the surrounding country rock. On both contact surfaces of the country rock into which the sill has intruded, evidence of heating will be observed (contact metamorphism). Lava flows will show this evidence only on the lower side of the flow. In addition, lava flows will typically show evidence of vesicles (bubbles) where gases escaped into the atmosphere. Because sills generally form at shallow depths (up to many kilometers) below the surface, the pressure of overlying rock prevents this from happening much, if at all. Lava flows will also typically show evidence of weathering on their upper surface, whereas sills, if still covered by country rock, typically do not.

Associated Ore Deposits



Figure 5. Mid-Carboniferous dolerite sill cutting Lower Carboniferous shales and sandstones, Horton Bluff, Minas Basin South Shore, Nova Scotia

Certain layered intrusions are a variety of sill that often contain important ore deposits. Precambrian examples include the Bushveld, Insizwa and the Great Dyke complexes of southern Africa, the Duluth intrusive complex of the Superior District, and the Stillwater igneous complex of the United States. Phanerozoic examples are usually smaller and include the Rùm peridotite complex of Scotland and the Skaergaard igneous complex of east Greenland. These intrusions often contain concentrations of gold, platinum, chromium and other rare elements.

Transgressive Sills

Despite their concordant nature, many large sills change stratigraphic level within the intruded sequence, with each concordant part of the intrusion linked by relatively short dike-like segments. Such sills are known as transgressive, examples include the Whin Sill and sills within the Karoo basin.[3] The geometry of large sill complexes in sedimentary basins has become clearer with the availability of 3D seismic reflection data.[4] Such data has shown that many sills have an overall saucer shape and that many others are at least in part transgressive.[5]

Other Meanings

“Sill” may also refer to the rise in depth near the mouth of a fjord caused by the terminal moraine of the previous glacier.

---

1. Essentials of Geology, 3rd Ed, Stephen Marshak ↩
2. Pilkington, Mark and Walter R. Roest; *Removing varying directional trends in aeromagnetic data,* Geophysics, vol. 63 no. 2 (1998), pp. 446–453. abstract↩
3. Holmes, A. 1978. *Principles of Physical Geology,* Van Nostrand Reinhold (UK) Co.Ltd., 720 pp. See also, Polteau, S., Mazzini, A., Galland, O., Planke, S. & Malthe-Sørenssen, A. 2008. Saucer-shaped intrusions: Occurrences, emplacement and implications, *Earth and Planetary Science Letters* 266, 195–204. ↩
4. Thomson, K. & Hutton, D. 2004. Geometry and growth of sill complexes: insights using 3D seismic from the North Rockall Trough, *Bulletin of Volcanology, 66*, 364–375. ↩
5. Planke, S., Rasmussen, T., Rey, S.S. & Myklebust, R. 2005. Seismic characteristics and distribution of volcanic intrusions and hydrothermal vent complexes in the Vøring and Møre basins, In: Doré, A. G. & Vining, B. A. (eds) *Petroleum Geology: North-West Europe and Global Perspectives — Proceedings of the 6th Petroleum Geology Conference*, 833–844. Petroleum Geology Conferences Ltd. Published by the Geological Society, London. ↩

Contributors and Attributions

CC licensed content, Shared previously

- Dike (geology). **Provided by**: Wikipedi. **Located at**: https://en.Wikipedia.org/wiki/Dike_%28geology%29. **License**: *CC BY-SA: Attribution-ShareAlike*
- Sill (geology). **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Sill_%28geology%29. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.9.7: Dikes and Sills is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.8: Volcanic Landforms



Why is the Republic of Indonesia made of 17,508 islands?

Around the Pacific Rim is Indonesia, a nation built from the dotted volcanoes of an island arc. Indonesia is distinctive for its rich volcanic soil, tropical climate, tremendous biodiversity, and volcanoes. These volcanoes are in Java, Indonesia.

Landforms from Lava

Volcanoes and Vents

The most obvious landforms created by lava are volcanoes, most commonly as cinder cones, composite volcanoes, and shield volcanoes. Eruptions also take place through other types of vents, commonly from fissures (Figure 1). The eruptions that created the entire ocean floor are essentially fissure eruptions.



Figure 1. A fissure eruption on Mauna Loa in Hawaii travels toward Mauna Kea on the Big Island.

Lava Domes

Viscous lava flows slowly. If there is not enough magma or enough pressure to create an explosive eruption, the magma may form a lava dome. Because it is so thick, the lava does not flow far from the vent. (Figure 2).



Figure 2. Lava domes are large, round landforms created by thick lava that does not travel far from the vent.

Lava flows often make mounds right in the middle of craters at the top of volcanoes, as seen in the Figure 3.



Figure 3. Lava domes may form in the crater of composite volcanoes as at Mount St. Helens.

Lava Plateaus

A lava plateau forms when large amounts of fluid lava flow over an extensive area (Figure 4). When the lava solidifies, it creates a large, flat surface of igneous rock.



Figure 4. Layer upon layer of basalt have created the Columbia Plateau, which covers more than 161,000 square kilometers (63,000 square miles) in Washington, Oregon, and Idaho.

Land

Lava creates new land as it solidifies on the coast or emerges from beneath the water (Figure 5).



Figure 5. Lava flowing into the sea creates new land in Hawaii.

Over time the eruptions can create whole islands. The Hawaiian Islands are formed from shield volcano eruptions that have grown over the last 5 million years (Figure 6).

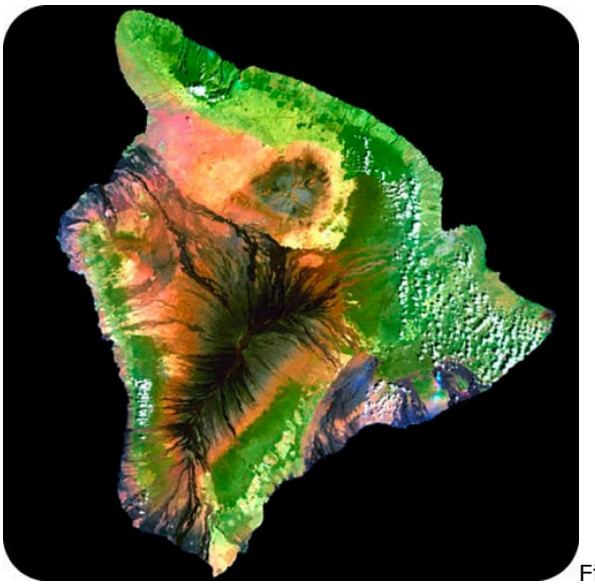


Figure 6. The island of Hawaii was created by hotspot volcanism. You can see some of the volcanoes (both active and extinct) in this mosaic of false-color composite satellite images.

Landforms from Magma

Magma intrusions can create landforms. Shiprock in New Mexico is the neck of an old volcano that has eroded away (Figure 7). The volcanic neck is the remnant of the conduit the magma traveled up to feed an eruption.



Figure 7. The aptly named Shiprock in New Mexico.

Summary

- Landforms created by lava include volcanoes, domes, and plateaus.
- New land can be created by volcanic eruptions.
- Landforms created by magma include volcanic necks and domes.

Contributors and Attributions

CC licensed content, Shared previously

- Volcanic Landforms. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Volcanic-Landforms/lesson/Volcanic-Landforms-HS-ES. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.9.8: Volcanic Landforms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.9: Igneous Features

Check Your Understanding

Answer the question(s) below to see how well you understand the topics covered in the previous section. This short quiz does **not** count toward your grade in the class, and you can retake it an unlimited number of times.

Use this quiz to check your understanding and decide whether to (1) study the previous section further or (2) move on to the next section.

https://assessments.lumenlearning.co...sessments/1114

Contributors and Attributions

CC licensed content, Original

- **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 4.9.9: Igneous Features is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.10: Volcanic Gases, Pyroclastic Flow, and Tephra

Volcanic Gases and Their Effects

Magma contains dissolved gases that are released into the atmosphere during eruptions. Gases are also released from magma that either remains below ground (for example, as an intrusion) or is rising toward the surface. In such cases, gases may escape continuously into the atmosphere from the soil, volcanic vents, fumaroles, and hydrothermal systems.

At high pressures deep beneath the earth's surface, volcanic gases are dissolved in molten rock. But as magma rises toward the surface where the pressure is lower, gases held in the melt begin to form tiny bubbles. The increasing volume taken up by gas bubbles makes the magma less dense than the surrounding rock, which may allow the magma to continue its upward journey. Closer to the surface, the bubbles increase in number and size so that the gas volume may exceed the melt volume in the magma, creating a magma foam. The rapidly expanding gas bubbles of the foam can lead to explosive eruptions in which the melt is fragmented into pieces of volcanic rock, known as tephra. If the molten rock is not fragmented by explosive activity, a lava flow will be generated.

Together with the tephra and entrained air, volcanic gases can rise tens of kilometers into Earth's atmosphere during large explosive eruptions. Once airborne, the prevailing winds may blow the eruption cloud hundreds to thousands of kilometers from a volcano. The gases spread from an erupting vent primarily as acid aerosols (tiny acid droplets), compounds attached to tephra particles, and microscopic salt particles.

Volcanic gases undergo a tremendous increase in volume when magma rises to the Earth's surface and erupts. For example, consider what happens if one cubic meter of 900°C rhyolite magma containing five percent by weight of dissolved water were suddenly brought from depth to the surface. The one cubic meter of magma now would occupy a volume of 670 m^3 as a mixture of water vapor and magma at atmospheric pressure (Sparks et. al., 1997)! The one meter cube at depth would increase to 8.75 m on each side at the surface. Such enormous expansion of volcanic gases, primarily water, is the main driving force of explosive eruptions.

The most abundant gas typically released into the atmosphere from volcanic systems is water vapor (H_2O), followed by carbon dioxide (CO_2) and sulfur dioxide (SO_2). Volcanoes also release smaller amounts of others gases, including hydrogen sulfide (H_2S), hydrogen (H_2), carbon monoxide (CO), hydrogen chloride (HCL), hydrogen fluoride (HF), and helium (He).

Examples of volcanic gas compositions, in volume percent concentrations (from Symonds et. al., 1994)

Volcano Tectonic Style Temperature	**Kilauea Summit** Hot Spot 1170°C	**Erta` Ale** Divergent Plate 1130°C	**Momotombo** Convergent Plate 820°C
H_2O	37.1	77.2	97.1
CO_2	48.9	11.3	1.44
SO_2	11.8	8.34	0.50
H_2	0.49	1.39	0.70
CO	1.51	0.44	0.01
H_2S	0.04	0.68	0.23
HCl	0.08	0.42	2.89
HF	—	—	0.26

The volcanic gases that pose the greatest potential hazard to people, animals, agriculture, and property are sulfur dioxide, carbon dioxide, and hydrogen fluoride. Locally, sulfur dioxide gas can lead to acid rain and air pollution downwind from a volcano. Globally, large explosive eruptions that inject a tremendous volume of sulfur aerosols into the stratosphere can lead to lower

surface temperatures and promote depletion of the Earth's ozone layer. Because carbon dioxide gas is heavier than air, the gas may flow into in low-lying areas and collect in the soil. The concentration of carbon dioxide gas in these areas can be lethal to people, animals, and vegetation. A few historic eruptions have released sufficient fluorine-compounds to deform or kill animals that grazed on vegetation coated with volcanic ash; fluorine compounds tend to become concentrated on fine-grained ash particles, which can be ingested by animals.

Sulfur dioxide (SO_2)

The effects of SO_2 on people and the environment vary widely depending on (1) the amount of gas a volcano emits into the atmosphere; (2) whether the gas is injected into the troposphere or stratosphere; and (3) the regional or global wind and weather pattern that disperses the gas. Sulfur dioxide (SO_2) is a colorless gas with a pungent odor that irritates skin and the tissues and mucous membranes of the eyes, nose, and throat. Sulfur dioxide chiefly affects upper respiratory tract and bronchi. The World Health Organization recommends a concentration of no greater than 0.5 ppm over 24 hours for maximum exposure. A concentration of 6-12 ppm can cause immediate irritation of the nose and throat; 20 ppm can cause eye irritation; 10,000 ppm will irritate moist skin within minutes.

Emission rates of SO_2 from an active volcano range from <20 tonnes/day to >10 million tonnes/day according to the style of volcanic activity and type and volume of magma involved. For example, the large explosive eruption of Mount Pinatubo on 15 June 1991 expelled 3-5 km^3 of dacite magma and injected about 20 million metric tons of SO_2 into the stratosphere. The sulfur aerosols resulted in a 0.5-0.6°C cooling of the Earth's surface in the Northern Hemisphere. The sulfate aerosols also accelerated chemical reactions that, together with the increased stratospheric chlorine levels from human-made chlorofluorocarbon (CFC) pollution, destroyed ozone and led to some of the lowest ozone levels ever observed in the atmosphere.

At Kilauea Volcano, the recent effusive eruption of about 0.0005 km^3/day (500,000 m^3) of basalt magma releases about 2,000 tonnes of SO_2 into the lower troposphere. Downwind from the vent, acid rain and air pollution is a persistent health problem when the volcano is erupting.

- **SO_2 causes air pollutionVolcanic smog.** Eruptions of Kilauea Volcano release large quantities of sulfur dioxide gas into the atmosphere that can lead to volcanic air pollution on the Island of Hawai`i. Sulfur dioxide gas reacts chemically with sunlight, oxygen, dust particles, and water to form volcanic smog known as vog.
- **SO_2 effects Earth's surface temperatureGlobal cooling and ozone depletion.** Measurements from recent eruptions such as Mount St. Helens, Washington (1980), El Chichon, Mexico (1982), and Mount Pinatubo, Philippines (1991), clearly show the importance of sulfur aerosols in modifying climate, warming the stratosphere, and cooling the troposphere. Research has also shown that the liquid drops of sulfuric acid promote the destruction of the Earth's ozone layer.

Hydrogen sulfide (H_2S)

Hydrogen sulfide (H_2S) is a colorless, flammable gas with a strong offensive odor. It is sometimes referred to as sewer gas. At low concentrations it can irritate the eyes and acts as a depressant; at high concentrations it can cause irritation of the upper respiratory tract and, during long exposure, pulmonary edema. A 30-minute exposure to 500 ppm results in headache, dizziness, excitement, staggering gait, and diarrhea, followed sometimes by bronchitis or bronchopneumonia.

Carbon dioxide (CO_2)

Volcanoes release more than 130 million tonnes of CO_2 into the atmosphere every year. This colorless, odorless gas usually does not pose a direct hazard to life because it typically becomes diluted to low concentrations very quickly whether it is released continuously from the ground or during episodic eruptions. But in certain circumstances, CO_2 may become concentrated at levels lethal to people and animals. Carbon dioxide gas is heavier than air and the gas can flow into in low-lying areas; breathing air with more than 30% CO_2 can quickly induce unconsciousness and cause death. In volcanic or other areas where CO_2 emissions occur, it is important to avoid small depressions and low areas that might be CO_2 traps. The boundary between air and lethal gas can be extremely sharp; even a single step upslope may be adequate to escape death.

CO_2 trapped in depressions can be lethal to people and animals. When a burning piece of cloth is lowered into a hole that has a high concentration of CO_2, the fire goes out. Such a condition can be lethal to people and animals.

Air with 5% CO_2 causes perceptible increased respiration; 6-10% results in shortness of breath, headaches, dizziness, sweating, and general restlessness; 10-15% causes impaired coordination and abrupt muscle contractions; 20-30% causes loss of consciousness and convulsions; over 30% can cause death (Hathaway et. al., 1991).

Please see the web article, "Volcanic Gases and Climate Change Overview" for more information on Volcanic versus anthropogenic CO_2emissions.

Here is a historical example of the effects of carbon dioxide gas: Mammoth Mountain in Long Valley Caldera, California kills trees near Mammoth Mountain, California

Hydrogen Chloride (HCl)

Chlorine gas is emitted from volcanoes in the form of hydrochloric acid (HCl). Exposure to the gas irritates mucous membranes of the eyes and respiratory tract. Concentrations over 35 ppm cause irritation of the throat after short exposure; >100 ppm results in pulmonary edema, and often laryngeal spasm. It also causes acid rain downwind from volcanoes because HCl is extremely soluble in condensing water droplets and it is a very "strong acid" (it dissociates extensively to give H^+ ions in the droplets).

Hydrogen Fluoride (HF)

Fluorine is a pale yellow gas that attaches to fine ash particles, coats grass, and pollutes streams and lakes. Exposure to this powerful caustic irritant can cause conjunctivitis, skin irritation, bone degeneration and mottling of teeth. Excess fluorine results in a significant cause of death and injury in livestock during ash eruptions. Even in areas that receive just a millimeter of ash, poisoning can occur where the fluorine content of dried grass exceeds 250 ppm. Animals that eat grass coated with fluorine-tainted ash are poisoned. Small amounts of fluorine can be beneficial, but excess fluorine causes fluorosis, an affliction that eventually kills animals by destroying their bones. It also promotes acid rain effects downwind of volcanoes, like HCl.

Secondary Gas Emissions

Another type of gas release occurs when lava flows reach the ocean. Extreme heat from molten lava boils and vaporizes seawater, leading to a series of chemical reactions. The boiling and reactions produce a large white plume, locally known as lava haze or laze, containing a mixture of hydrochloric acid and concentrated seawater.

- **Laze plumes are very acidic.** Extreme heat from lava entering the sea rapidly boils and vaporizes seawater, leading to a series of chemical reactions. The boiling and reactions produce a large white plume, locally known as lava haze or laze, which contains a mixture of hydrochloric acid (HCl) and concentrated seawater. This is a short-lived local phenomenon that only affects people or vegetation directly under the plume.
  The hydrochloric acid (HCl) comes from the breakdown of seawater-derived chlorides during sudden boiling. Because the lava is largely degassed by the time it reaches the sea, any HCL coming from it is insignificant by comparison. Analyzed samples of the plume show that is is a brine with a salinity of about 2.3 times that of seawater and a pH of 1.5-2.0.
- **Key seawater chloride breakdown reactions that produce HCl gas**
  - $MgCl_2$ (sea salt) + H_2O (steam) = MgO (periclase) + 2HCl (HCl gas)
  - 2 NaCl (sea salt) + H_2O (steam) = Na_2O (sodium oxide) + 2 HCL (HCl gas)
  - $CaCl_2$ (sea salt) + H_2O (steam) = CaO (lime) + 2 HCL (HCl gas)
- **Avoid standing beneath a laze plume.** Dense laze plumes, such as that shown here (Photograph by C.C. Heliker, February 10, 1994) contain as much as 10-15 parts per million of hydrochloric acid. These values drop off sharply as the plume moves away from the lava entry areas. During along-shore or on-shore winds, this plume produces acid rain that may fall on people and land along the coast. This rain (pH 1.5 to 2), often more acidic that lime juice or stomach acid, is very corrosive to the skin and clothing. Visitors to the lava entry areas should avoid standing directly in, under, or downwind of the laze plume.

Pyroclastic Flows and Their Effects

About Pyroclastic Flows

Pyroclastic flows are high-density mixtures of hot, dry rock fragments and hot gases that move away from the vent that erupted them at high speeds. They may result from the explosive eruption of molten or solid rock fragments, or both. They may also result

from the nonexplosive eruption of lava when parts of dome or a thick lava flow collapses down a steep slope. Most pyroclastic flows consist of two parts: a basal flow of coarse fragments that moves along the ground, and a turbulent cloud of ash that rises above the basal flow. Ash may fall from this cloud over a wide area downwind from the pyroclastic flow.


Figure 1. Pyroclastic Flows destroy by direct impact.

Figure 1 shows a remnant of a building in Francisco Leon that was destroyed by pyroclastic flows and surges during the eruption of El Chichon volcano in southeastern Mexico between March 29 and April 4, 1982. Francisco Leon was located about 5 km SSE of the volcano. The reinforcement rods in the concrete wall are bent in the direction of flow (right to left).

Several pyroclastic surges swept down all sides of the volcano, reaching as far as 2 to 8 km from the volcano; pyroclastic flows represent only about 30 percent of the total combined mass of surges and flows. In addition to Francisco Leon, eight other villages were destroyed, claiming the lives of up to 2,000 people.


Figure 2. Pyroclastic Flows bury sites with hot rock debris.

Pyroclastic flows erupted by Mount Pinatubo on June 15, 1991, buried the Marella River valley (SW of Pinatubo) with pumice, ash, and other volcanic rocks to depths of between 50 and 200 m (as shown in figure 2). This eruption was one of the largest in the 20th century, depositing about 5.5 km^3 of rock debris over nearly 400 km^2. The pyroclastic flows traveled as far as 12 to 16 km from the volcano in all directions. Unlike river valleys, the steep slopes around the volcano were veneered with very thin, discontinuous pyroclastic-flow deposits.

A series of pyroclastic flows from Redoubt Volcano in Alaska between December 1989 and April 1990 rapidly melted snow and ice that generated lahars in Drift River (valley in figure 3).

 Figure 3. Pyroclastic Flows melt snow and ice to form lahars.

The lahars swept 40 km to Cook Inlet. Most of the pyroclastic flows were caused by the repeated collapse of a lava dome growing high on the volcano's north flank.

This view across the upper Drift River valley is to the SW; the north flank is on the right side of the volcano, but the dome is not visible. As each pyroclastic flow swept down the volcano's snow- and glacier-covered north flank, the hot lava-dome fragments eroded and mixed with the snow and ice to form a torrent of water that swept into Drift River. The sudden surges of water eroded loose sediment on the valley floor and transformed into lahars.

In this part of the Drift River valley, many of the lahars covered the entire valley floor, but they were generally no more than a few meters thick. The lahars provided an excellent opportunity to test a new lahar detection system.

 Figure 4. Pyroclastic Flows burn forests, crops, and buildings.

A small but highly fluidized pyroclastic flow traveled down the narrow Belham stream channel as far as 5 to 6 km from a lava dome growing at the summit of Soufriere Hills volcano in Montserrat (shown in figure 4). The basal part of the pyroclastic flow was confined to the channel bottom, but the overriding hot ash cloud burned and killed vegetation along the channel. The pyroclastic flow was triggered by a partial collapse of the dome. With initial temperatures higher than 900°C, the rock debris and gas carried by a pyroclastic flow or surge easily remain hot enough to burn or singe combustible material even when they travel more than 20 km from a vent.

Effects of Pyroclastic Flows

A pyroclastic flow will destroy nearly everything in its path. With rock fragments ranging in size from ash to boulders traveling across the ground at speeds typically greater than 80 km per hour, pyroclastic flows knock down, shatter, bury or carry away nearly all objects and structures in their way. The extreme temperatures of rocks and gas inside pyroclastic flows, generally between 200°C and 700°C, can cause combustible material to burn, especially petroleum products, wood, vegetation, and houses.

Pyroclastic flows vary considerably in size and speed, but even relatively small flows that move less than 5 km from a volcano can destroy buildings, forests, and farmland. And on the margins of pyroclastic flows, death and serious injury to people and animals

may result from burns and inhalation of hot ash and gases.

Pyroclastic flows generally follow valleys or other low-lying areas and, depending on the volume of rock debris carried by the flow, they can deposit layers of loose rock fragments to depths ranging from less than one meter to more than 200 m. Such loose layers of ash and volcanic rock debris in valleys and on hillslopes can lead to lahars indirectly by:

1. Damming or blocking tributary streams, which may cause water to form a lake behind the blockage, overtop and erode the blockage, and mix with the rock fragments as it rushes downstream (for example, see this case study at Pinatubo Volcano, Philippines)
2. Increasing the rate of stream runoff and erosion during subsequent rainstorms. Hot pyroclastic flows and surges can also directly generate lahars by eroding and mixing with snow and ice on a volcano's flanks, thereby sending a sudden torrent of water surging down adjacent valleys (see case study from Nevado del Ruiz volcano, Colombia).

Tephra and Their Effects

Tephra is a general term for fragments of volcanic rock and lava regardless of size that are blasted into the air by explosions or carried upward by hot gases in eruption columns or lava fountains. Such fragments range in size from less than 2 mm (ash) to more than 1 m in diameter. Large-sized tephra typically falls back to the ground on or close to the volcano and progressively smaller fragments are carried away from the vent by wind. Volcanic ash, the smallest tephra fragments, can travel hundreds to thousands of kilometers downwind from a volcano.

Tephra consists of a wide range of rock particles (size, shape, density, and chemical composition), including combinations of pumice, glass shards, crystals from different types of minerals, and shattered rocks of all types (igneous, sedimentary, and metamorphic). A great variety of terms are used to describe the range of rock fragments thrown into the air by volcanoes. The terms classify the fragments according to size, shape, or the way in which they form and travel.

Mount St. Helens Tephra: block	Mount St. Helens Tephra: ash & pumice
	

Volcanic Ash

How far will it fall downwind from an erupting volcano?

Ash usually covers a much larger area and disrupts the lives of far more people than the other more lethal types of volcano hazards. Unfortunately, the size of ash particles that fall to the ground and the thickness of ashfall downwind from an erupting volcano are difficult to predict in advance. Not only is there a wide range in the size of an eruption that might occur and the amount of tephra injected into the atmosphere, but the direction and strength of the prevailing wind can vary widely.

Case Histories: Tephra Distribution Downwind from Eruption

- Tephra Falls of the 1991 Eruptions of Mount Pinatubo
  online report from Fire and Mud: eruptions and lahars of Mount Pinatubo, Philippines.

- Tephra fall from Mount St. Helens, Washington, on May 18, 1980
  graph, ash thickness and particle size downwind from volcano

Potential Effects of Volcanic Ash

Volcanic ash is highly disruptive to economic activity because it covers just about everything, infiltrates most openings, and is highly abrasive. Airborne ash can obscure sunlight to cause temporary darkness and reduce visibility to zero. Ash is slippery, especially when wet; roads, highways, and airport runways may become impassable. Automobile and jet engines may stall from ash-clogged air filters and moving parts can be damaged from abrasion, including bearings, brakes, and transmissions.

More effects include the following:

- Daylight turns into darkness.
- Roofs may collapse from added weight.
- Machinery and vehicles will be abraded.
- Farmland will be covered.
- Roads will be slippery, blocked, or blocked.
- Power systems may shut down.
- Waste-water systems may clog.
- Gutters may fill and collapse.

Contributors and Attributions

Public domain content

- Types of Volcano Hazards. **Provided by**: USGS. **Located at**: http://volcanoes.usgs.gov/hazards/. **License**: *Public Domain: No Known Copyright*

This page titled 4.9.10: Volcanic Gases, Pyroclastic Flow, and Tephra is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.11: Lahars, Landslides, and Lavaflow

Lahars and Their Effects



Figure 1. Gualí River valley. Photograph by R. Janda on December 18, 1985

Lahar is an Indonesian term that describes a hot or cold mixture of water and rock fragments flowing down the slopes of a volcano and (or) river valleys. When moving, a lahar looks like a mass of wet concrete that carries rock debris ranging in size from clay to boulders more than 10 m in diameter. Lahars vary in size and speed. Small lahars less than a few meters wide and several centimeters deep may flow a few meters per second. Large lahars hundreds of meters wide and tens of meters deep can flow several tens of meters per second–much too fast for people to outrun.

As a lahar rushes downstream from a volcano, its size, speed, and the amount of water and rock debris it carries constantly change. The beginning surge of water and rock debris often erodes rocks and vegetation from the side of a volcano and along the river valley it enters. This initial flow can also incorporate water from melting snow and ice (if present) and the river it overruns. By eroding rock debris and incorporating additional water, lahars can easily grow to more than 10 times their initial size. But as a lahar moves farther away from a volcano, it will eventually begin to lose its heavy load of sediment and decrease in size.

Eruptions may trigger one or more lahars directly by quickly melting snow and ice on a volcano or ejecting water from a crater lake. More often, lahars are formed by intense rainfall during or after an eruption–rainwater can easily erode loose volcanic rock and soil on hillsides and in river valleys. Some of the largest lahars begin as landslides of saturated and hydrothermally altered rock on the flank of a volcano or adjacent hillslopes. Landslides are triggered by eruptions, earthquakes, precipitation, or the unceasing pull of gravity on the volcano.

Lahars almost always occur on or near stratovolcanoes because these volcanoes tend to erupt explosively and their tall, steep cones are either snow covered, topped with a crater lake, constructed of weakly consolidated rock debris that is easily eroded, or internally weakened by hot hydrothermal fluids. Lahars are also common from the snow- and ice-covered shield volcanoes in Iceland where eruptions of fluid basalt lava frequently occur beneath huge glaciers.

The scenarios listed below illustrate most of the mechanisms by which lahars are generated. For convenience, we've grouped the mechanisms according to whether a volcano is erupting, has erupted, or is quiet. Each mechanism is illustrated with one or more case studies.

Volcano Landslides and Their Effects

Landslides are large masses of rock and soil that fall, slide, or flow very rapidly under the force of gravity. These mixtures of debris move in a wet or dry state, or both. Landslides commonly originate as massive rockslides or avalanches which disintegrate during movement into fragments ranging in size from small particles to enormous blocks hundreds of meters across. If the moving rock debris is large enough and contains a large content of water and fine material (typically, >3-5 percent of clay-sized particles), the landslide may transform into a lahar and flow downvalley more than 100 km from a volcano!

Volcano landslides range in size from less than 1 km3 to more than 100 km3. The high velocity (>100 km/hr) and great momentum of landslides allows them to run up slopes and to cross valley divides up to several hundred meters high. For example, the landslide at Mount St. Helens on May 18, 1980, had a volume of 2.5 km3, reached speeds of 50-80 m/s (180-288 km/hr), and surged up and over a 400 m tall ridge located about 5 km from the volcano!

Landslides are common on volcanoes because their massive cones (1) typically rise hundreds to thousands of meters above the surrounding terrain; and (2) are often weakened by the very process that created them–the rise and eruption of molten rock. Each time magma moves toward the surface, overlying rocks are shouldered aside as the molten rock makes room for itself, often creating internal shear zones or oversteepening one or more sides of the cone. Magma that remains within the cone releases volcanic gases that partially dissolve in groundwater, resulting in a hot acidic hydrothermal system that weakens rock by altering rock minerals to clay. Furthermore, the tremendous mass of thousands of layers lava and loose fragmented rock debris can lead to internal faults and fault zones that move frequently as the cone "settles" under the downward pull of gravity.

These conditions permit a number of factors to trigger a landslide or to allow part of a volcano's cone to simply collapse under the influence of gravity:

- intrusion of magma into a volcano
- explosive eruptions (magmatic or phreatic–steam-driven explosions)
- large earthquake directly beneath a volcano or nearby (typically >M5)
- intense rainfall that saturates a volcano or adjacent tephra-covered hillslopes with water, especially before or during a large earthquake.

A landslide typically destroys everything in its path and may generate a variety of related activity. Historically, landslides have caused explosive eruptions, buried river valleys with tens of meters of rock debris, generated lahars, triggered waves and tsunami, and created deep horseshoe-shaped craters.

By removing a large part of a volcano's cone, a landslide may abruptly decrease pressure on the shallow magmatic and hydrothermal systems, which can generate explosions ranging from a small steam explosion to large steam- and magma-driven directed blasts. A large landslide often buries valleys with tens to hundreds of meters of rock debris, forming a chaotic landscape marked by dozens of small hills and closed depressions. If the deposit is thick enough, it may dam tributary streams to form lakes in the subsequent days to months; the lakes may eventually drain catastrophically and generate lahars and floods downstream.

Landslides also generate some of the largest and most deadly lahars, either by transforming directly into a lahar or, after it stops moving, from dewatering of the deposit. Historically, however, the most deadly volcano landslide occurred in 1792 when sliding debris from Mt. Mayuyama near Unzen Volcano in Japan slammed into the Ariaka Sea and generated a wave on the opposite side that killed nearly 15,000 people.

On a volcano, landslides typically carve deep gashes into its cone or create large horseshoe-shaped craters hundreds of meters deep and more than a kilometer in width.

Volcanic landslides can do the following:

- Trigger volcanic explosions
- Generate lahars that travel far downstream
- Cause waves and tsunamis in a lake or ocean
- Bury river valleys with rock debris
- Dam tributary streams to form lakes
- Create a crater or scar on the volcano

Lava Flows and Their Effects

Lava flows are streams of molten rock that pour or ooze from an erupting vent. Lava is erupted during either nonexplosive activity or explosive lava fountains. Lava flows destroy everything in their path, but most move slowly enough that people can move out of the way. The speed at which lava moves across the ground depends on several factors, including (1) type of lava erupted and its viscosity; (2) steepness of the ground over which it travels; (3) whether the lava flows as a broad sheet, through a confined channel, or down a lava tube; and (4) rate of lava production at the vent.

Fluid basalt flows can extend tens of kilometers from an erupting vent. The leading edges of basalt flows can travel as fast as 10

km/hour on steep slopes but they typically advance less than 1 km/hour on gentle slopes. But when basalt lava flows are confined within a channel or lava tube on a steep slope, the main body of the flow can reach velocities >30 km/hour.

Viscous andesite flows move only a few kilometers per hour and rarely extend more than 8 km from their vents. Viscous dacite and rhyolite flows often form steep-sided mounds called lava domes over an erupting vent. Lava domes often grow by the extrusion of many individual flows >30 m thick over a period of several months or years. Such flows will overlap one another and typically move less than a few meters per hour.



Figure 2. `A`a lava flow moves through an intersection in the Royal Garderns subdivision on the south flank of Kilauea Volcano, Hawai`i. Photograph by J.D. Griggs in 1984.

Everything in the path of an advancing lava flow will be knocked over, surrounded, or buried by lava, or ignited by the extremely hot temperature of lava. When lava erupts beneath a glacier or flows over snow and ice, meltwater from the ice and snow can result in far-reaching lahars. If lava enters a body of water or water enters a lava tube, the water may boil violently and cause an explosive shower of molten spatter over a wide area. Methane gas, produced as lava buries vegetation, can migrate in subsurface voids and explode when heated. Thick viscous lava flows, especially those that build a dome, can collapse to form fast-moving pyroclastic flows.

Deaths caused directly by lava flows are uncommon because most move slowly enough that people can move out the way easily and flows usually don't travel far from the vent. Death and injury can result when onlookers approach an advancing lava flow too closely or their retreat is cut off by other flows. Deaths attributed to lava flows are often due to related causes, such as explosions when lava interacts with water, the collapse of an active lava delta, asphyxiation due to accompanying toxic gases, pyroclastic flows from a collapsing dome, and lahars from meltwater.

Other natural phenomena such as hurricanes, tornadoes, tsunami, fires, and earthquakes often destroy buildings, agricultural crops, and homes, but the owner(s) can usually rebuild or repair structures and their businesses in the same location. Lava flows, however, can bury homes and agricultural land under tens of meters of hardened black rock; landmarks and property lines become obscured by a vast, new hummocky landscape. People are rarely able to use land buried by lava flows or sell it for more than a small fraction of its previous worth.


Figure 3. Lava buries or surrounds everything.

Lava erupted from Kilauea Volcano covers part of Highway 130 on the southeast coast of the Island of Hawai`i (see figure 3). By the end of 1998, lava from the Pu'u' 'O'o-Kupaianaha eruption covered about 13 km of the highway to depths as great as 25 m. Between 1983 and 1998, lava flows covered an area of 99.7 km^2 (38.5 mi^2).

Because lava flows can completely block roads and highways that may serve as the only evacuation route for people threatened by an advancing flow, it is vital for communities that could be inundated with lava to develop emergency-response plans.

One of the chief threats of lava flows to property owners is that the flows may burn buildings and homes even if the flow doesn't reach the structure. This house caught fire from the intense heat of an advancing `a`a flow (note red glow of flow left of the house).

Basalt has the highest temperature of any lava, typically between about 1170-1100°C (~2140-2000°F). The other lava types (andesite, dacite, and rhyolite) form cooler flows with temperatures between about 1000-800°C (~1800-1500°F); some flows can still move slowly at temperatures as low as about 600°C (~1100°F).


Figure 4. View of eruption crater and ash-covered Vatnajökull glacier about 36 hours after the eruption had broken through the ice.

Lava melts snow and ice to form jökulhlaups and lahars. By this time, the area of subsidence had grown to about 9 km long and 2-3 km wide. The eruption continued for about another 10 days, and meltwater from the glacier flowed into the Grímsvötn caldera. On October 1, water level in the caldera's subglacial lake was about 1410 m; by October 16, the water level had risen to 1504 m, an increase of 94 m! According to scientists monitoring the activity, lava erupting from the fissure was piled up on the ground beneath the glacier, "forming a mountain ridge which in places is expected to be 200 m high."

On October 16, scientists stated that the meltwater, which had been accumulating under the ice shelf in the Grímsvötn caldera lake, could begin draining at any time to trigger a jökulhlaup (glacial outburst flood). On November 5 the expected jökulhlaup began.

Collapsing lava flows trigger pyroclastic flows. The fast-moving pyroclastic flow was caused by the collapse of a thick lava flow that was extruding from the summit area and oozing down the volcano's steep upper cone. When the lava flow collapsed, the hot lava broke apart into fragments ranging in size from boulders to tiny ash particles and swept down the volcano under the influence of gravity to form the pyroclastic flow; the flow reached a maximum distance of 4.5 km from the summit.

Contributors and Attributions

Public domain content

- Types of Volcano Hazards. **Provided by**: USGS. **Located at**: http://volcanoes.usgs.gov/hazards/. **License**: *Public Domain: No Known Copyright*

This page titled 4.9.11: Lahars, Landslides, and Lavaflow is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.12: Signs of an Eruption

In 2005 USGS geologist Chris Newhall made a list of the six most important signs of an imminent volcanic eruption. They are as follows:

1. *Gas leaks* — the release of gases (mostly H_2O, CO_2, and SO_2) from the magma into the atmosphere through cracks in the overlying rock
2. *Bit of a bulge* — the deformation of part of the volcano, indicating that a magma chamber at depth is swelling or becoming more pressurized
3. *Getting shaky* — many (hundreds to thousands) of small earthquakes, indicating that magma is on the move. The quakes may be the result of the magma forcing the surrounding rocks to crack, or a harmonic vibration that is evidence of magmatic fluids moving underground.
4. *Dropping fast* — a sudden decrease in the rate of seismicity, which may indicate that magma has stalled, which could mean that something is about to give way
5. *Big bump* — a pronounced bulge on the side of the volcano (like the one at Mt. St. Helens in 1980), which may indicate that magma has moved close to surface
6. *Blowing off steam* — steam eruptions (a.k.a. **phreaticeruptions**) that happen when magma near the surface heats groundwater to the boiling point. The water eventually explodes, sending fragments of the overlying rock far into the air.

Contributors and Attributions

CC licensed content, Shared previously

- 4.5 Monitoring Volcanoes and Predicting Eruptions. **Authored by**: Steven Earle. **Provided by**: BC Campus. **Located at**: https://opentextbc.ca/geology/chapter/4-5-monitoring-volcanoes-and-predicting-eruptions/. **Project**: Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.9.12: Signs of an Eruption is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.9.13: Tools to Monitor Volcanoes

Here is a list of the equipment we should have and the actions we can take to monitor a volcano and predict when it might erupt.

Assessing Seismicity



Figure 1. A seismometer installed in 2007 in the vicinity of the Nazco Cone, B.C. [photo Cathie Hickson, used with permission]

The simplest and cheapest way to monitor a volcano is with seismometers. In an area with several volcanoes that have the potential to erupt (e.g., the Squamish-Pemberton area), a few well-placed seismometers can provide us with an early warning that something is changing beneath one of the volcanoes, and that we need to take a closer look. There are currently enough seismometers in the Lower Mainland and on Vancouver Island to provide this information.[1]

If there is seismic evidence that a volcano is coming to life, more seismometers should be placed in locations within a few tens of kilometers of the source of the activity (Figure 1). This will allow geologists to determine the exact location and depth of the seismic activity so that they can see where the magma is moving.

Detecting Gases

Water vapor quickly turns into clouds of liquid water droplets and is relatively easy to detect just by looking, but CO_2 and SO_2 are not as obvious. It's important to be able to monitor changes in the composition of volcanic gases, and we need instruments to do that. Some can be monitored from a distance (from the ground or even from the air) using infrared devices, but to obtain more accurate data, we need to sample the air and do chemical analysis. This can be achieved with instruments placed on the ground close to the source of the gases, or by collecting samples of the air and analyzing them in a lab.

Measuring Deformation



Figure 2. A GPS unit installed at Hualalai volcano, Hawaii. The dish-shaped antenna on the right is the GPS receiver. The antenna on the left is for communication with a base station. [from USGS at: http://hvo.wr.usgs.gov/volcanowatch/view.php?id=173]

There are two main ways to measure ground deformation at a volcano. One is known as a **tiltmeter**, which is a sensitive three-directional level that can sense small changes in the tilt of the ground at a specific location. Another is through the use of GPS (global positioning system) technology (Figure 2). GPS is more effective than a tiltmeter because it provides information on how far the ground has actually moved — east-west, north-south, and up-down.

By combining information from these types of sources, along with careful observations made on the ground and from the air, and a thorough knowledge of how volcanoes work, geologists can get a good idea of the potential for a volcano to erupt in the near future (months to weeks, but not days). They can then make recommendations to authorities about the need for evacuations and restricting transportation corridors. Our ability to predict volcanic eruptions has increased dramatically in recent decades because of advances in our understanding of how volcanoes behave and in monitoring technology. Providing that careful work is done, there is no longer a large risk of surprise eruptions, and providing that public warnings are issued and heeded, it is less and less likely that thousands will die from sector collapse, pyroclastic flows, ash falls, or lahars. Indirect hazards are still very real, however, and we can expect the next eruption like the one at Laki in 1783 to take an even greater toll than it did then, especially since there are now roughly eight times as many people on Earth.

Think about It: **Volcano Alert!**

You're the chief volcanologist for the Geological Survey of Canada (GSC), based in Vancouver. At 10:30 a.m. on a Tuesday, you receive a report from a seismologist at the GSC in Sidney saying that there has been a sudden increase in the number of small earthquakes in the vicinity of Mt. Garibaldi. You have two technicians available, access to some monitoring equipment, and a four-wheel-drive vehicle. At noon, you meet with your technicians and a couple of other geologists. By the end of the day, you need to have a plan to implement, starting tomorrow morning, and a statement to release to the press. What should your first day's fieldwork include? What should you say later today in your press release?

---

1. See: http://www.earthquakescanada.nrcan.gc.ca↵

Contributors and Attributions

CC licensed content, Shared previously

- 4.5 Monitoring Volcanoes and Predicting Eruptions. **Authored by**: Steven Earle. **Provided by**: BC Campus. **Located at**: https://opentextbc.ca/geology/chapter/4-5-monitoring-volcanoes-and-predicting-eruptions/. **Project**: Physical Geology.

License: *CC BY: Attribution*

This page titled 4.9.13: Tools to Monitor Volcanoes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.10: MASS MOVEMENT

Learning Objectives

Define and interpret the causes, types, effects and hazards of mass movement and land subsidence.

- Identify and describe types of mass movement.
- Identify the various causes of mass movement, as well as ways to diminish the damage caused.

4.10.1: Slides

4.10.2: Falls

4.10.3: Slump and Creep

4.10.4: Natural Mass Wasting

4.10.5: Human Impact on Mass Wasting

This page titled Chapter 4.10: Mass Movement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.10.1: Slides

 Figure 1. Pieces of rock regularly fall to the base of cliffs to form talus slopes.

Rocks that fall to the base of a cliff make a **talus slope** (figure 1). Sometimes as one rock falls, it hits another rock, which hits another rock, and begins a landslide.

Landslides and **avalanches** are the most dramatic, sudden, and dangerous examples of earth materials moved by gravity. Landslides are sudden falls of rock, whereas avalanches are sudden falls of snow.

- Here is a video of a snow avalanche.

Landslides

Although many types of mass movements are included in the general term “landslide,” the more restrictive use of the term refers only to mass movements, where there is a distinct zone of weakness that separates the slide material from more stable underlying material. The two major types of slides are rotational slides and translational slides. Rotational slide: This is a slide in which the surface of rupture is curved concavely upward and the slide movement is roughly rotational about an axis that is parallel to the ground surface and transverse across the slide (figure 2a). Translational slide: In this type of slide, the landslide mass moves along a roughly planar surface with little rotation or backward tilting (figure 2b). A block slide is a translational slide in which the moving mass consists of a single unit or a few closely related units that move downslope as a relatively coherent mass (figure 2c).

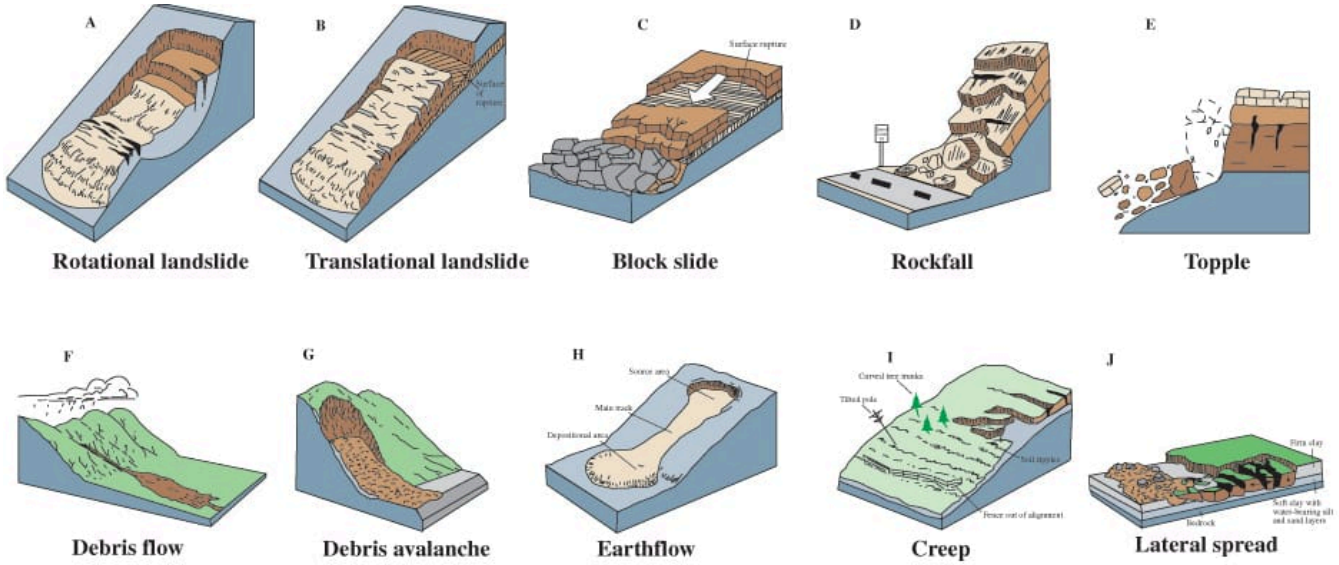


Figure 1. These schematics illustrate the major types of landslide movement.

When large amounts of rock suddenly break loose from a cliff or mountainside, they move quickly and with tremendous force (figure 2). Air trapped under the falling rocks acts as a cushion that keeps the rock from slowing down. Landslides and avalanches can move as fast as 200 to 300 km/hour.



Figure 2. (a) Landslides are called rock slides by geologists. (b) A snow avalanche moves quickly down slope, burying everything in its path.

Landslides are exceptionally destructive. Homes may be destroyed as hillsides collapse. Landslides can even bury entire villages. Landslides may create lakes when the rocky material dams a stream. If a landslide flows into a lake or bay, they can trigger a tsunami (figure 3).



Figure 3. The 1958 landslide into Lituya Bay, Alaska, created a 524m tsunami that knocked down trees at elevations higher than the Empire State Building (light gray).

Landslides often occur on steep slopes in dry or semi-arid climates. The California coastline, with its steep cliffs and years of drought punctuated by seasons of abundant rainfall, is prone to landslides. At-risk communities have developed landslide warning systems. Around San Francisco Bay, the National Weather Service and the U.S. Geological Survey use rain gauges to monitor soil moisture. If soil becomes saturated, the weather service issues a warning. Earthquakes, which may occur on California's abundant faults, can also trigger landslides.

- Here is a video of rapid downslope movement of material.

KQED: Landslide Detectives

Hillside properties in the San Francisco Bay Area and elsewhere may be prone to damage from landslides. Geologists are studying the warning signs and progress of local landslides to help reduce risks and give people adequate warnings of these looming threats. You can learn more here.

Contributors and Attributions

CC licensed content, Shared previously

- 10.5: Erosion and Deposition by Gravity. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/10.5/. **License**: *CC BY-NC: Attribution-NonCommercial*

Public domain content

- Landslide Types and Processes. **Provided by**: USGS. **Located at**: http://pubs.usgs.gov/fs/2004/3072/fs-2004-3072.html. **License**: *Public Domain: No Known Copyright*

This page titled 4.10.1: Slides is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.10.2: Falls



Figure 1. The white areas on green hillsides mark scars from numerous mudflows.

Falls are abrupt movements of masses of geologic materials, such as rocks and boulders, that become detached from steep slopes or cliffs. Separation occurs along discontinuities such as fractures, joints, and bedding planes, and movement occurs by free-fall, bouncing, and rolling. Falls are strongly influenced by gravity, mechanical weathering, and the presence of interstitial water.

Lahars and Mudflow

Added water creates natural hazards produced by gravity (figure 1). On hillsides with soils rich in clay, little rain, and not much vegetation to hold the soil in place, a time of high precipitation will create a **mudflow**. Mudflows follow river channels, washing out bridges, trees, and homes that are in their path.

- This video shows a debris flow.

A lahar is mudflow that flows down a composite volcano (figure 2). Ash mixes with snow and ice melted by the eruption to produce hot, fast-moving flows. The lahar caused by the eruption of Nevado del Ruiz in Columbia in 1985 killed more than 23,000 people.



Figure 2. A lahar is a mudflow that forms from volcanic ash and debris.

Contributors and Attributions

CC licensed content, Shared previously

- 10.5: Erosion and Deposition by Gravity. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/10.5/. **License**: *CC BY-NC: Attribution-NonCommercial*

Public domain content

- Landslide Types and Processes. **Provided by**: USGS. **Located at**: http://pubs.usgs.gov/fs/2004/3072/fs-2004-3072.html. **License**: *Public Domain: No Known Copyright*

This page titled 4.10.2: Falls is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.10.3: Slump and Creep

Less dramatic types of downslope movement move earth materials slowly down a hillside. **Slump** moves materials as a large block along a curved surface (figure 1). Slumps often happen when a slope is undercut, with no support for the overlying materials, or when too much weight is added to an unstable slope.



Figure 1. Slump material moves as a whole unit, leaving behind a crescent shaped scar.



Figure 2. Trees with curved trunks are often signs that the hillside is slowly creeping downhill.

Creep is the imperceptibly slow, steady, downward movement of slope-forming soil or rock. Movement is caused by shear stress sufficient to produce permanent deformation, but too small to produce shear failure. There are generally three types of creep:

1. Seasonal, where movement is within the depth of soil affected by seasonal changes in soil moisture and soil temperature
2. Continuous, where shear stress continuously exceeds the strength of the material
3. Progressive, where slopes are reaching the point of failure as other types of mass movements. Creep is indicated by curved tree trunks, bent fences or retaining walls, tilted poles or fences, and small soil ripples or ridges

Curves in tree trunks indicate creep because the base of the tree is moving downslope while the top is trying to grow straight up (figure 2). Tilted telephone or power company poles are also signs of creep.

Contributors and Attributions

CC licensed content, Shared previously

- 10.5: Erosion and Deposition by Gravity. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/10.5/. **License**: *CC BY-NC: Attribution-NonCommercial*

Public domain content

- Landslide Types and Processes. **Provided by**: USGS. **Located at**: http://pubs.usgs.gov/fs/2004/3072/fs-2004-3072.html. **License**: *Public Domain: No Known Copyright*

This page titled 4.10.3: Slump and Creep is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.10.4: Natural Mass Wasting

Contributing Factors

There are several factors that increase the chance that a landslide will occur. Some of these we can prevent and some we cannot.

Water

A little bit of water helps to hold grains of sand or soil together. For example, you can build a larger sand castle with slightly wet sand than with dry sand. However too much water causes the sand to flow quickly away. Rapid snow melt or rainfall adds extra water to the soil, which increases the weight of the slope and makes sediment grains lose contact with each other, allowing flow.

Rock Type

Layers of weak rock, such as clay, also allow more landslides. Wet clay is very slippery, which provides an easy surface for materials to slide over.

Undercutting

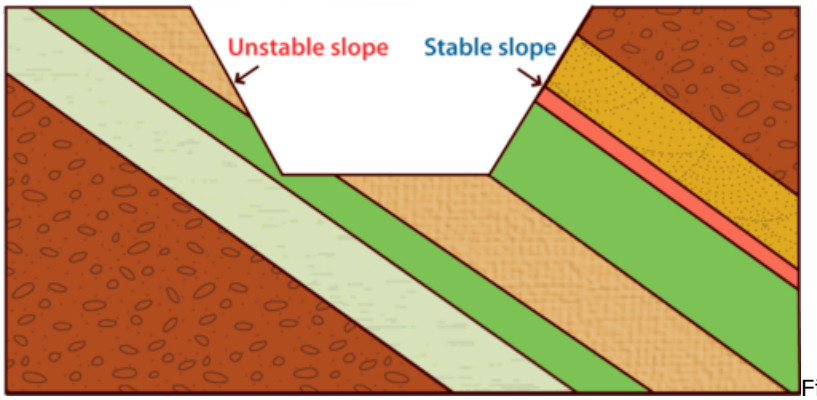


Figure 8. The slope of underlying materials must be considered when making road cuts.

If people dig into the base of a slope to create a road or a homesite, the slope may become unstable and move downhill. This is particularly dangerous when the underlying rock layers slope towards the area (figure 8).

- Ocean waves undercut cliffs and cause landslides on beaches as in this video.

When construction workers cut into slopes for homes or roads, they must stabilize the slope to help prevent a landslide (figure 9). Trees roots or even grasses can bind soil together. It is also a good idea to provide drainage so that the slope does not become saturated with water.

 Figure 9. A rock wall stabilizes a slope that has been cut away to make a road.

Ground shaking

An earthquake, volcanic eruption, or even just a truck going by can shake unstable ground loose and cause a slide. Skiers and hikers may disturb the snow they travel over and set off an avalanche.

A very good introduction to the topic, "Landslide 101," is a video seen on National Geographic Videos, Environment Video, Natural Disasters, Landslides, and more.

Landslides

Although there are multiple types of causes of landslides, the three that cause most of the damaging landslides around the world are these:

Landslides and Water

Slope saturation by water is a primary cause of landslides. This effect can occur in the form of intense rainfall, snowmelt, changes in ground-water levels, and water-level changes along coastlines, earth dams, and the banks of lakes, reservoirs, canals, and rivers.

Landsliding and flooding are closely allied because both are related to precipitation, runoff, and the saturation of ground by water. In addition, debris flows and mudflows usually occur in small, steep stream channels and often are mistaken for floods; in fact, these two events often occur simultaneously in the same area.

Landslides can cause flooding by forming landslide dams that block valleys and stream channels, allowing large amounts of water to back up. This causes backwater flooding and, if the dam fails, subsequent downstream flooding. Also, solid landslide debris can "bulk" or add volume and density to otherwise normal streamflow or cause channel blockages and diversions creating flood conditions or localized erosion. Landslides can also cause overtopping of reservoirs and/or reduced capacity of reservoirs to store water.

Landslides and Seismic Activity

Many mountainous areas that are vulnerable to landslides have also experienced at least moderate rates of earthquake occurrence in recorded times. The occurrence of earthquakes in steep landslide-prone areas greatly increases the likelihood that landslides will occur, due to ground shaking alone or shaking-caused dilation of soil materials, which allows rapid infiltration of water. The 1964 Great Alaska Earthquake caused widespread landsliding and other ground failure, which caused most of the monetary loss due to the earthquake. Other areas of the United States, such as California and the Puget Sound region in

Washington, have experienced slides, lateral spreading, and other types of ground failure due to moderate to large earthquakes. Widespread rockfalls also are caused by loosening of rocks as a result of ground shaking. Worldwide, landslides caused by earthquakes kill people and damage structures at higher rates than in the United States.

Landslides and Volcanic Activity

Landslides due to volcanic activity are some of the most devastating types. Volcanic lava may melt snow at a rapid rate, causing a deluge of rock, soil, ash, and water that accelerates rapidly on the steep slopes of volcanoes, devastating anything in its path. These volcanic debris flows (also known as lahars) reach great distances, once they leave the flanks of the volcano, and can damage structures in flat areas surrounding the volcanoes. The 1980 eruption of Mount St. Helens, in Washington triggered a massive landslide on the north flank of the volcano, the largest landslide in recorded times.

Geological and Morphological Causes

Geological Causes

1. Weak or sensitive materials
2. Weathered materials
3. Sheared, jointed, or fissured materials
4. Adversely oriented discontinuity (bedding, schistosity, fault, unconformity, contact, and so forth)
5. Contrast in permeability and/or stiffness of materials

Morphological Causes

1. Tectonic or volcanic uplift
2. Glacial rebound
3. Fluvial, wave, or glacial erosion of slope toe or lateral margins
4. Subterranean erosion (solution, piping)
5. Deposition loading slope or its crest
6. Vegetation removal (by fire, drought)
7. Thawing
8. Freeze-and-thaw weathering
9. Shrink-and-swell weathering

Contributors and Attributions

CC licensed content, Shared previously

- 10.5: Erosion and Deposition by Gravity. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/10.5/. **License**: *CC BY: Attribution*

Public domain content

- Landslide Types and Processes. **Provided by**: USGS. **Located at**: http://pubs.usgs.gov/fs/2004/3072/fs-2004-3072.html. **License**: *Public Domain: No Known Copyright*

This page titled 4.10.4: Natural Mass Wasting is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.10.5: Human Impact on Mass Wasting

Causing Mass Wasting

Humans can contribute to mass wasting in a few different ways:

1. Excavation of slope or its toe
2. Loading of slope or its crest
3. Drawdown (of reservoirs)
4. Deforestation
5. Irrigation
6. Mining
7. Artificial vibration
8. Water leakage from utilities

Prevention and Awareness

As already noted, we cannot prevent mass wasting in the long term as it is a natural and ongoing process; however, in many situations there are actions that we can take to reduce or mitigate its damaging effects on people and infrastructure. Where we can neither delay nor mitigate mass wasting, we should consider moving out of the way.



Figure 1. Site of the 2008 rock slide at Porteau Cove. Notice the prominent fracture set parallel to the surface of the slope. The slope has been stabilized with rock bolts (top) and holes have been drilled into the rock to improve drainage (one is visible in the lower right). Risk to passing vehicles from rock fall has been reduced by hanging mesh curtains (background). [SE photo 2012]

It is comforting to think that we can prevent some effects of mass wasting by mechanical means, such as the rock bolts in the road cut at Porteau Cove (Figure 1), or the drill holes used to drain water out of a slope, as was done at the Downie Slide, or the building of physical barriers, such as retaining walls. What we have to remember is that the works of humans are limited compared to the works of nature. The rock bolts in the road cut at Porteau Cove will slowly start to corrode after a few years, and within a few decades, many of them will begin to lose their strength. Unless they are replaced, they will no longer support that slope. Likewise, drainage holes at the Downie Slide will eventually become plugged with sediment and chemical precipitates, and unless they are periodically unplugged, their effectiveness will decrease. Eventually, unless new holes are drilled, the drainage will be so compromised that the slide will start to move again. This is why careful slope monitoring by geological and geotechnical engineers is important at these sites. The point here is that our efforts to “prevent” mass wasting are only as good as our resolve to maintain those preventive measures.

Delaying mass wasting is a worthy endeavor, of course, because during the time that the measures are still effective they can save lives and reduce damage to property and infrastructure. The other side of the coin is that we must be careful to avoid activities that could make mass wasting more likely. One of the most common anthropogenic causes of mass wasting is road construction, and this applies both to remote gravel roads built for forestry and mining and large urban and regional highways. Road construction is a potential problem for two reasons. First, creating a flat road surface on a slope inevitably involves creating a cut bank that is steeper than the original slope. This might also involve creating a filled bank that is both steeper and weaker than the original slope (Figure 2). Second, roadways typically cut across natural drainage features, and unless great care is taken to reroute the runoff water and prevent it from forming concentrated flows, over-saturating fill of materials can result.

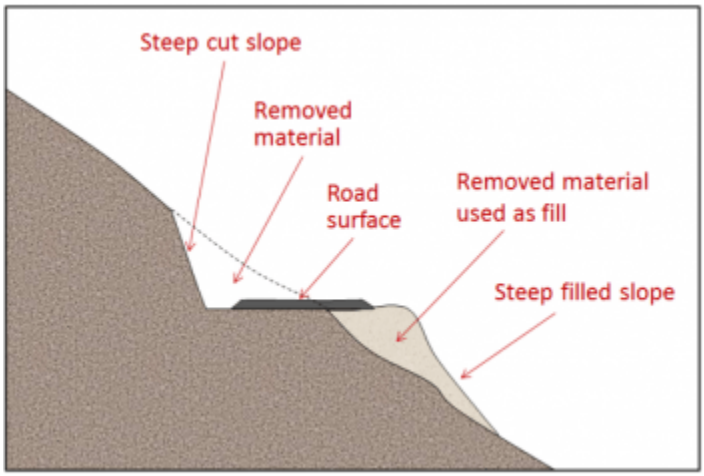
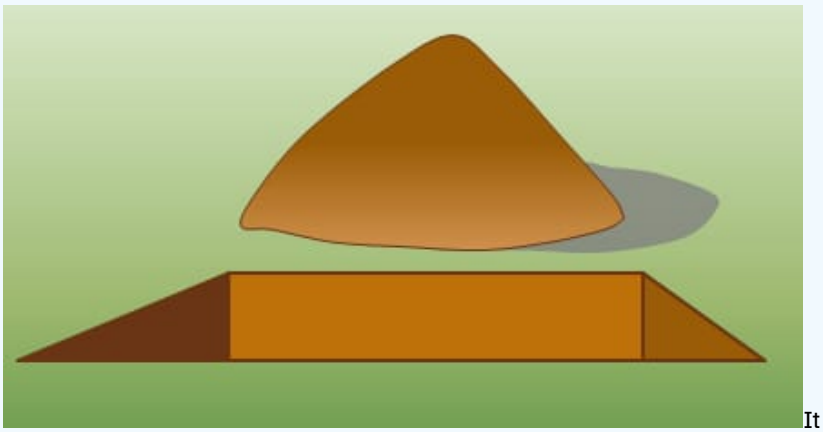


Figure 2. An example of a road constructed by cutting into a steep slope and the use of the cut material as fill. [SE]

Apart from water issues, engineers building roads and other infrastructure on bedrock slopes have to be acutely aware of the geology, and especially of any weaknesses or discontinuities in the rock related to bedding, fracturing, or foliation. If possible, situations like that at Porteau Cove (Figure 1) should be avoided — by building somewhere else — rather than trying to stitch the slope back together with rock bolts.

It is widely believed that construction of buildings on the tops of steep slopes can contribute to the instability of the slope. This is probably true, but not because of the weight of the building. As you'll see by completing the exercise below, a typical house isn't usually heavier than the fill that was removed from the hole in the ground made to build it. A more likely contributor to instability of the slope around a building is the effect that it and the changes made to the surrounding area have on drainage.

Exercise: How Much Does a House Weigh and Can It Contribute to a Slope Failure?


It is commonly believed that building a house (or some other building) at the top of a slope will add a lot of extra weight to the slope, which could contribute to slope failure. But what does a house actually weigh? A typical 150 m^2 (approximately 1,600 ft^2) wood-frame house with a basement and a concrete foundation weighs about 145 t (metric tonnes). But most houses are built on foundations that are excavated into

the ground. This involves digging a hole and taking some material away, so we need to subtract what that excavated material weighs. Assuming our 150 m2 house required an excavation that was 15 m by 11 m by 1 m deep, that's 165 m^3 of "dirt," which typically has a density of about 1.6 t per m^3.

Calculate the weight of the soil that was removed and compare that with the weight of the house and its foundation.

If you're thinking that building a bigger building is going to add more weight, consider that bigger buildings need bigger and deeper excavations, and in many cases the excavations will be into solid rock, which is much heavier than surficial materials.

You may wish to consider how a building might change the drainage on a slope. There are a number ways. Water can be collected by roofs, go into downspouts, and form concentrated flows that are directed onto or into the slope. Likewise drainage from nearby access roads, lawn irrigation, leaking pools, and septic systems can all alter the surface and groundwater flow in a slope.

Monitoring Mass Wasting



Figure 3. Part of a motion-monitoring device at the Checkerboard Slide near Revelstoke, B.C. The lower end of the cable is attached to a block of rock that is unstable. Any incremental motion of that block will move the cable, which will be detectable on this device. [SE]

In some areas, it is necessary to establish warning systems so that we know if conditions have changed at a known slide area, or if a rapid failure, such as a debris flow, is actually on its way downslope. The Downie Slide above the Revelstoke Resevoir is monitored 24/7 with a range of devices, such as inclinometers (slope-change detectors), bore-hole motion sensors, and GPS survey instruments. A simple mechanical device for monitoring the nearby Checkerboard Slide (which is also above the Revelstoke Reservoir) is shown in Figure 3. Both of these are very slow-moving rock slides, but it's very important to be able to detect changes in their rates of motion because at both of these locations a rapid failure would result in large bodies of rock plunging into the reservoir and sending a wall of water over the Revelstoke Dam, potentially destroying the nearby town of Revelstoke.

Mt. Rainier, a glacier-covered volcano in Washington State, has the potential to produce massive mudflows or debris flows (lahars) with or without a volcanic eruption. Over 100,000 people in the Tacoma, Puyallup, and Sumner areas are in harm's way because they currently reside on deposits from past lahars (Figure 4). In 1998, a network of acoustic monitors was established around Mt. Rainier. The monitors are embedded in the ground adjacent to expected lahar paths. They are intended to provide warnings to emergency officials, and when a lahar is detected, the residents of the area will have anywhere from 40 minutes to three hours to get to safe ground.



Figure 4. Mt. Rainier, Washington, from Tacoma

Mitigating the Impacts of Mass Wasting

In situations where we can't predict, prevent, or delay mass-wasting hazards, some effective measures can be taken to minimize the associated risk. For example, many highways in B.C. and western Alberta have avalanche shelters like that shown in Figure 5. In some parts of the world, similar features have been built to protect infrastructure from other types of mass wasting.



Figure 5. A snow avalanche shelter on the Coquihalla Highway. The expected path of the avalanche is the steep untreed slope above. [SE]

Debris flows are inevitable, unpreventable, and unpredictable in many parts of B.C., but nowhere more so than along the Sea-to-Sky Highway between Horseshoe Bay and Squamish. The results have been deadly and expensive many times in the past. It would be very expensive to develop a new route in this region, so provincial authorities have taken steps to protect residents and traffic on the highway and the railway. Debris-flow defensive structures have been constructed in several drainage basins, as shown in Figure 6. One strategy is to allow the debris to flow quickly through to the ocean along a smooth channel. Another is to capture the debris within a constructed basin that allows the excess water to continue through, but catches the debris materials.



Figure 6. Two strategies for mitigating debris flows on the Sea-to-Sky Highway. Left: A concrete –lined channel on Alberta Creek allows debris to flow quickly through to the ocean. Right: A debris-flow catchment basin on Charles Creek. In 2010, a debris flow filled the basin to the level of the dotted white line. [SE]

Landslide Mitigation—How to Reduce the Effects of Landslides

Vulnerability to landslide hazards is a function of location, type of human activity, use, and frequency of landslide events. The effects of landslides on people and structures can be lessened by total avoidance of landslide hazard areas or by restricting, prohibiting, or imposing conditions on hazard-zone activity. Local governments can reduce landslide effects through land-use policies and regulations. Individuals can reduce their exposure to hazards by educating themselves on the past hazard history of a site and by making inquiries to planning and engineering departments of local governments. They can also obtain the professional services of an engineering geologist, a geotechnical engineer, or a civil engineer, who can properly evaluate the hazard potential of a site, built or unbuilt.

The hazard from landslides can be reduced by avoiding construction on steep slopes and existing landslides, or by stabilizing the slopes. Stability increases when ground water is prevented from rising in the landslide mass by (1) covering the landslide with an impermeable membrane, (2) directing surface water away from the landslide, (3) draining ground water away from the landslide, and (4) minimizing surface irrigation. Slope stability is also increased when a retaining structure and/or the weight of a soil/rock berm are placed at the toe of the landslide or when mass is removed from the top of the slope.

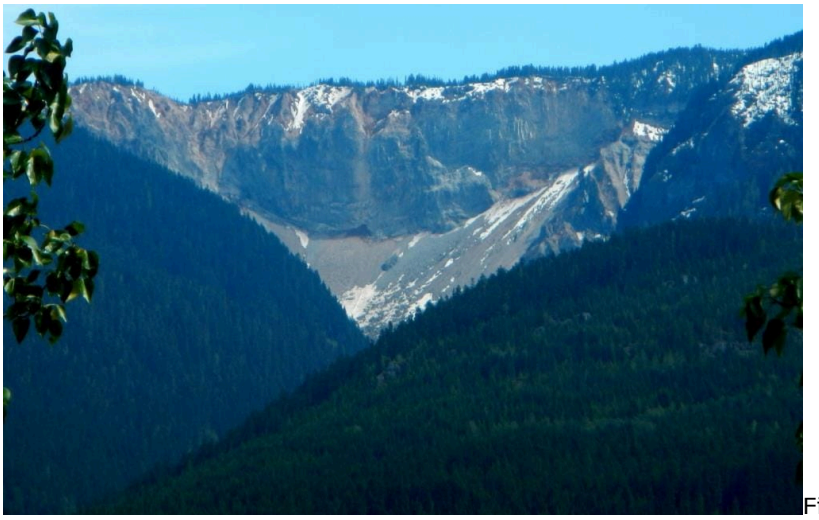


Figure 7. The Barrier, south of Whistler, B.C., was the site of a huge rock avalanche in 1855, which extended from the cliff visible here 4 km down the valley and across the current location of the Sea-to-Sky Highway and the Cheakamus River. [SE]

Finally, in situations where we can't do anything to delay, predict, contain, or mitigate slope failures, we simply have to have the sense to stay away. There is a famous example of this in B.C. at a site known as Garibaldi, 25 km south of Whistler. In the early 1980s the village of Garibaldi had a population of about 100, with construction underway on some new homes, and plans for many

more.

In the months that followed the deadly 1980 eruption of Mt. St. Helens in Washington State, the B.C. Ministry of Transportation commissioned a geological study that revealed that a steep cliff known as The Barrier (Figure 7) had collapsed in 1855, leading to a large rock avalanche, and that it was likely to collapse again unpredictably, putting the village of Garibaldi at extreme risk. In an ensuing court case, it was ruled that the Garibaldi site was not a safe place for people to live. Those who already had homes there were compensated, and everyone else was ordered to leave.

Contributors and Attributions

CC licensed content, Original

- Revision, adaption, and original content. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Preventing, Delaying, Monitoring, and Mitigating Mass Wasting. **Authored by**: Steven Earle. **Provided by**: BC Campus. **Located at**: https://opentextbc.ca/geology/chapter/15-3-preventing-delaying-monitoring-and-mitigating-mass-wasting/. **Project**: Physical Geology. **License**: *CC BY: Attribution*

Public domain content

- Landslide Types and Processes. **Provided by**: USGS. **Located at**: http://pubs.usgs.gov/fs/2004/3072/fs-2004-3072.html. **License**: *Public Domain: No Known Copyright*
- Mount Rainier over Tacoma. **Authored by**: Lyn Topinka. **Provided by**: USGS. **Located at**: http://vulcan.wr.usgs.gov/Imgs/Jpg/Rainier/Images/Rainier84_mount_rainier_and_tacoma_08-20-84.jpg. **License**: *Public Domain: No Known Copyright*

This page titled 4.10.5: Human Impact on Mass Wasting is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.11: HYDROLOGY

Learning Objectives

Understand the roles, processes and effects of streams and groundwater.

- Describe and model processes involved throughout Earth's hydrologic cycle.
- Identify and describe channel types and sedimentary loads formed and carried by rivers and streams.
- Describe critical components of groundwater.
- Identify various geologic features associated with groundwater.
- Recognize features associated with oceans and shorelines.

4.11.1: Phases of the Hydrologic Cycle

4.11.2: Types of Streams and Rivers

4.11.3: Drainage Basins

4.11.4: Geologic Processes and Flowing Water

4.11.5: Groundwater

4.11.6: Porosity and Permeability

4.11.7: Aquifers and Well-Potential

4.11.8: Groundwater Withdrawal

4.11.9: Quality of Groundwater

4.11.10: Karst Topography

4.11.11: Geysers

4.11.12: Springs

4.11.13: Geologic Features

4.11.14: Oceans and Shorelines

4.11.15: Ocean Movement

4.11.16: Shorelines

4.11.17: Coastal Hazards

4.11.18: Human Modifications of Coastal Processes

This page titled Chapter 4.11: Hydrology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.1: Phases of the Hydrologic Cycle

Because of the unique properties of water, water molecules can cycle through almost anywhere on Earth. The water molecule found in your glass of water today could have erupted from a volcano early in Earth history. In the intervening billions of years, the molecule probably spent time in a glacier or far below the ground. The molecule surely was high up in the atmosphere and maybe deep in the belly of a dinosaur. Where will that water molecule go next?

Three States of Water

Water is the only substance on Earth that is present in all three states of matter—as a solid, liquid or gas. (And Earth is the only planet where water is present in all three states.) Because of the ranges in temperature in specific locations around the planet, all three phases may be present in a single location or in a region. The three phases are solid (ice or snow), liquid (water), and gas (**water vapor**). See ice, water, and clouds (figure 2).



Figure 2. Can you find all three phases of water in this image? (a) Ice floating in the sea. (b) Liquid water. (c) Water vapor is invisible, but clouds that form when water vapor condenses are not.

The Water Cycle

Because Earth's water is present in all three states, it can get into a variety of environments around the planet. The movement of water around Earth's surface is the **hydrologic (water) cycle** (figure 3).

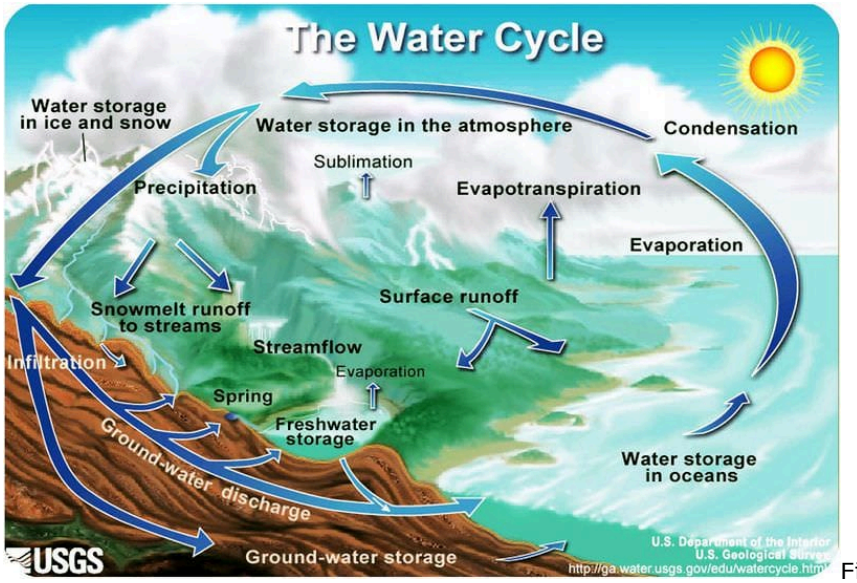


Figure 3. Because it is a cycle, the water cycle has no beginning and no end.

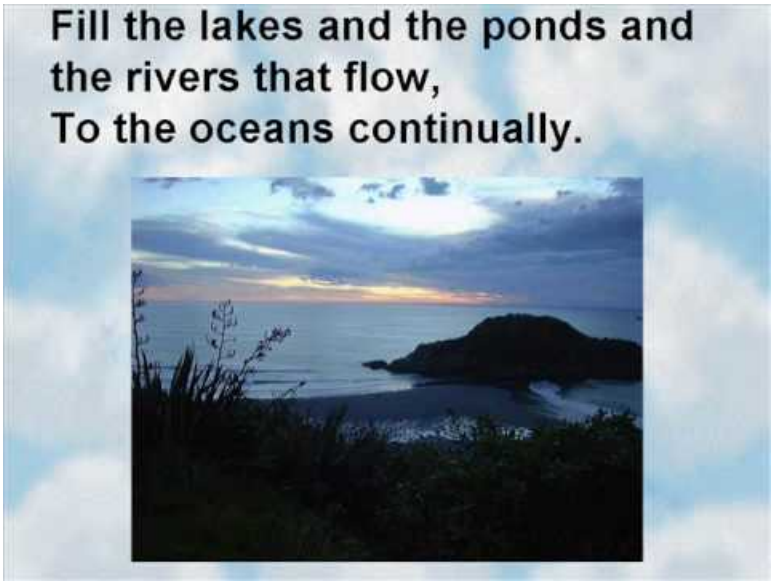
Most of Earth's water is stored in the oceans where it can remain for hundreds or thousands of years. The oceans are discussed

in detail in the chapter Earth's Oceans.

Water changes from a liquid to a gas by **evaporation** to become water vapor. The Sun's energy can evaporate water from the ocean surface or from lakes, streams, or puddles on land. Only the water molecules evaporate; the salts remain in the ocean or a fresh water reservoir.

The water vapor remains in the atmosphere until it undergoes **condensation** to become tiny droplets of liquid. The droplets gather in clouds, which are blown about the globe by wind. As the water droplets in the clouds collide and grow, they fall from the sky as precipitation. **Precipitation** can be rain, sleet, hail, or snow. Sometimes precipitation falls back into the ocean and sometimes it falls onto the land surface.

For a little fun, watch this video. This water cycle song focuses on the role of the sun in moving H_2O from one reservoir to another. The movement of all sorts of matter between reservoirs depends on Earth's internal or external sources of energy:



A YouTube element has been excluded from this version of the text. You can view it online here: http://pb.libretexts.org/geo/?p=396

This animation shows the annual cycle of monthly mean precipitation around the world.

When water falls from the sky as rain it may enter streams and rivers that flow downward to oceans and lakes. Water that falls as snow may sit on a mountain for several months. Snow may become part of the ice in a glacier, where it may remain for hundreds or thousands of years. Snow and ice may go directly back into the air by **sublimation**, the process in which a solid changes directly into a gas without first becoming a liquid. Although you probably have not seen water vapor sublimating from a glacier, you may have seen dry ice sublimate in air.

Snow and ice slowly melt over time to become liquid water, which provides a steady flow of fresh water to streams, rivers, and lakes below. A water droplet falling as rain could also become part of a stream or a lake. At the surface, the water may eventually evaporate and reenter the atmosphere.

A significant amount of water infiltrates into the ground. Soil moisture is an important reservoir for water (figure 4). Water trapped in soil is important for plants to grow.

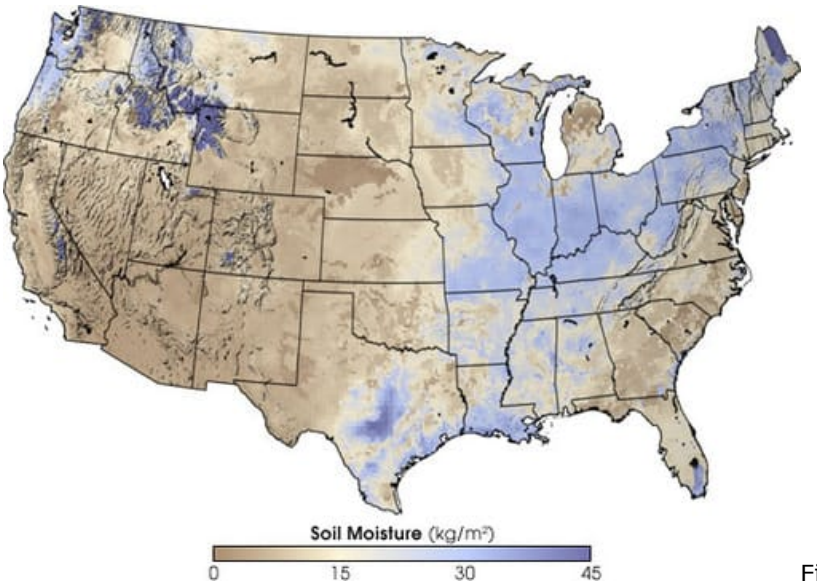


Figure 4. The moisture content of soil in the United States varies greatly.

Water may seep through dirt and rock below the soil through pores infiltrating the ground to go into Earth's groundwater system. Groundwater enters aquifers that may store fresh water for centuries. Alternatively, the water may come to the surface through springs or find its way back to the oceans.

Plants and animals depend on water to live and they also play a role in the water cycle. Plants take up water from the soil and release large amounts of water vapor into the air through their leaves (figure 5), a process known as **transpiration**.

An online guide to the hydrologic cycle from the University of Illinois is found here.

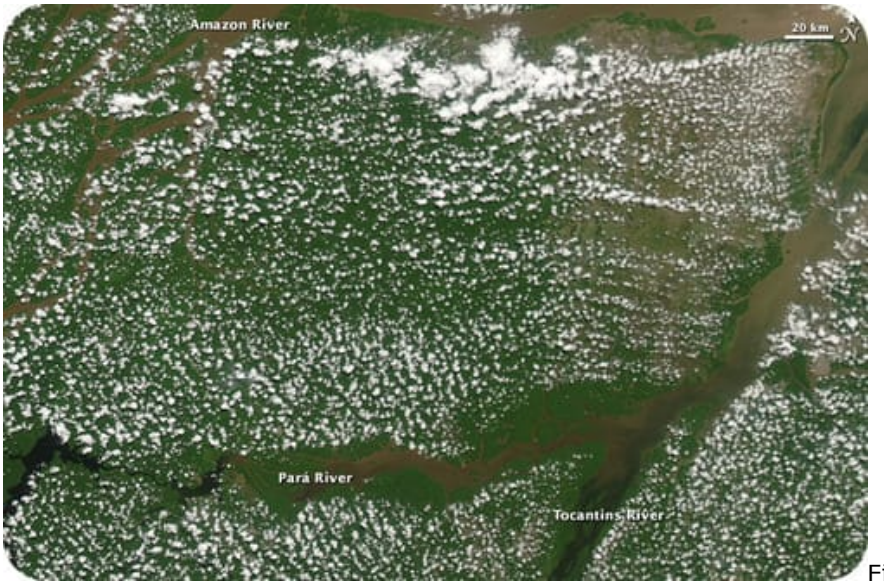


Figure 5. Clouds form above the Amazon Rainforest even in the dry season because of moisture from plant transpiration.

People also depend on water as a natural resource. Not content to get water directly from streams or ponds, humans create canals, aqueducts, dams, and wells to collect water and direct it to where they want it (figure 6).



Figure 6. The Pont du Gard aqueduct in France was constructed during the Roman Empire.

Table 1. Water Use in the United States and Globally

Use	United States	Global
Agriculture	34%	70%
Domestic (drinking, bathing)	12%	10%
Industry	5%	20%
Power plant cooling	49%	small

It is important to note that water molecules cycle around. If climate cools and glaciers and ice caps grow, there is less water for the oceans and sea level will fall. The reverse can also happen.

KQED: Tracking Raindrops

How the water cycle works and how rising global temperatures will affect the water cycle, especially in California, are the topics of this Quest video. Learn more here.

Contributors and Attributions

CC licensed content, Shared previously

- The Water Cycle. **Provided by**: CK-12. **Located at**: http://www.ck12.org/student/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.11.1: Phases of the Hydrologic Cycle is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.2: Types of Streams and Rivers

Introduction

Streams have a major role in geology. Streams sculpt and shape the earth's surface by eroding, transporting, and depositing sediment. By eroding sediment from uplifted areas and creating landforms made of deposited sediment in lower areas, streams shape the earth's surface more than glaciers do, more than waves on a beach do, and far more than wind does.

What Are Streams?

A stream is flow of water, driven by gravity, in a natural channel, on land. A small brook in a meadow and the Amazon River are both streams. It is interesting to watch water on a recently bulldozed construction site with a slope. At first the water saturates the ground and begins to flow downhill across the surface of the slope in a thin sheet. Soon, the water excavates small channels, known as rills, in the dirt. Rills coalesce to form larger channels. A network of streams, including tributaries, has formed. If not prevented, the channels may continue to deepen and erode soil from the construction site.

Over longer intervals of time the same processes we've imagined on the construction site have built systems of streams and stream valleys on the surface of the earth. Most valleys on earth are the product of streams. Streams erode dirt and rocks, transport the sediment, and redeposit it in new locations, shaping the earth's surface into a system of stream valleys.

Streams flow downhill due to the force of gravity. The higher the hill, the more gravitational energy there is to drive the stream. Where the slopes are steepest and the hills the highest, the streams will be the most energetic and the rate of erosion will be fastest.

Drainage Area

The drainage area of a stream encompasses all the land from which surface runoff flows into that stream. A stream drainage area is also called a watershed. Boundaries between stream drainage areas are called drainage divides. What stream drainage do you live in?

Stream Order

It is common for one stream to flow into another. The smaller of the two streams is a tributary of the larger stream. A stream with no tributaries is a first order stream. A stream with only first-order tributaries is a second order stream. A stream that has any second-order tributaries and none higher is a third-order stream, and so on. The Mississippi River is a tenth order stream, one of the highest order streams on earth. As more and more tributaries join together a larger stream network is formed and the master stream, the highest order stream in the system has a discharge that is the sum of all the tributary discharges. When flooding occurs, higher order streams take longer to build up to flood stage than lower order streams and longer for the flood to subside.

Drainage Patterns

A stream system that includes multiple tributaries exhibits a distinct drainage pattern as seen on a map. The drainage pattern depends on the rock types and geologic structures underlying the stream system. Some types of rock are harder and more resistant to erosion than others. If the geology underlying a stream system is fairly uniform—rocks equally resistant to erosion in all directions—a dendritic drainage pattern will develop, as shown in figure 1. A dendritic drainage pattern is the most common type.

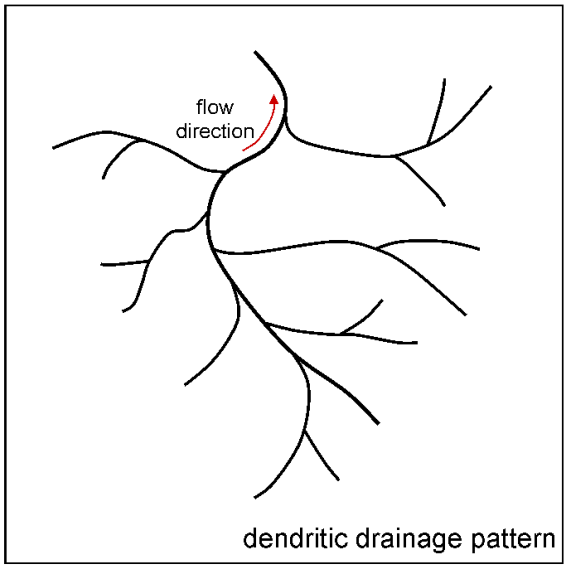


Figure 1

If a region is underlain by layered formations of rock that have been folded, and the layers have different degrees of resistance to erosion, the stream valleys will tend to follow the layers of less resistant rock, and the layers of harder rock will become ridges. This results in a trellis drainage pattern, as show in figure 2.

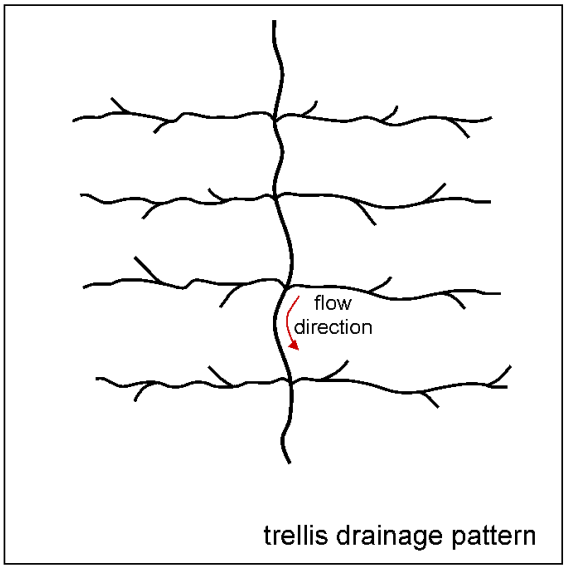


Figure 2.

In some places the geology consists of a single type of rock that is resistant to erosion but the rock contains sets of parallel joints where it erodes more easily. The sets of joints typically intersect each other at high angles. As stream valleys develop in the joint system a rectangular drainage pattern develops, as shown in figure 3. The stream valleys will bend sharply where they switch from following one joint set to another.

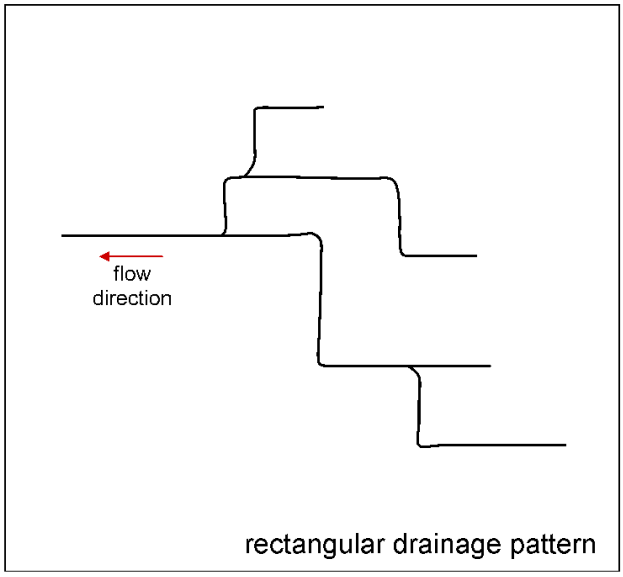


Figure 3.

Streams will radiate in all directions from the center of a broad, high-elevation area, such as a composite cone. This is known as a radial drainage pattern.

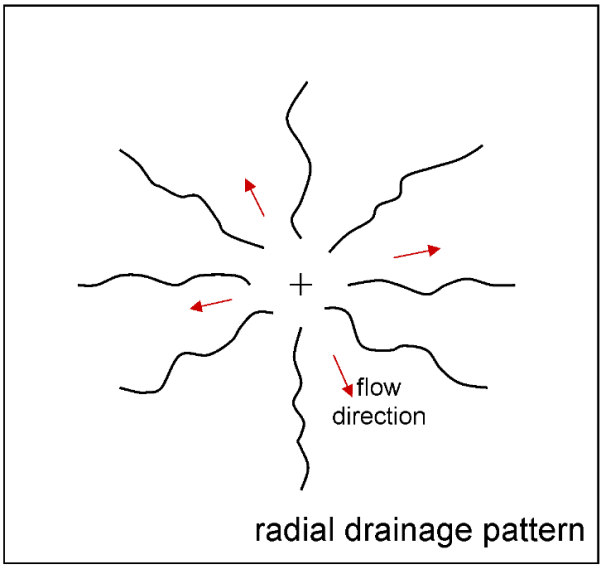


Figure 4.

Graded Profile

Because streams erode more and remove more sediment where the stream gradient is higher, and deposit more sediment where the stream gradient is lower, a stream will develop a graded profile as shown. The graded profile shows how the elevation of the stream changes along the length of the stream, from its beginning at the highest elevation to its base level where it ends at the lowest elevation it reaches.

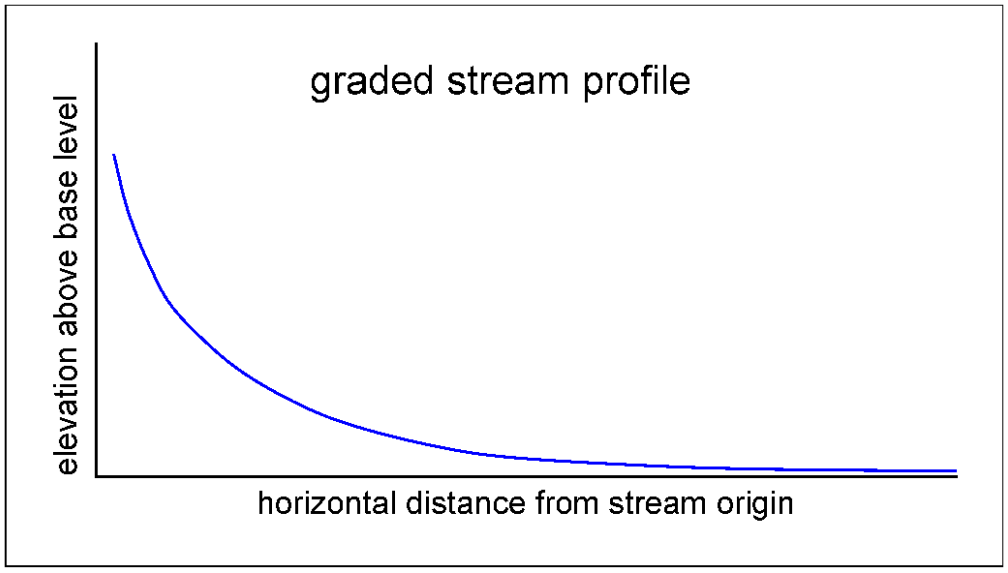


Figure 5.

A graded profile starts with a steep slope at the beginning of the stream and tapers to a gentle slope at the base level of the stream. Imagine a hiker following a stream up to its beginning and tracking her progress on a topographic map. She starts out where the stream valley is low and wide and the gradient is gentle so the hiking is easy. As the day goes on, the going gets steeper as she approaches the stream source. Checking the topographic map, she sees that the contour lines are much closer together than they were earlier in the day and farther down the valley. As she approaches the small lake on the side of the mountain where the stream begins, the slope is so steep that nearly loses her footing. She has experienced for herself a typical stream profile that steepens from near the base-level to the origin of the stream.

Lakes and waterfalls are temporary features in a stream drainage. If a lake forms the stream feeding it will slow down and deposit sediment until the lake has been filled with sediment. Where a waterfall forms the energy of the stream going over the waterfall is high will erode the base of the waterfall, causing the waterfall to retreat upstream until a graded profile is established.

Stream Discharge

The amount of water flowing through a stream and speed at which it is moving is expressed as the discharge of the stream. Discharge is measured by multiplying the cross-sectional area of a stream by the average speed of water through that cross-section. Multiplying the area of the cross-section by the average speed of the water results in units of volume/time. For example, if a stream has a cross-sectional area of 150 square feet (ft^2) and is moving at an average speed of 10 feet per second (ft/s), multiplying the area and the speed results in a discharge of 1,500 cubic feet per second (ft^3/s).

Sediment Load

The water flowing in streams erodes, transports, and deposits sediment. Most rocks and minerals are much denser than water. Sufficient energy in the stream is required to dislodge rocks from the earth and move them. The faster a stream flows, the more energy it has and the larger pieces of sediment it can transport. The *competence* of a stream refers to the maximum size of the pieces of sediment it can move. Faster-moving water has greater competence and can move larger pieces of sediment.

The *capacity* of a stream is the total amount of sediment it can move. Capacity depends on how fast the stream is moving and its total discharge. As a stream slows down, its competence and capacity are reduced. The stream starts to deposit sediment, beginning with the largest pieces. Several factors cause a stream to slow down, including the stream channel widening, the stream overflowing its banks and spreading into a floodplain, the stream gradient (downhill slope) diminishing, and the stream emptying into a larger, slower moving body of water.

Streams transport clastic sediment in two ways depending on the size of the particles. The coarser sediment is called bed load and consists of particles too large to be suspended in the moving current of water for an extended length of time. The bed load particles, the largest particles of sediment transported by a stream, spend most of their time on the bottom of the stream channel,

rolling, sliding, or bouncing downstream in fits and starts. The finer sediment is called suspended load and consists of particles small enough to be suspended in the moving current of water. Suspended load moves at about the same rate as the flowing water. In a muddy river, the mud is suspended load

Floodplains

Streams build floodplains through a combination of erosion and deposition at lower gradient stretches of stream valleys. Although a floodplain has a general downhill slope consistent with the overall stream gradient, a floodplain is relatively flat.

Floodplains are filled with sediments spread by the stream. These sediments are known as alluvium. Because alluvium is loose material that is easy for the stream to erode and redeposit, the location of a stream channel in a floodplain changes frequently.

Meanders

A stream running down a slope, even the gentle slope of a floodplain, will seldom follow a straight path for very long. Depending on the distribution of sediments and turbulence of the stream, one side of the channel may erode more easily than the other. The stream will migrate toward the area undergoing erosion, developing a curve in that direction. Once the stream channel has begun to curve, the energy of the water is concentrated on the outside of the curve.

The diagram shows a stretch of stream channel with a significant bend, also known as meander. The blue line shows how erosive energy is concentrated along the outside of each bend in the stream. As erosion occurs on the outside bank of a meander, deposition occurs on the inside bank where the water slows and drops sediment.

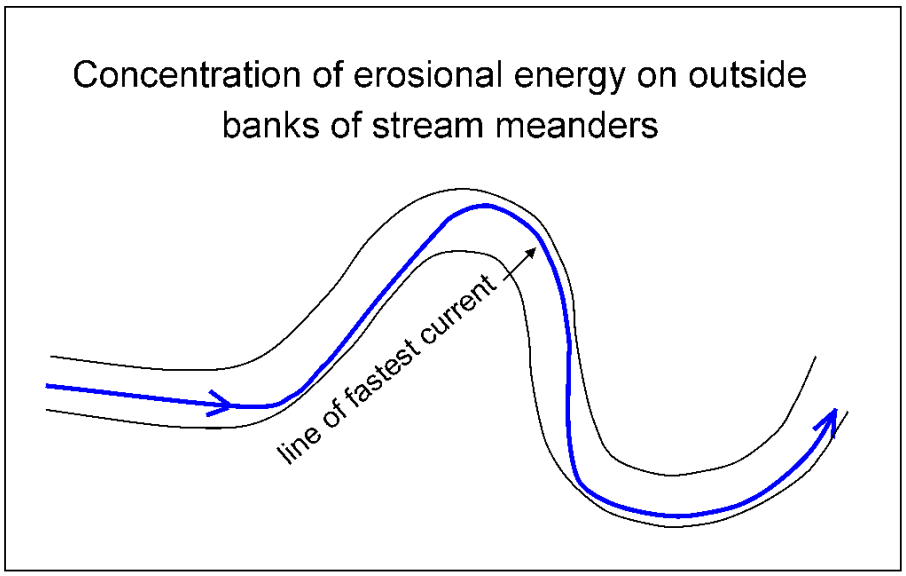


Figure 6.

The diagram below shows two well-developed meanders that have formed in a stream. Along each meander, the outer stream bank that is being cut into by erosion is called a cut bank. The inner bank, which has grown by accretion of deposited sediment, is called a point bar.

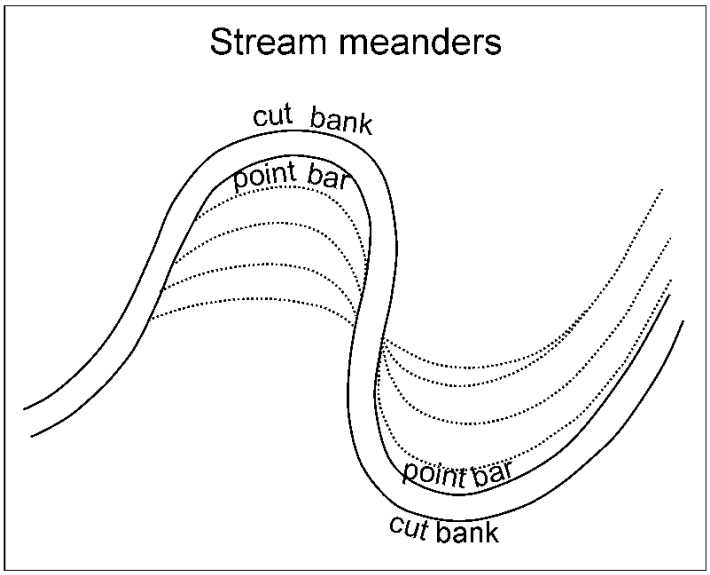


Figure 7.

In a stream, meanders enlarge and migrate downstream because the stream continually erodes its cut banks and grows its point bars. The diagram below shows the enlargement and downstream migration of a meander in a stream channel. As the meander is enlarged, its neck gets narrower. Eventually, the stream may cut through the neck of the meander, either as a result of gradual erosion and channel migration, or abruptly during high water and flooding. Once the stream has cut through the neck of the meander, the openings get filled with sediment dropped by water that slows down as it enters from the main stream. The sediment deposits will separate the cut off meander from the river channel and turn it into an oxbow lake. As the years go by, the oxbow lake will eventually be completely filled in with sediment because it is a low spot on the floodplain where any water that enters, such as during flooding, will come to a standstill and deposit its sediment load.

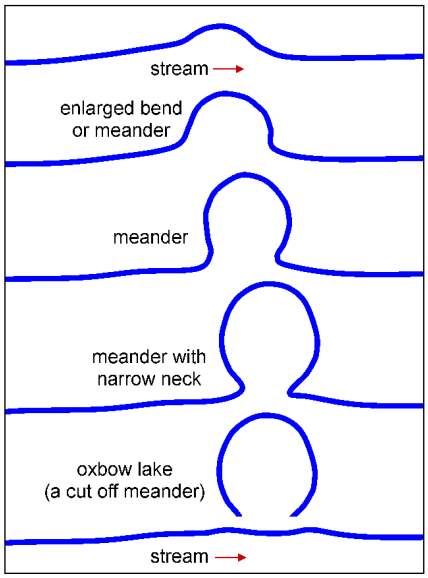


Figure 8.

Entrenched Meanders

Typical meandering stream channels flow through broad flood plains full of alluvial sediment. However, in some situations meanders may cut directly into bedrock. A meander that has cut into bedrock is known as an incised or entrenched meander. In contrast to meanders in alluvium that erode and migrate rapidly or get cut off at the neck abruptly, entrenched meanders are relatively fixed. This is because entrenched meanders are walled in by bedrock on both sides and have little floodplain to easily erode and redeposit.

Entrenched meanders form as a result of tectonic uplift of the stream drainage area. The uplift increases the gravity-driven energy

of the stream causing it to incise rapidly down through the flood plain alluvium into the bedrock beneath. Entrenched meanders are striking landscape features because they are unusual and they provide strong evidence of tectonic activity in a region. Classic examples of entrenched meanders include the Goosenecks of the San Juan River, which are incised in the Colorado Plateau east of the Grand Canyon, and a stretch of the Yakima River with entrenched meanders, which are incised in a recently uplifted ridge of basalt in the Columbia Plateau of eastern Washington state.

Braided Streams

Rather than a single channel, some streams have multiple channels that weave in and out of each other forming what is known as a braided stream. Braided streams are associated with excessive amounts of sediment entering a stream system. Valleys draining alpine glaciers are common settings for braided streams. The glaciers deposit more sediment into the meltwater stream system than a stream of that discharge has capacity to transport in a single channel system. Braided stream systems are indicators that there is an additional source of sediment in the system besides the stream itself. Sources of excess sediment that lead to braided streams include glaciers, eruptions of pyroclastic material by volcanoes and landslides.

Flooding and Flood Frequency

Streams flood. Flooding is a normal part of stream behavior. The rate at which streams erode, transport, and deposit sediments greatly increases during flooding. A flood occurs when the water depth in a stream exceeds the depth of the stream channel and spreads beyond the stream channel onto the surrounding land. When a stream completely fills its channel it is said to be at bankfull stage. When a stream surpasses bankfull stage it is said to be in flood. When the flood spreads widely enough to cause property damage, a stream is said to be at flood stage.

Different streams have different flood behaviors. Some rivers tend to have an annual flood associated with a rainy season or snowmelt season in the higher part of the drainage area. Some streams only flood sporadically, often with years between floods. In the United States all the larger streams have their flow measured at gaging stations that are installed and monitored by the US Geological Survey. Records of stream flow must be kept for at least 30 years to be statistically valid. The likelihood of a flood of a specified magnitude occurring on a stream in a given year can be calculated from the stream flow data. The magnitude of a flood that has a 1 out of 100 chance of happening in a given year—also known as the 100-year flood—is calculated for a stream from it record of stream flow. The 100-year flood magnitude, and maps that show what would be covered by a 100-year flood, are key elements in granting permits for land use and building construction in floodplains.

Low order streams may flood rapidly during or immediately after heavy rainfall. Such rapid floods are called flash floods. Flash floods are often unpredictable. If a river drains through a canyon from a high elevation area, people may be walking along a stream in the canyon in the sunshine and not realize that a thunderstorm is occurring at the stream's source. They may be surprised by a flash flood sweeping through the canyon. A slow building flood on a higher order stream can usually be predicted because major streams are monitored and it takes longer for water from heavy rainfall or rapidly melting snow to move through the higher order stream system. Floods on the highest order streams, such as the Mississippi River, may take several weeks to crest and subside.

Changes to the land in the drainage area of a stream, such as the addition of buildings and roads, can change how a stream floods. Buildings and pavement that cover the ground prevent infiltration and cause increased surface run off. Increased stream runoff means smaller amount of rain will cause the stream to reach flood stage than before the alterations to the land. The stream will flood more frequently. For the same amount and rate of rainfall that caused flooding prior to building, the stream will reach flood stage quicker and the flood will be deeper. The diagram below shows the flood response of a stream to a given amount of rainfall before and after the stream drainage area was urbanized by removal of forest and addition of buildings and roads. After urbanization, the stream reaches its peak flood level quicker and rises to a higher level than before urbanization. Unless countermeasures are engineered and installed, an urbanized area will flood more frequently and severely than the same area when it was forested.

Figure 9.

Alluvial Fans



Figure 10.

At the location where a stream reaches its base level, it slows down and deposits nearly all of the sediment it is carrying. A stream that comes down a canyon and enters a flat valley or plain builds a fan shaped deposit of sediment known as an alluvial fan. Alluvial fans are built mostly during flash floods. Alluvial fans are easy to recognize in arid areas but they form in wetter climates as well.

The diagram in figure 11 shows in cross-section how an alluvial fan develops over time as sediment is eroded from higher elevation and deposited on the adjacent lower elevation plain.

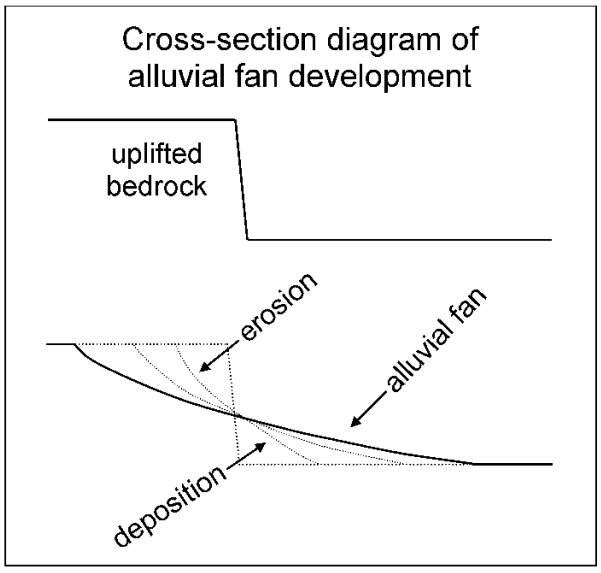


Figure 11.

Deltas

Deltas are important landforms to civilization. They provide fertile soils, flat land, and water for agriculture, as well as river channels for transportation.

A delta is a landform composed of sediment deposited where a stream enters a larger, slower moving body of water, such as an ocean, a lake, or a larger river. The term delta comes from the triangular shape of the Greek letter delta (Δ). Ancient Greek geographers recognized the triangular shape of the land created by the Nile River where it emptied into the Mediterranean Sea and gave the name delta to that landform. The Nile River delta is one of several types of deltas that are defined by the predominate processes that shape them. The Nile delta is a wave-dominated delta. Waves of the Mediterranean Sea have pushed and distributed sediment along the coast, flattening the seaward side of delta.

The Mississippi River delta is an example of a stream dominated delta. Deposition of sediments has built the delta into the Gulf of Mexico faster than waves or tides could redistribute the sediment. As commonly occurs in a delta, the Mississippi River splits in the downstream direction into several branches that discharge across the delta into the Gulf of Mexico. These branches are known as distributaries. The mouth of each distributary has built part of the delta farther out into the Gulf of Mexico forming what is known as a bird's foot delta, another name for a stream dominated delta based on the way it looks on a map.

The Ganges River delta is a tide dominated delta formed from sediment eroded from the Himalaya Mountains, the largest mountain range in the world. The mouth of the Ganges River is at the northern end of the Bay of Bengal, a large embayment of the Indian Ocean. The shape of this large bay has a magnifying effect on the tides. The combination of strong tides and the consistently high discharge from a river caring a large sediment load create a branching pattern of distributaries, in effect a braided stream system across the delta.

Deltas are lowlands that lie barely above sea level and are at high risk of being submerged under water. There are several ways in which deltas can be inundated by rising water. Floods coming down the river can cover a delta. Marine deltas can be subject to storm surges when extreme winds raise sea level along the coast and push ocean water inland. In the last several decades a new risk of submergence has arisen for marine deltas. Many marine deltas are undergoing gradual submergence as global sea level rises.

Contributors and Attributions

CC licensed content, Shared previously

- Basics -- Streams. **Authored by**: Ralph L. Dawes and Cheryl D. Dawes. **Provided by**: Wenatchee Valley College. **Located at**: http://commons.wvc.edu/rdawes/G101OCL/Basics/streams.html. **Project**: Geology 101 - Introduction to Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.11.2: Types of Streams and Rivers is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.3: Drainage Basins

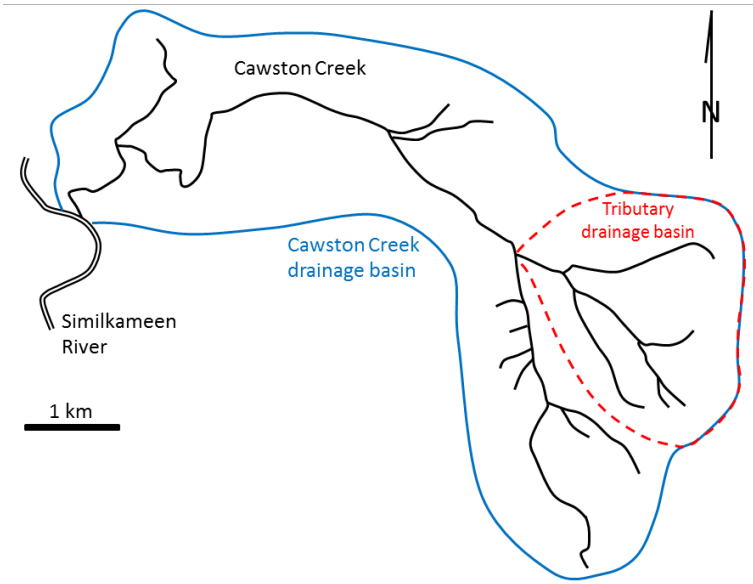


Figure 1. Cawston Creek near Keremeos, B.C. The blue line shows the extent of the drainage basin. The dashed red line is the drainage basin of one of its tributaries. [SE]

A **stream** is a body of flowing surface water of any size, ranging from a tiny trickle to a mighty river. The area from which the water flows to form a stream is known as its **drainage basin**. All of the precipitation (rain or snow) that falls within a drainage basin eventually flows into its stream, unless some of that water is able to cross into an adjacent drainage basin via groundwater flow. An example of a drainage basin is shown in Figure 1.

Cawston Creek is a typical small drainage basin (approximately 25 km^2) within a very steep glaciated valley. As shown in Figure 2, the upper and middle parts of the creek have steep **gradients** (averaging about 200 m/km but ranging from 100 to 350 m/km), and the lower part, within the valley of the Similkameen River, is relatively flat (<5 m/km). The shape of the valley has been controlled first by tectonic uplift (related to plate convergence), then by pre-glacial stream erosion and mass wasting, then by several episodes of glacial erosion, and finally by post-glacial stream erosion. The lowest elevation of Cawston Creek (275 m at the Similkameen River) is its **base level**. Cawston Creek cannot erode below that level unless the Similkameen River erodes deeper into its flood plain (the area that is inundated during a flood).

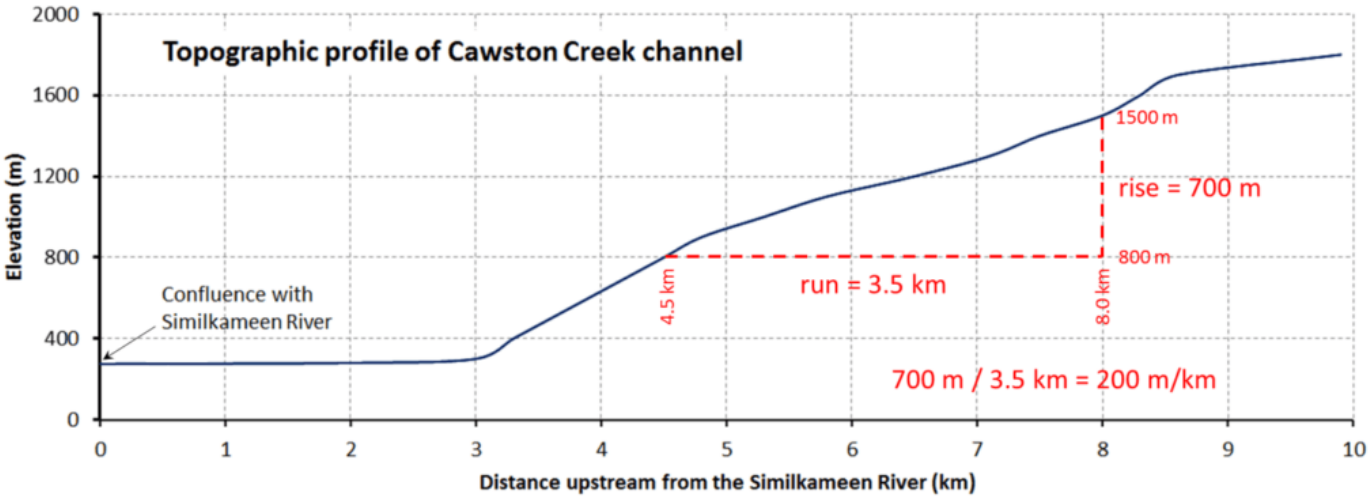


Figure 2. Profile of the main stem of Cawston Creek near Keremeos, B.C. The maximum elevation of the drainage basin is about 1,840 m, near Mount Kobau. The base level is 275 m, at the Similkameen River. As shown, the gradient of the stream can be determined by dividing the change in elevation between any two points (rise) by the distance between those two points (run). [SE]

Metro Vancouver's water supply comes from three large drainage basins on the north shore of Burrard Inlet. The boundary between two drainage basins is the height of land between them. A drop of water falling on the boundary between the Capilano and Seymour drainage basins (a.k.a., watersheds), for example, could flow into either one of them.

The pattern of tributaries within a drainage basin depends largely on the type of rock beneath, and on structures within that rock (folds, fractures, faults, etc.). The three main types of drainage patterns are illustrated in Figure 3. **Dendritic** patterns, which are by far the most common, develop in areas where the rock (or unconsolidated material) beneath the stream has no particular fabric or structure and can be eroded equally easily in all directions. Examples would be granite, gneiss, volcanic rock, and sedimentary rock that has not been folded. Most areas of British Columbia have dendritic patterns, as do most areas of the prairies and the Canadian Shield. **Trellis** drainage patterns typically develop where sedimentary rocks have been folded or tilted and then eroded to varying degrees depending on their strength. The Rocky Mountains of B.C. and Alberta are a good example of this, and many of the drainage systems within the Rockies have trellis patterns. **Rectangular** patterns develop in areas that have very little topography and a system of bedding planes, fractures, or faults that form a rectangular network. Rectangular drainage patterns are rare in Canada.

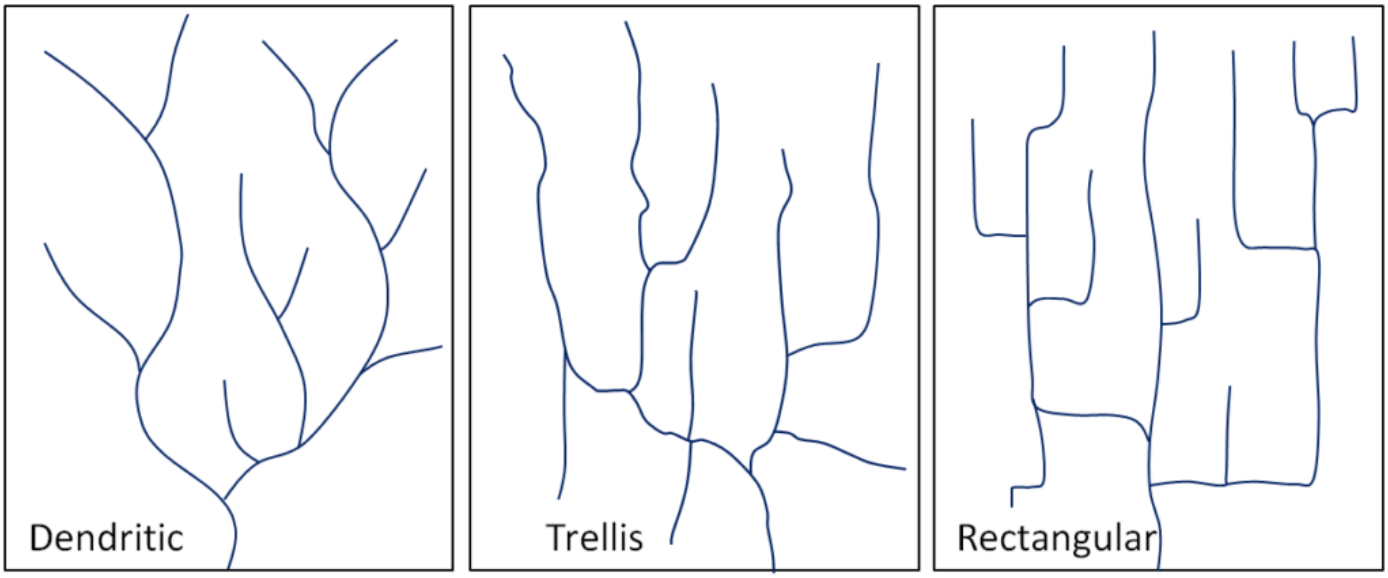


Figure 3. Typical dendritic, trellis, and rectangular stream drainage patterns. [SE]

In many parts of Canada, especially relatively flat areas with thick glacial sediments, and throughout much of Canadian Shield in eastern and central Canada, drainage patterns are chaotic, or what is known as **deranged** (Figure 4, left). Lakes and wetlands are common in this type of environment. A fourth type of drainage pattern, which is not specific to a drainage basin, is known as **radial** (Figure 4, right). Radial patterns form around isolated mountains (such as volcanoes) or hills, and the individual streams typically have dendritic drainage patterns.

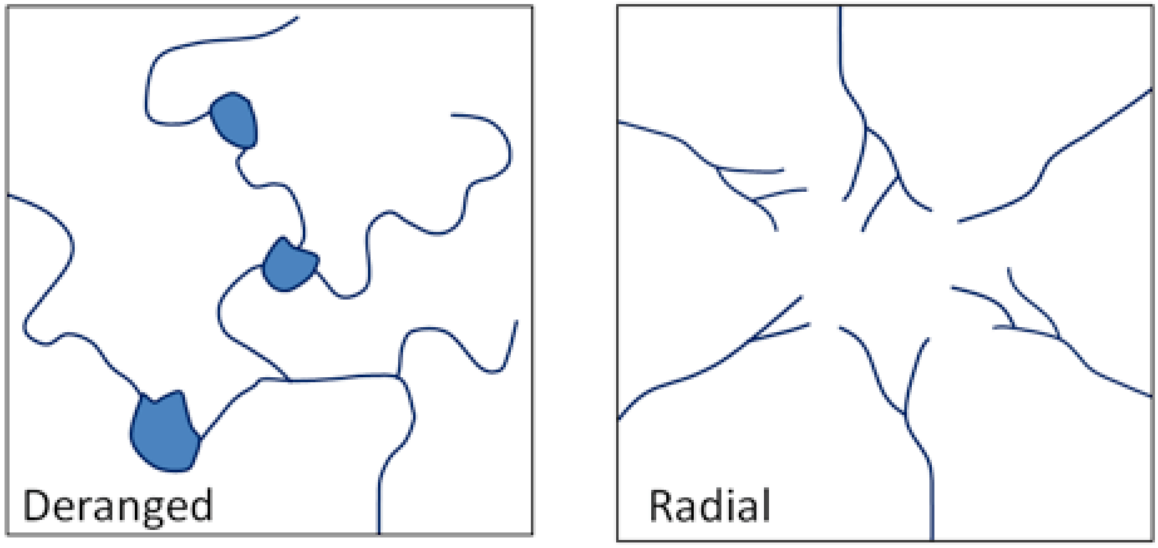


Figure 4. Left: a typical deranged pattern; right: a typical radial drainage pattern developed around a mountain or hill. [SE]

Over geological time, a stream will erode its drainage basin into a smooth profile similar to that shown in Figure 5. If we compare this with an ungraded stream like Cawston Creek (Figure 1), we can see that graded streams are steepest in their headwaters and their gradient gradually decreases toward their mouths. Ungraded streams have steep sections at various points, and typically have rapids and waterfalls at numerous locations along their lengths.

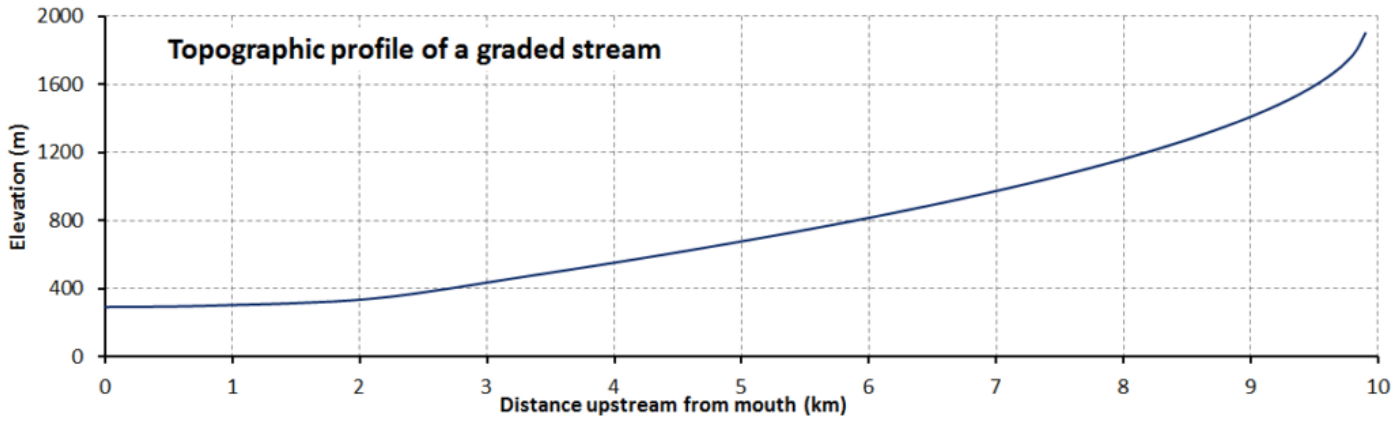


Figure 5. The topographic profile of a typical graded stream. [SE]

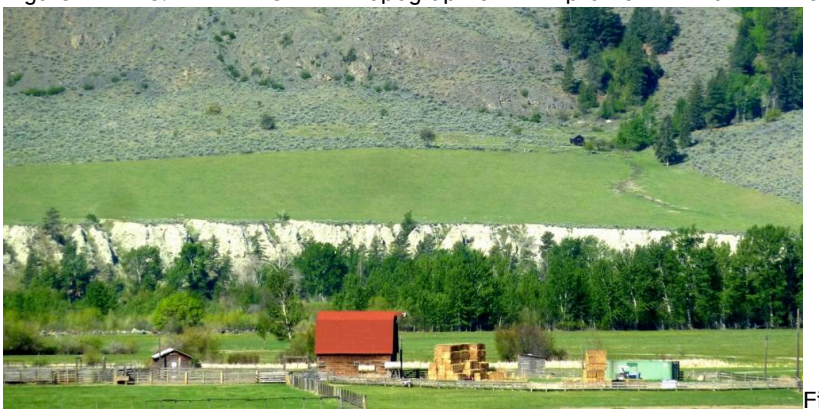


Figure 6. An example of a change in the base level of a small stream that flows into the Similkameen river near Keremeos. The previous base level was near the top of the sandy bank. The current base level is the river. [SE]

A graded stream can become ungraded if there is renewed tectonic uplift, or if there is a change in the base level, either

because of tectonic uplift or some other reason. As stated earlier, the base level of Cawston Creek is defined by the level of the Similkameen River, but this can change, and has done so in the past. Figure 6 shows the valley of the Similkameen River in the Keremeos area. The river channel is just beyond the row of trees. The green field in the distance is underlain by material eroded from the hills behind and deposited by a small creek (not Cawston Creek) adjacent to the Similkameen River when its level was higher than it is now. Sometime in the past several centuries, the Similkameen River eroded down through these deposits (forming the steep bank on the other side of the river), and the base level of the small creek was lowered by about 10 m. Over the next few centuries, this creek will seek to become graded again by eroding down through its own alluvial fan.

Another example of a change in base level can be seen along the Juan de Fuca Trail on southwestern Vancouver Island. As shown in Figure 7, many of the small streams along this part of the coast flow into the ocean as waterfalls. It is evident that the land in this area has risen by about 5 m in the past few thousand years, probably in response to deglaciation. The streams that used to flow directly into the ocean now have a lot of down-cutting to do to become regraded.



Figure 7. Two streams with a lowered base level on the Juan de Fuca Trail, southwestern Vancouver Island. [SE]

The ocean is the ultimate base level, but lakes and other rivers act as base levels for many smaller streams. We can create an artificial base level on a stream by constructing a dam.

Exercise: **The Effect of a Dam on Base Level**



Figure 8. Revelstoke Dam and Revelstoke Lake on the Columbia River at Revelstoke, BC [SE]

When a dam is built on a stream, a reservoir (artificial lake) forms behind the dam, and this temporarily (for many decades at least) creates a new base level for the part of the stream above the reservoir.

- How does the formation of a reservoir affect the stream where it enters the reservoir, and what happens to the sediment it

was carrying?
- The water leaving the dam has no sediment in it. How does this affect the stream below the dam?

Take a moment to think about your answers to these questions. You can record them in the space below, if you'd like:

[practice-area rows="4"][/practice-area]


Figure 9. Terraces on the Fraser River at High Bar. [Marie Betcher photo, used with permission]

Sediments accumulate within the flood plain of a stream, and then, if the base level changes, or if there is less sediment to deposit, the stream may cut down through those existing sediments to form terraces. A terrace on the Similkameen River is shown in Figure 6 and some on the Fraser River are shown in Figure 9. The Fraser River photo shows at least two levels of terraces.

In the late nineteenth century, American geologist William Davis proposed that streams and the surrounding terrain develop in a cycle of erosion (Figure 10). Following tectonic uplift, streams erode quickly, developing deep V-shaped valleys that tend to follow relatively straight paths. Gradients are high, and profiles are ungraded. Rapids and waterfalls are common. During the mature stage, streams erode wider valleys and start to deposit thick sediment layers. Gradients are slowly reduced and grading increases. In old age, streams are surrounded by rolling hills, and they occupy wide sediment-filled valleys. Meandering patterns are common.

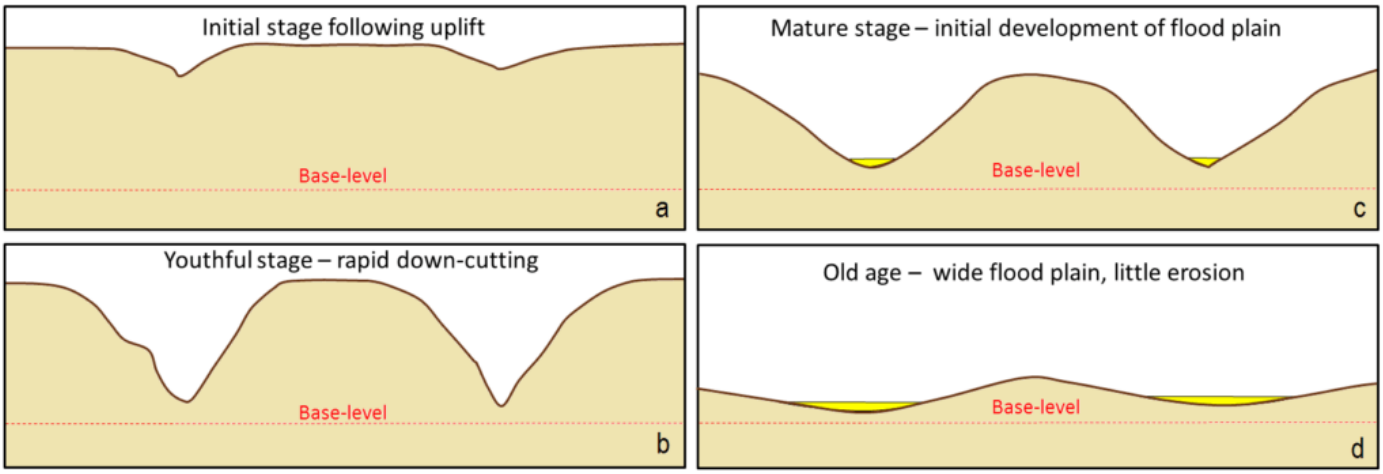


Figure 10. A depiction of the Davis cycle of erosion: a: initial stage, b: youthful stage, c: mature stage, and d: old age. [SE]

Davis's work was done long before the idea of plate tectonics, and he was not familiar with the impacts of glacial erosion on streams and their environments. While some parts of his theory are out of date, it is still a useful way to understand streams and their evolution.

Contributors and Attributions

CC licensed content, Shared previously

- Drainage Basins. **Authored by**: Steven Earle. **Provided by**: BC Campus. **Located at**: https://opentextbc.ca/geology/chapter/13-2-drainage-basins/. **Project**: Physical Geology. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.11.3: Drainage Basins is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.4: Geologic Processes and Flowing Water

 Figure 1. As streams flow towards the ocean, they carry weathered materials.

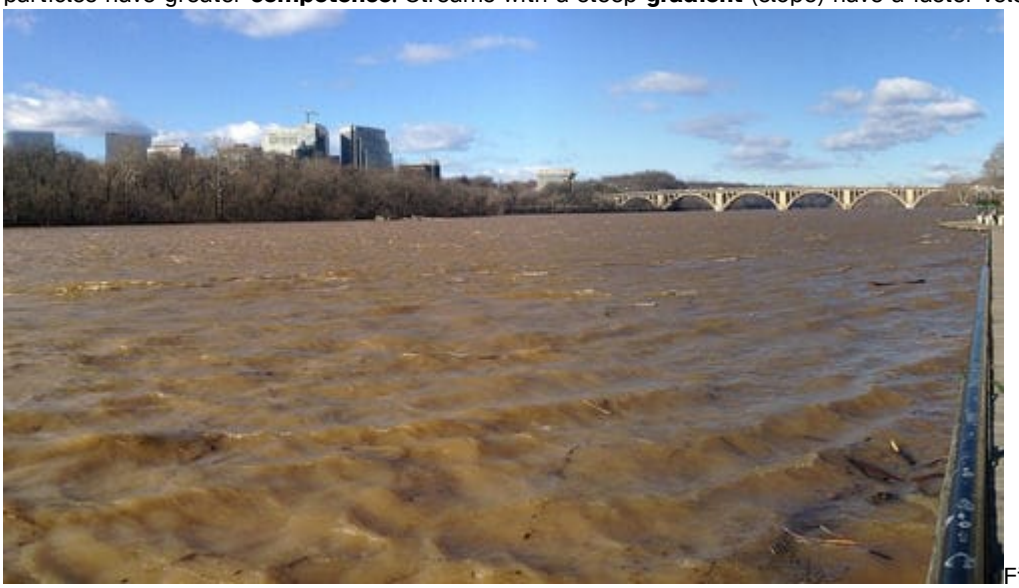
Streams—any running water from a rivulet to a raging river—complete the hydrologic cycle by returning precipitation that falls on land to the oceans (figure 1). Some of this water moves over the surface and some moves through the ground as **groundwater.** Flowing water does the work of both erosion and deposition.

Erosion and Deposition by Streams

Erosion by Streams

Flowing streams pick up and transport weathered materials by eroding sediments from their banks. Streams also carry ions and ionic compounds that dissolve easily in the water. Sediments are carried as:

- **Dissolved load**: Dissolved load is composed of ions in solution. These ions are usually carried in the water all the way to the ocean.
- **Suspended load**: Sediments carried as solids as the stream flows are suspended load. The size of particles that can be carried is determined by the stream's velocity (figure 2). Faster streams can carry larger particles. Streams that carry larger particles have greater **competence**. Streams with a steep **gradient** (slope) have a faster velocity and greater competence.

 Figure 2. Rivers carry sand, silt and clay as suspended load. During flood stage, the suspended load greatly increases as stream velocity increases.

- **Bed load**: Particles that are too large to be carried as suspended load are bumped and pushed along the stream bed as bed load. Bed load sediments do not move continuously. This intermittent movement is called **saltation**. Streams with high velocities and steep gradients do a great deal of down cutting into the stream bed, which is primarily accomplished by

movement of particles that make up the bed load.

- Here is a video of bedload transport.

Stages of Streams

As a stream flows from higher elevations, like in the mountains, towards lower elevations, like the ocean, the work of the stream changes. At a stream's **headwaters**, often high in the mountains, gradients are steep (figure 3). The stream moves fast and does lots of work eroding the stream bed.



Figure 3. This stream begins as snow melt from the mountains.

As a stream moves into lower areas, the gradient is not as steep. Now the stream does more work eroding the edges of its banks. Many streams develop curves in their channels called **meanders** (figure 4).



Figure 4. (a) At a meander, a stream actively erodes its outer banks and deposits material along the inside curves. This causes these meanders to migrate laterally over time. (b) This stream has deposited larger materials such as gravel and pebbles along the inside curve of a meander. (c) This image is a topographic map. The San Juan River eroded the land surface as the Colorado Plateau uplifted. The river's meanders were preserved as a feature called incised meanders.

As the river moves onto flatter ground, the stream erodes the outer edges of its banks to carve a **floodplain**, which is a flat level area surrounding the stream channel (figure 5).

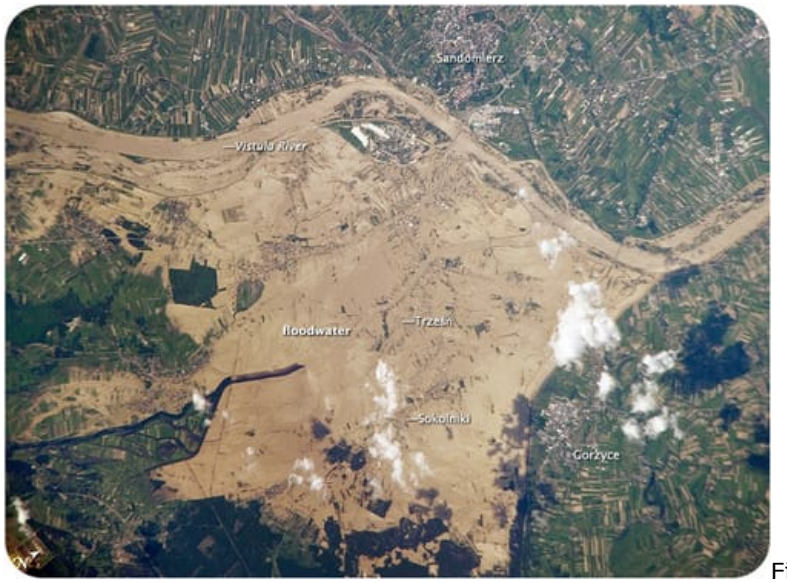


Figure 5. The Vistula River in Poland flows onto its floodplain.

Base level is where a stream meets a large body of standing water, usually the ocean, but sometimes a lake or pond. Streams work to down cut in their stream beds until they reach base level. The higher the elevation, the farther the stream is from where it will reach base level and the more cutting it has to do.

Stream Deposition

As a stream gets closer to base level, its gradient lowers and it deposits more material than it erodes. On flatter ground, streams deposit material on the inside of meanders. Placer mineral deposits, described in the Earth's Minerals chapter, are often deposited there. A stream's floodplain is much broader and shallower than the stream's channel. When a stream flows onto its floodplain, its velocity slows and it deposits much of its load. These sediments are rich in nutrients and make excellent farmland (figure 6).

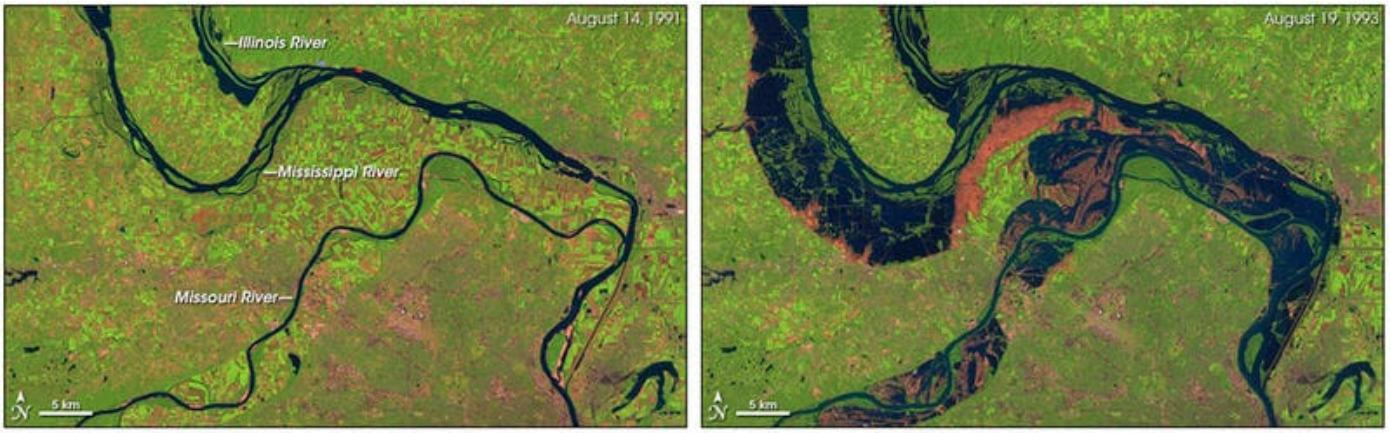


Figure 6. The Mississippi floodplain is heavily farmed. Flooding can wipe out farms and towns, but the stream also deposits nutrient-rich sediments that enrich the floodplain.

A stream at flood stage carries lots of sediments. When its gradient decreases, the stream overflows its banks and broadens its channel. The decrease in gradient causes the stream to deposit its sediments, the largest first. These large sediments build a higher area around the edges of the stream channel, creating **natural levees** (figure 7).

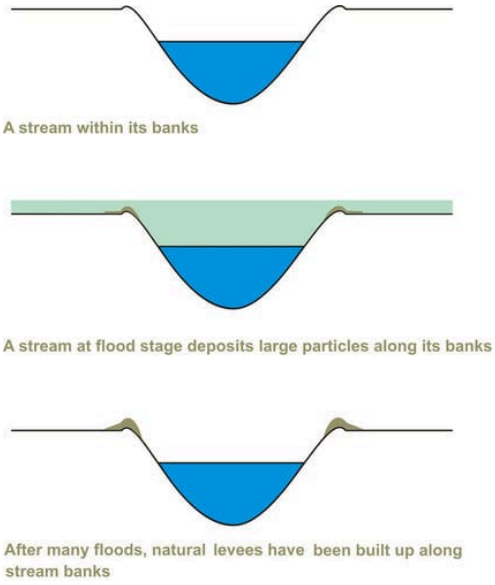


Figure 7. After many floods, a stream builds natural levees along its banks.

When a river enters standing water, its velocity slows to a stop. The stream moves back and forth across the region and drops its sediments in a wide triangular-shaped deposit called a **delta** (figure 8).



Figure 8. (a) The Nile River delta has a classic triangular shape, like the capital Greek letter delta. (b) Sediment in the Yellow River delta. The main stream channel splits into many smaller distributaries.

If a stream falls down a steep slope onto a broad flat valley, an **alluvial fan** develops (figure 9). Alluvial fans generally form in arid regions.



Figure 9. An alluvial fan in Iran. The mountains are in the lower right corner of the photograph.

Contributors and Attributions

CC licensed content, Shared previously

- 10.1: Water Erosion and Deposition. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/10.1/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.11.4: Geologic Processes and Flowing Water is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.5: Groundwater

Introduction

Although this may seem surprising, water beneath the ground is commonplace. Usually groundwater travels slowly and silently beneath the surface, but in some locations it bubbles to the surface at springs. The products of erosion and deposition by groundwater were described in the Erosion and Deposition chapter.

Groundwater

Groundwater is the largest reservoir of liquid fresh water on Earth and is found in **aquifers**, porous rock and sediment with water in between. Water is attracted to the soil particles and **capillary action**, which describes how water moves through a porous media, moves water from wet soil to dry areas.

Aquifers are found at different depths. Some are just below the surface and some are found much deeper below the land surface. A region may have more than one aquifer beneath it and even most deserts are above aquifers. The source region for an aquifer beneath a desert is likely to be far from where the aquifer is located; for example, it may be in a mountain area.

The amount of water that is available to enter groundwater in a region is influenced by the local climate, the slope of the land, the type of rock found at the surface, the vegetation cover, land use in the area, and water retention, which is the amount of water that remains in the ground. More water goes into the ground where there is a lot of rain, flat land, porous rock, exposed soil, and where water is not already filling the soil and rock.

The residence time of water in a groundwater aquifer can be from minutes to thousands of years. Groundwater is often called “fossil water” because it has remained in the ground for so long, often since the end of the ice ages.

Aquifers

Features of an Aquifer

To be a good aquifer, the rock in the aquifer must have good:

- **porosity**: small spaces between grains
- **permeability**: connections between pores

This animation shows porosity and permeability. The water droplets are found in the pores between the sediment grains, which is porosity. When the water can travel between ores, that’s permeability.

To reach an aquifer, surface water infiltrates downward into the ground through tiny spaces or pores in the rock. The water travels down through the permeable rock until it reaches a layer that does not have pores; this rock is **impermeable** (figure 1). This impermeable rock layer forms the base of the aquifer. The upper surface where the groundwater reaches is the **water table**.

Figure 1. Groundwater is found beneath the solid surface. Notice that the water table roughly mirrors the slope of the land’s surface. A well penetrates the water table.

The Water Table

For a groundwater aquifer to contain the same amount of water, the amount of recharge must equal the amount of discharge. What are the likely sources of recharge? What are the likely sources of discharge?

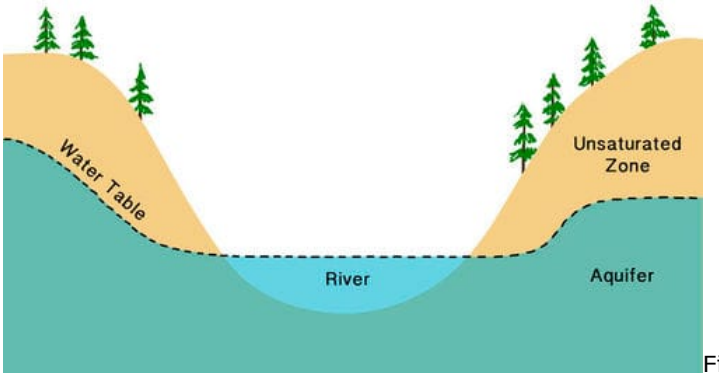


Figure 2. The top of the stream is the top of the water table. The stream feeds the aquifer.

In wet regions, streams are fed by groundwater; the surface of the stream is the top of the water table (figure 2). In dry regions, water seeps down from the stream into the aquifer. These streams are often dry much of the year. Water leaves a groundwater reservoir in streams or springs. People take water from aquifers, too.

What happens to the water table when there is a lot of rainfall? What happens when there is a drought? Although groundwater levels do not rise and fall as rapidly as at the surface, over time the water table will rise during wet periods and fall during droughts.

One of the most interesting, but extremely atypical types of aquifers is found in Florida. Although aquifers are very rarely underground rivers, in Florida water has dissolved the limestone so that streams travel underground and above ground (figure 3).

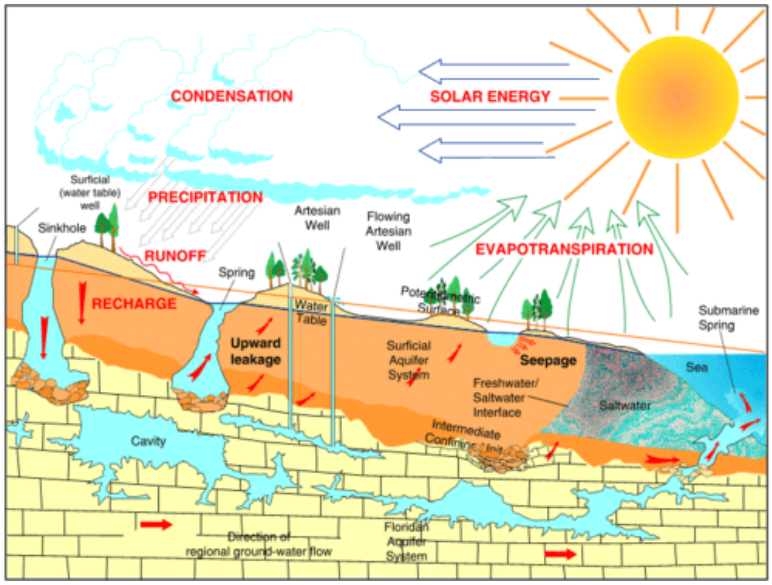


Figure 3. In Florida, groundwater is sometimes not underground.

Contributors and Attributions

CC licensed content, Shared previously

- 13.3: Groundwater. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/13.3/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.11.5: Groundwater is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.6: Porosity and Permeability

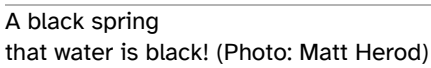

A black spring

Figure 1. A spring coming out of the shale near Red Creek. Yes, that water is black! (Photo: Matt Herod)

As we've learned, groundwater is simply water that exists underground. However, there are still lots of misconceptions about how people envision groundwater. Many envision large underground lakes and rivers, and while those do exist, they represent an infinitesimally small percentage of all groundwater. Generally speaking groundwater exists in the pore spaces between grains of soil and rocks. Imagine a water filled sponge. All of the holes in that sponge are water-filled. By squeezing that sponge we force the water out, similarly, by pumping an aquifer we force the water out of pore spaces.

There are lots of terms in hydrogeology, most of which are very simple, but essential. Here are a few of the big ones and their meanings.

Porosity

Porosity is an intrinsic property of every material. It refers to the amount of empty space within a given material. In a soil or rock the porosity (empty space) exists between the grains of minerals. In a material like gravel the grains are large and there is lots of empty space between them since they don't fit together very well. However, in a material like a gravel, sand and clay mixture the porosity is much less as the smaller grains fill the spaces. The amount of water a material can hold is directly related to the porosity since water will try and fill the empty spaces in a material. We measure porosity by the percentage of empty space that exists within a particular porous media.

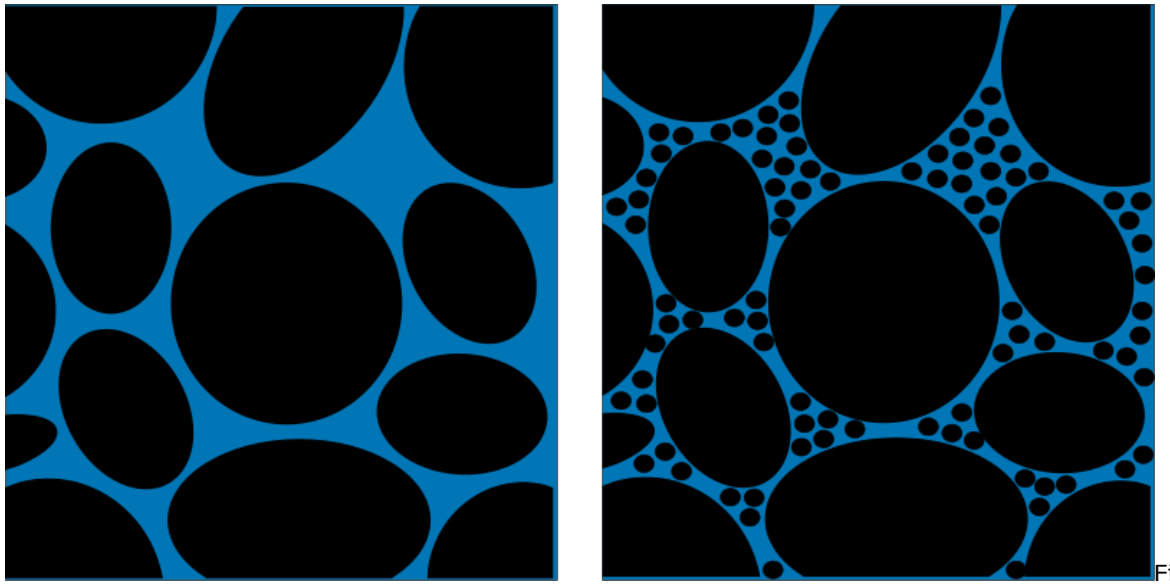


Figure 2. Porosity in two different media. The image on the left is analagous to gravel whereas on the right smaller particles are filling some of the pores and displacing water. Therefore, the water content of the material on the right is less. (Source: Wikipedia)

Permeability

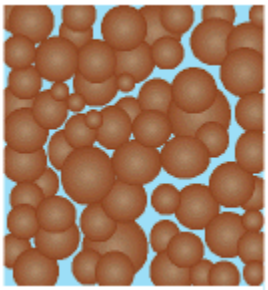


Figure 3. Video showing how connected pores have high permeability and can transport water easily. Note that some pores are isolated and cannot transport water trapped within them.

Permeability is another intrinsic property of all materials and is closely related to porosity. Permeability refers to how connected pore spaces are to one another. If the material has high permeability than pore spaces are connected to one another allowing water to flow from one to another, however, if there is low permeability then the pore spaces are isolated and water is trapped within them. For example, in a gravel all of the pores well connected one another allowing water to flow through it, however, in a clay most of the pore spaces are blocked, meaning water cannot flow through it easily.

Aquifer

An aquifer is a term for a type of soil or rock that can hold and transfer water that is completely saturated with water. That means that all it is simply a layer of soil or rock that has a reasonably high porosity and permeability that allows it to contain water and transfer it from pore to pore relatively quickly and all of the pore spaces are filled with water. Good examples of aquifers are glacial till or sandy soils which have both high porosity and high permeability. Aquifers allows us to recover groundwater by pumping quickly and easily. However, overpumping can easily reduce the amount of water in an aquifer and cause it to dry up. Aquifers are replenished when surface water infiltrates through the ground and refills the pore spaces in the aquifer. This process is called recharge. It is especially important to ensure that recharge is clean and uncontaminated or the entire aquifer could become polluted. There are two main types of aquifer. An unconfined aquifer is one that does not have an aquitard above it but usually does below it.

When a water-bearing rock readily transmits water to wells and springs, it is called an aquifer. Wells can be drilled into the aquifers and water can be pumped out. Precipitation eventually adds water (recharge) into the porous rock of the aquifer. The rate of recharge is not the same for all aquifers, though, and that must be considered when pumping water from a well. Pumping too much water too fast draws down the water in the aquifer and eventually causes a well to yield less and less water and even run dry. In fact, pumping your well too fast can even cause your neighbor's well to run dry if you both are pumping from the same aquifer.

In the diagram below, you can see how the ground below the water table (the blue area) is saturated with water. The "unsaturated zone" above the water table (the greenish area) still contains water (after all, plants' roots live in this area), but it is not totally saturated with water. You can see this in the two drawings at the bottom of the diagram, which show a close-up of how water is stored in between underground rock particles.

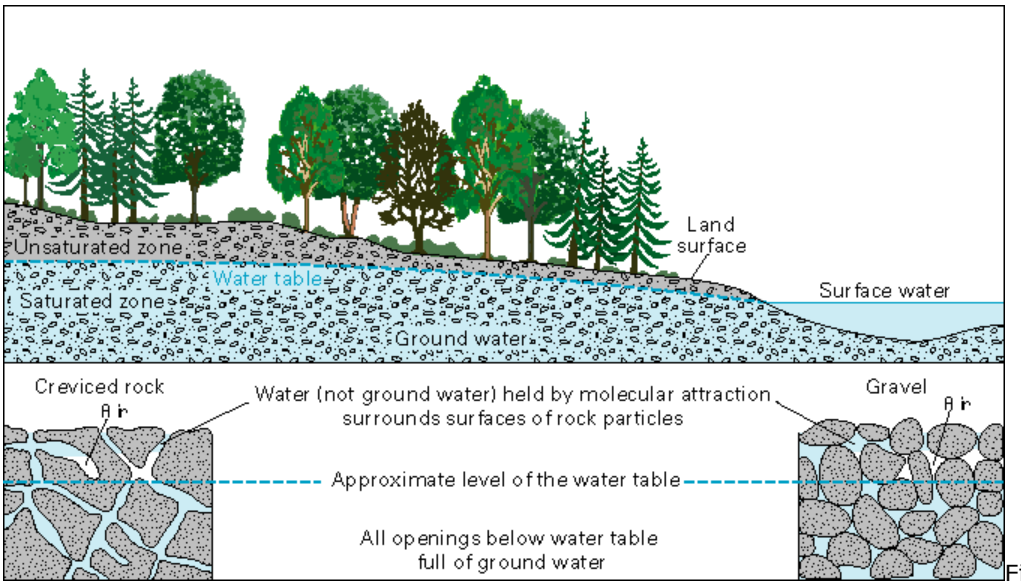


Figure 2.

Sometimes the porous rock layers become tilted in the earth. There might be a confining layer of less porous rock both above and below the porous layer. This is an example of a confined aquifer. In this case, the rocks surrounding the aquifer confines the pressure in the porous rock and its water. If a well is drilled into this “pressurized” aquifer, the internal pressure might (depending on the ability of the rock to transport water) be enough to push the water up the well and up to the surface without the aid of a pump, sometimes completely out of the well. This type of well is called artesian. The pressure of water from an artesian well can be quite dramatic.

A relationship does not necessarily exist between the water-bearing capacity of rocks and the depth at which they are found. A very dense granite that will yield little or no water to a well may be exposed at the land surface. Conversely, a porous sandstone, such as the Dakota Sandstone mentioned previously, may lie hundreds or thousands of feet below the land surface and may yield hundreds of gallons per minute of water. Rocks that yield freshwater have been found at depths of more than 6,000 feet, and salty water has come from oil wells at depths of more than 30,000 feet. On the average, however, the porosity and permeability of rocks decrease as their depth below land surface increases; the pores and cracks in rocks at great depths are closed or greatly reduced in size because of the weight of overlying rocks.

Water Movement in Aquifers

Water movement in aquifers is highly dependent of the permeability of the aquifer material. Permeable material contains interconnected cracks or spaces that are both numerous enough and large enough to allow water to move freely. In some permeable materials groundwater may move several metres in a day; in other places, it moves only a few centimeters in a century. Groundwater moves very slowly through relatively impermeable materials such as clay and shale.

After entering an aquifer, water moves slowly toward lower lying places and eventually is discharged from the aquifer from springs, seeps into streams, or is withdrawn from the ground by wells. Groundwater in aquifers between layers of poorly permeable rock, such as clay or shale, may be confined under pressure. If such a confined aquifer is tapped by a well, water will rise above the top of the aquifer and may even flow from the well onto the land surface. Water confined in this way is said to be under artesian pressure, and the aquifer is called an artesian aquifer.

Visualizing Artesian Pressure

Here’s a little experiment to show you how artesian pressure works. Fill a plastic sandwich baggie with water, put a straw in through the opening, tape the opening around the straw closed, do *not* point the straw towards your teacher or parents, and then squeeze the baggie. Artesian water is pushed out through the straw.

Aquitard

The other type is a confined aquifer that has an aquitard above and below it. An aquitard is basically the opposite of an aquifer with one key exception. Aquitards have very low permeability and do not transfer water well at all. In fact, in the ground they often act as a barrier to water flow and separate two aquifers. The one key exception is that aquitards can have high porosity and hold lots of water however, due to the their low permeability they are unable to transmit it from pore to pore and therefore water cannot flow within an aquitard very well. A good example of an aquitard is a layer of clay. Clay often has high porosity but almost no permeability meaning it is essentially a barrier which water cannot flow through and the water within it is trapped. However, there is still limited water flow within aquitards due to other processes that I won't get into now.

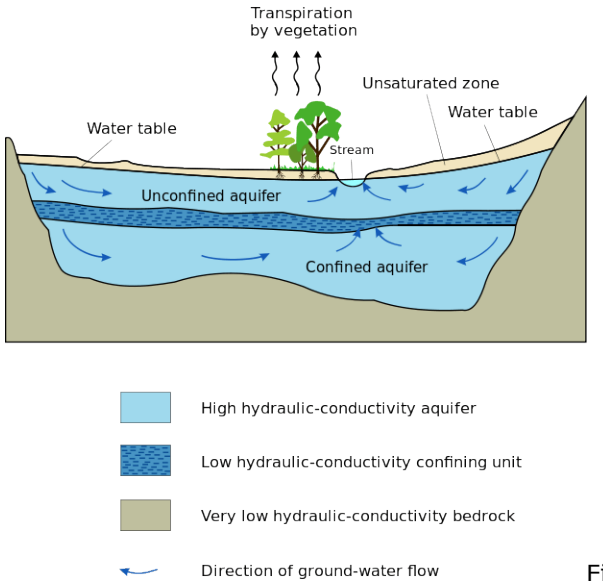


Figure 4.

Contributors and Attributions

CC licensed content, Shared previously

- Back to Basics on Groundwater. **Authored by**: Matt Herod. **Provided by**: European Geosciences Union. **Located at**: http://blogs.egu.eu/network/geosphere/2013/09/17/back-to-basics-on-groundwater/. **License**: *CC BY: Attribution*

Public domain content

- Aquifers. **Provided by**: USGS. **Located at**: http://water.usgs.gov/edu/earthgwaquifer.html. **License**: *Public Domain: No Known Copyright*

This page titled 4.11.6: Porosity and Permeability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.7: Aquifers and Well-Potential

There's a good chance that the average Joe who had to dig a well in ancient Egypt probably did the work with his hands, a shovel, and a bucket. He would have kept digging until he reached the water table and water filled the bottom of the hole. Some wells are still dug by hand today, but more modern methods are available. It's still a dirty job, though!

Wells are extremely important to all societies. In many places wells provide a reliable and ample supply of water for home uses, irrigation, and industries. Where surface water is scarce, such as in deserts,people couldn't survive and thrive without groundwater.

Types of Wells

Digging a well by hand is becoming outdated today (would YOU want to do it?). Modern wells are more often drilled by a truck-mounted drill rig. Still, there are many ways to put in a well—here are some of the common methods.

Dug Wells

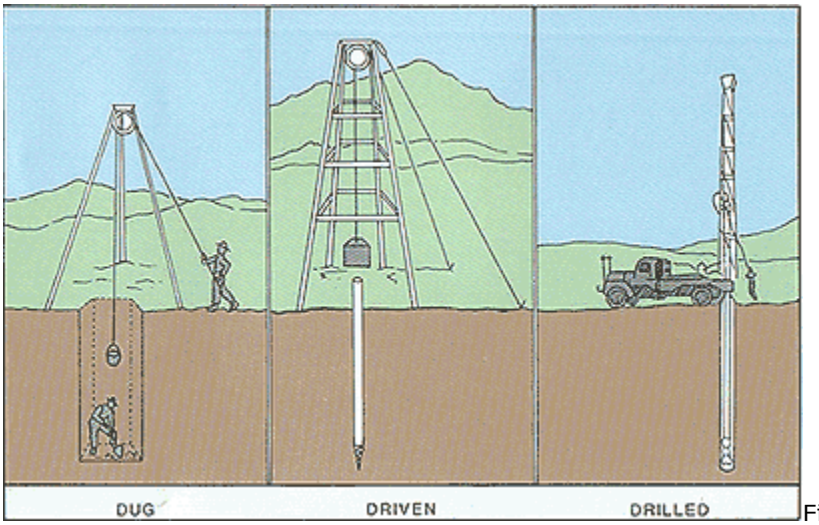


Figure 1. Well Types

Hacking at the ground with a pick and shovel is one way to dig a well. If the ground is soft and the water table is shallow,then dug wells can work. Historically, dug wells were excavated by hand shovel to below the water table until incoming water exceeded the digger's bailing rate . The well was lined with stones, brick, tile, or other material to prevent collapse, and was covered with a cap of wood, stone, or concrete. They cannot be dug much deeper than the water table — just as you cannot dig a hole very deep when you are at the beach... it keeps filling up with water!

Driven Wells

Driven wells are still common today. They are built by driving a small-diameter pipe into soft earth, such as sand or gravel. A screen is usually attached to the bottom of the pipe to filter out sand and other particles. Problems? They can only tap shallow water, and because the source of the water is so close to the surface, contamination from surface pollutants can occur.

Drilled Wells

Most modern wells are drilled, which requires a fairly complicated and expensive drill rig. Drill rigs are often mounted on big trucks. They use rotary drill bits that chew away at the rock, percussion bits that smash the rock, or, if the ground is soft,large auger bits. Drilled wells can be drilled more than 1,000 feet deep. Often a pump is placed at the bottom to push water up to the surface.

Water Levels in Wells

Ground-water users would find life easier if the water level in the aquifer that supplied their well always stayed the same. Seasonal variations in rainfall and the occasional drought affect the “height” of the underground water level. If a well is pumped at a faster rate than the aquifer around it is recharged by precipitation or other underground flow, then water levels around the well can be lowered. The water level in a well can also be lowered if other wells near it are withdrawing too much water. When water levels drop below the levels of the pump intakes, then wells will begin to pump air—they will “go dry.”

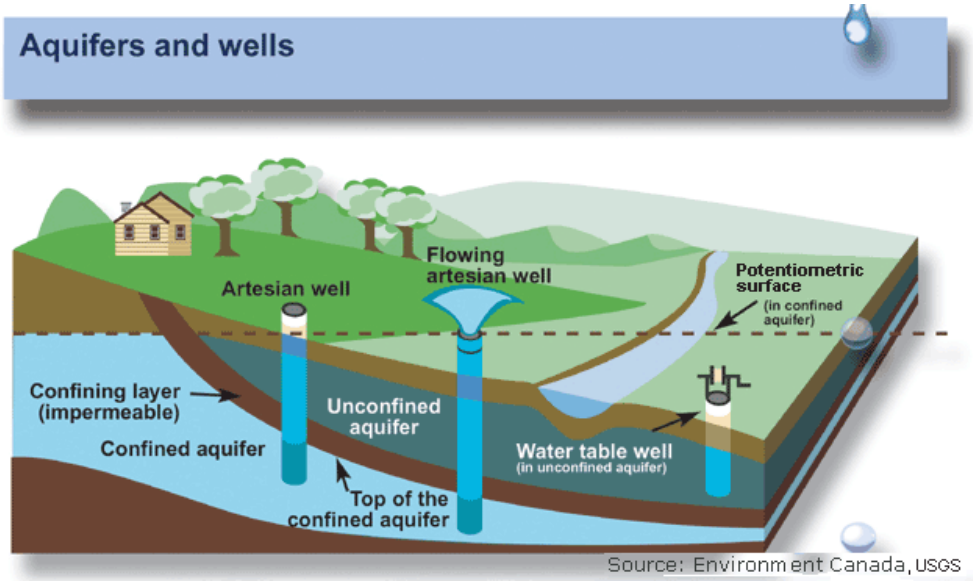


Figure 2. Aquifers and Wells

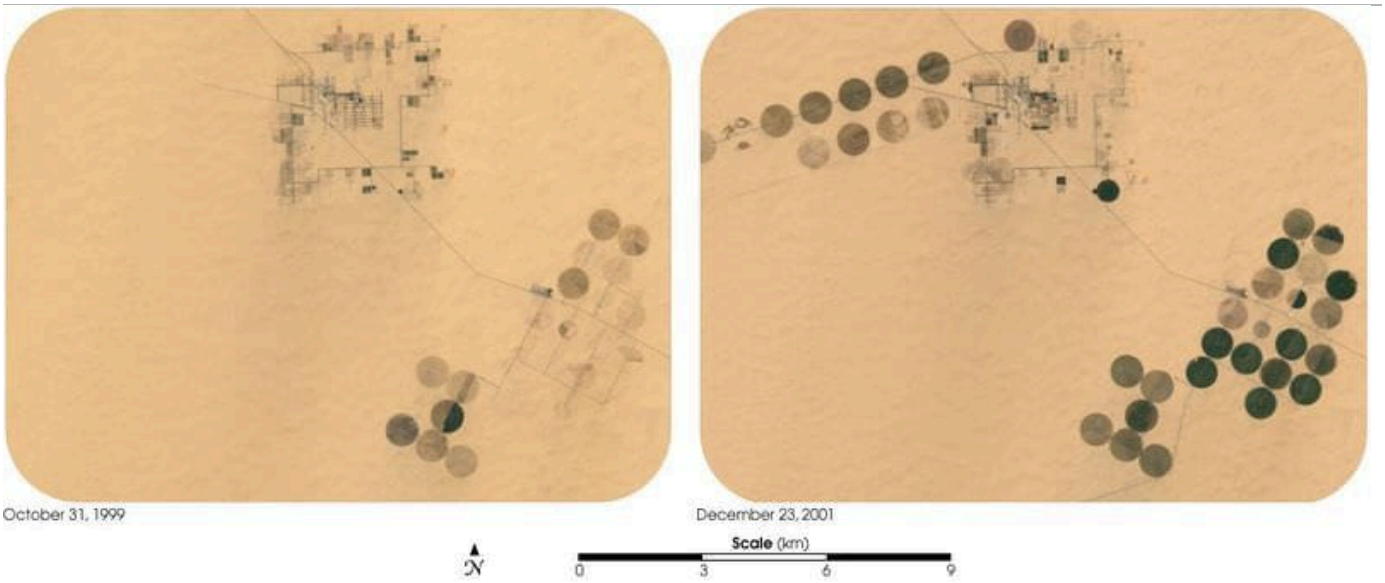
Contributors and Attributions

Public domain content

- Groundwater: Wells. **Authored by**: Roger M. Waller. **Provided by**: USGS. **Located at**: http://water.usgs.gov/edu/earthgwwells.html. **Project**: Ground Water and the Rural Homeowner. **License**: *Public Domain: No Known Copyright*

This page titled 4.11.7: Aquifers and Well-Potential is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.8: Groundwater Withdrawal



Is it good to make the desert bloom?

Many sunny, arid regions are good for growing crops as long as water can be added. Some of the increase in productivity is due to farming in regions that are technically too dry. Groundwater can be used to make the desert bloom, but at what cost? And for how long? Eventually the wells will run dry.

Groundwater Overuse

Some aquifers are overused; people pump out more water than is replaced. As the water is pumped out, the water table slowly falls, requiring wells to be dug deeper, which takes more money and energy. Wells may go completely dry if they are not deep enough to reach into the lowered water table.

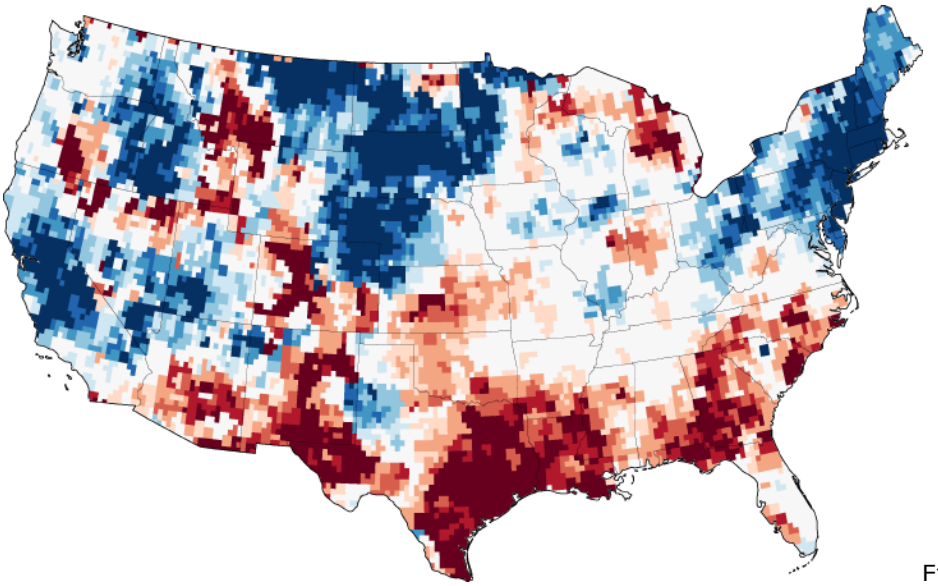


Figure 1. Intense drought has reduced groundwater levels in the southern U.S., particularly in Texas and New Mexico.

Other problems may stem from groundwater overuse. Subsidence and saltwater intrusion are two of them.

Ogallala Aquifer

The Ogallala Aquifer supplies about one-third of the irrigation water in the United States. The Ogallala Aquifer is widely used by people for municipal and agricultural needs. (Figure 2). The aquifer is found from 30 to 100 meters deep over an area of about 440,000 square kilometers!

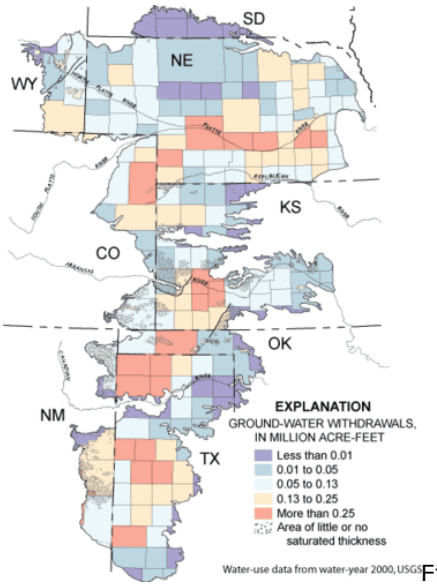
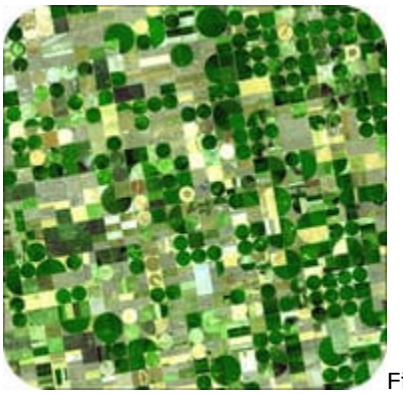


Figure 2. The Ogallala Aquifer is found beneath eight states and is heavily used.


Figure 3. Farms in Kansas

The water in the aquifer is mostly from the last ice age. About eight times more water is taken from the Ogallala Aquifer each year than is replenished. Much of the water is used for irrigation .Farms in Kansas use central pivot irrigation (Figure 3), which is more efficient since water falls directly on the crops instead of being shot in the air. These fields are between 800 and 1600 meters (0.5 and 1 mile) in diameter.

Subsidence

Lowering the water table may cause the ground surface to sink. Subsidence may occur beneath houses and other structures. The San Joaquin Valley of California is one of the world's major agricultural areas. So much groundwater has been pumped that the land has subsided many tens of feet.

Salt Water Intrusion

When coastal aquifers are overused, salt water from the ocean may enter the aquifer, contaminating the aquifer and making it less useful for drinking and irrigation. Salt water incursion is a problem in developed coastal regions, such as on Hawaii.

Summary

- When water is pumped from an aquifer, the water table declines and wells must be drilled deeper.
- The Ogallala Aquifer was filled in the ice age but is being used to irrigate the farms of the Midwestern U.S. at a rate far greater than it is being replenished.
- Ground subsidence and saltwater intrusion are two possible consequences of groundwater overuse.

Practice

Use this resource to answer the questions that follow.

A link to an interactive elements can be found at the bottom of this page.

1. How has irrigation changed farming?
2. What is leading to people's demands for additional water?
3. What do scientists need to see to better plan for future water use?
4. What is the GRACE satellite doing?
5. How does GRACE find groundwater aquifers?
6. How people know the aquifers are being depleted?
7. What is happening in India? what will happen if the water continues to decline?
8. What is the future of water?

Contributors and Attributions

CC licensed content, Shared previously

- Groundwater Depletion. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Groundwater-Depletion/lesson/Groundwater-Depletion-HS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*

All rights reserved content

- NASA | Science for a Hungry World: Part 5. **Authored by**: NASA Goddard. **Located at**: https://youtu.be/o1QsCa7RmmU. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

This page titled 4.11.8: Groundwater Withdrawal is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.9: Quality of Groundwater

As was noted at the very beginning of this chapter, one of the good things about groundwater as a source of water is that it is not as easily contaminated as surface water is. But there are two caveats to that: one is that groundwater can become naturally contaminated because of its very close connection to the materials of its aquifer, and the second is that once contaminated by human activities, groundwater is very difficult to clean up.

Natural Contamination of Groundwater

Groundwater moves slowly through an aquifer, and unlike the surface water of a stream, it has a lot of contact with the surrounding rock or sediment. In most aquifers, the geological materials that make up the aquifer are relatively inert, or are made up of minerals that dissolve very slowly into the groundwater. Over time, however, all groundwater gradually has more and more material dissolved within it as it remains in contact with the aquifer. In some areas, that rock or sediment includes some minerals that could potentially contaminate the water with elements that might make the water less than ideal for human consumption or agricultural use. Examples include copper, arsenic, mercury, fluorine, sodium, and boron. In some cases, contamination may occur because the aquifer material has particularly high levels of the element in question. In other cases, the aquifer material is just normal rock or sediment, but some particular feature of the water or the aquifer allows the contaminant to build up to significant levels.

An example of natural contamination takes place in the bedrock aquifers of the east coast of Vancouver Island and the adjacent Gulf Islands. The aquifer is the Cretaceous (90 Ma to 65 Ma) Nanaimo Group, which is made up of sandstone, mudstone, and conglomerate (Figure 1).



Figure 1. Cretaceous Nanaimo Group sandstone exposed in a Nanaimo parking lot [SE]

The rocks of the Nanaimo Group are not particularly enriched in any trace elements, but the submarine-fan sandstone that makes up much of the group is a lithic wacke, and therefore has relatively high levels of clay (for a sandstone). This clay is good at adsorbing[1] some elements from the water and desorbing others, and in the process, its pH goes up (it becomes alkaline). At high pH levels (some as high as 9 in the Nanaimo Group), the element fluorine that is present naturally in the rock (as it is in almost any rock) has an increased tendency to dissolve in the water. In some areas, groundwater in the Nanaimo Group has fluorine levels that are well above recommended levels for drinking water. The World Health Organization (WHO) maximum acceptable concentration (MAC) for fluorine is 1.5 mg/L (milligrams per litre). Between 5% and 10% of the domestic wells around Nanaimo and adjacent Gabriola Island have more than that, some as much as 10 mg/L. A small amount of fluorine in the human diet is considered important for maintaining dental health, but high levels can lead to malformation and discolouration of teeth, and long-term exposure can lead to other more serious health effects such as skeletal problems.

Nanaimo Group groundwater can also have elevated levels of boron, again related to pH and adsorption from clay minerals. While

boron at the levels found there is not toxic to humans, there is enough boron in some wells to be toxic to plants, and the water cannot be used for irrigation.

Figure 2. The distribution of arsenic in groundwater in Bangladesh. The WHO recommended safe level for arsenic is 10 µg/L. All of the green, orange, and red areas on the map exceed that limit. [From: BGS and DPHE. 2001. Arsenic contamination of groundwater in Bangladesh. Kinniburgh, D G and Smedley, P L, http://www.bgs.ac.uk/arsenic/bangladesh/.]

Rural residents in the densely populated country of Bangladesh (over 1,000 residents/km2, compared with 3.4/km2 in Canada) used to rely mostly on surface supplies for their drinking water, and many of these were subject to bacterial contamination. Infant mortality rates were among the highest in the world and other illnesses such as diarrhea, dysentery, typhoid, cholera, and hepatitis were common. In the 1970s, international agencies, including UNICEF, started a program of drilling wells to access abundant groundwater supplies at depths of 20 m to 100 m. Eventually over 8 million such wells were drilled. Infant mortality and illness rates dropped dramatically, but it was later discovered that the water from a high proportion of these wells has arsenic above safe levels (Figure 2).

Most of the wells in the affected areas are drilled into relatively recent sediments of the vast delta of the Ganges and Brahmaputra Rivers. While these sediments are not particularly enriched in arsenic, they have enough organic matter in them to use up any oxygen present. This leads to water with a naturally low oxidation potential (anoxic conditions); arsenic is highly soluble under these conditions, and so any arsenic present in the sediments easily gets dissolved into the groundwater. Arsenic poisoning leads to headaches, confusion, and diarrhea, and eventually to vomiting, stomach pain, and convulsions. If not treated, the final outcomes are heart disease, stroke, cancer, diabetes, coma, and death. There are ways to treat arsenic-rich groundwater, but it is a challenge in Bangladesh to implement the simple and effective technology that is available.

Anthropogenic Contamination of Groundwater

Groundwater can become contaminated by pollution at the surface (or at depth), and there are many different anthropogenic (human-caused) sources of contamination.

The vulnerability of aquifers to pollution depends on several factors, including the depth to the water table, the permeability of the material between the surface and the aquifer, the permeability of the aquifer, the slope of the surface, and the amount of precipitation. Confined aquifers tend to be much less vulnerable than unconfined ones, and deeper aquifers are less vulnerable than shallow ones. Steeper slopes mean that surface water tends to run off rather than infiltrate (and this can reduce the possibility of contamination). Contamination risk is also less in dry areas than in areas with heavy rainfall.

Studies of groundwater vulnerability have been completed for various regions of British Columbia. A groundwater vulnerability map for southern Vancouver Island is shown in Figure 3. The yellow to red areas are considered to have high vulnerability to pollution from surface sources, and most of these are where the aquifers are unconfined in quite permeable unconsolidated sediments of either glacial or fluvial origin, where the water table is relatively shallow and the terrain is relatively flat.

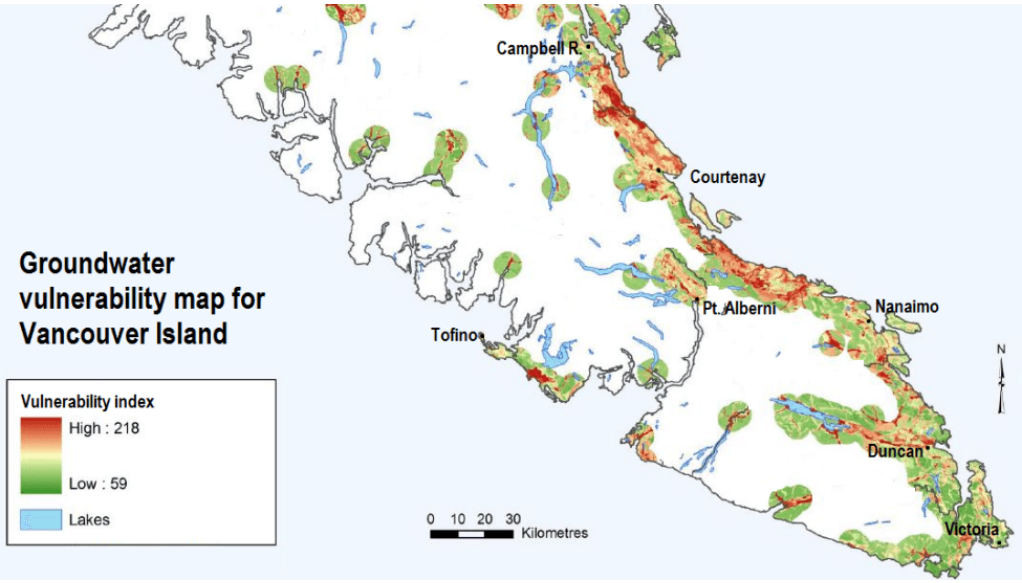


Figure 3. The vulnerability to anthropogenic contamination of aquifers on southern Vancouver Island. Much of the island is not mapped (shown as white) because of a lack of aquifer information in areas without wells. [From: Newton, P. and Gilchrist, A. 2010. Technical summary of intrinsic vulnerability mapping methods of Vancouver Island, Vancouver Island Water Resources Vulnerability Mapping Project, Vancouver Island University, 45pp. Used with permission. https://web.viu.ca/groundwater/PDF/VI_DRASTIC_Summary_Phase2_2010.pdf]

The important sources of anthropogenic groundwater contamination include the following:

- Chemicals and animal waste related to agriculture, and chemicals applied to golf courses and domestic gardens
- Landfills
- Industrial operations
- Mines, quarries, and other rock excavations
- Leaking fuel storage tanks (especially those at gas stations)
- Septic systems
- Runoff from roads (e.g., winter salting) or chemical spills of materials being transported

Agriculture

Intensive agricultural operations and golf courses can have a significant impact on the environment, especially where chemicals and other materials are used to enhance growth or control pests. An example of agricultural contamination is in the Abbotsford area of the Fraser Valley, where nitrate levels above the 44 mg/L maximum acceptable level (expressed as nitrate) in the Abbotsford-Sumas aquifer have been observed since the 1950s; however, the problem became much worse as agriculture intensity increased in the 1980s. By 2004, groundwater with nitrate levels in excess of 44 mg/L was reported over an area of about 75 km2 around Abbotsford, and the problem extended across the border into the Sumas area of Washington State.

This region is intensively used for berry crops (especially raspberries and blueberries) and large poultry operations, as well as lesser amounts of grazing and forage crops. Chicken manure is typically stored in fields adjacent to chicken barns, and may release nitrogen to the environment from runoff water, and from releases of ammonia gas. Over decades, both chemical fertilizers and chicken manure and other manures have been applied to the berry crops to provide extra nitrogen to help maximize berry growth. If the fertilizer added is in excess of what the plants need, or is poorly timed compared to when it is needed, then the extra nitrogen may be leached into the groundwater below. Berry crops are irrigated over the summer to help the crops grow. Summer irrigation and winter rainfall may carry excess nitrate from the near surface to the aquifer below.

Since the 1990s, agricultural practices have been tightened up to reduce the rate of groundwater contamination, but it will take decades for nitrate levels to drop in the Abbotsford-Sumas aquifer. Agriculture and Agri-Food Canada and many others are conducting research on better irrigation and nitrate management techniques to reduce the amount of nitrogen that leaches to groundwater.

Landfills

In the past, domestic and commercial refuse was commonly trucked to a “dump” (typically a hole in the ground), and when the hole was filled, it was covered with soil and forgotten. In situations like this, rain and melting snow can easily pass through the soil used to cover the refuse. This water passes into the waste itself, and the resulting landfill **leachate** that flows from the bottom of the landfill can seriously contaminate the surrounding groundwater and surface water. In the past few decades, regulations around refuse disposal have been significantly strengthened, and important steps have been taken to reduce the amount of landfill waste by diverting recyclable and compostable materials to other locations.

A modern engineered landfill has an impermeable liner (typically heavy plastic, although engineered clay liners or natural clay may be adequate in some cases), a plumbing system for draining leachate (the rainwater that flows through the refuse and becomes contaminated), and a network of monitoring wells both within and around the landfill (Figure 4). Once part or all of a landfill site is full, it is sealed over with a plastic cover, and a system is put in place to extract **landfill gas** (typically a mixture of carbon dioxide and methane). That gas can be sent to a nearby location where it is burned to create heat or used to generate electricity. The leachate must be treated, and that can be done in a normal sewage treatment plant.

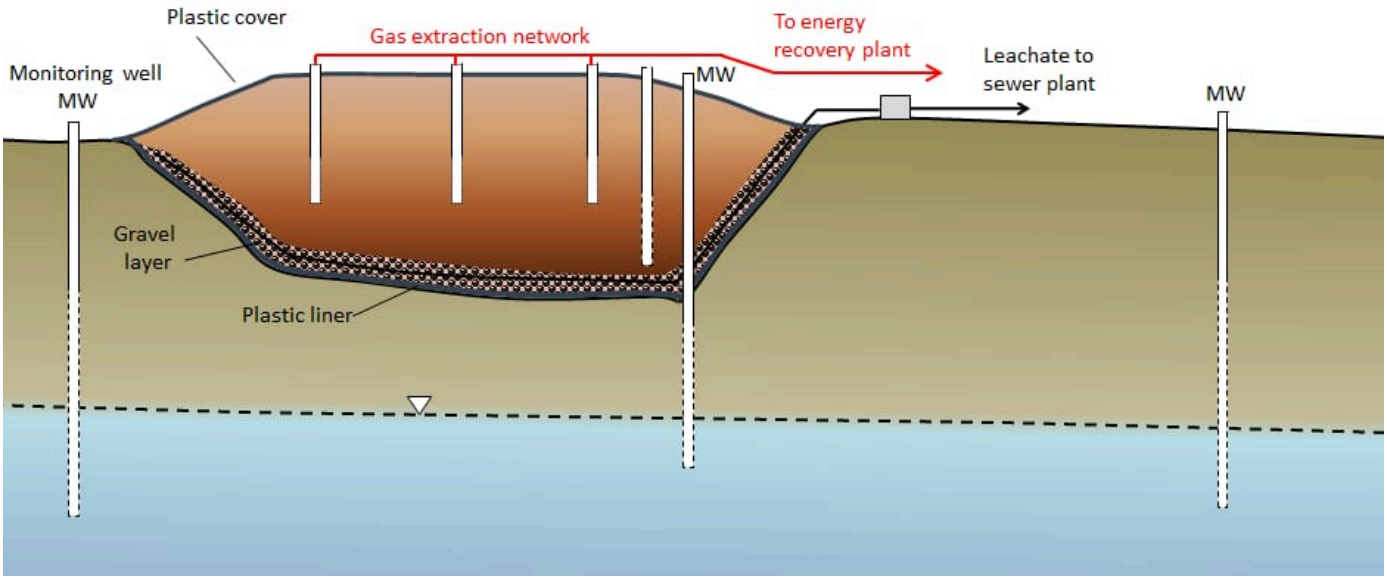


Figure 4 A cross-section of a typical modern landfill [SE]

The monitoring wells are used to assess the level of the water table around the landfill and to collect groundwater samples so that any leakage can be detected. Because some leakage is almost inevitable, the ideal placement for landfills is in areas where the depth to the water table is significant (tens of metres if possible) and where the aquifer material is relatively impermeable. Landfills should also be situated far from streams, lakes, or wetlands so that contamination of aquatic habitats can be avoided.

Today there are hundreds of abandoned dumps scattered across the country; most have been left to contaminate groundwater that we might wish to use sometime in the future. In many cases, it’s unlikely that we’ll be able to do so.

Exercises: What Goes on at Your Landfill?



Figure 5. A landfill [SE photo]

Unless you live in a remote rural area, there's a good chance that the refuse you can't recycle is picked up at the curb and taken to a landfill. Most landfills are operated by cities or regional districts, and you should be able to find information about yours on the appropriate local government website. See if you can answer some the following questions:

1. Which government body operates your landfill?
2. Where is the landfill situated?
3. Is your waste all placed in a landfill, or are there other processes in use (e.g., incineration or composting)?
4. Are landfill gases captured, and, if so, what is done with them?
5. What could be changed to improve the waste disposal situation in your community (e.g., more recycling, compost collection, waste-to-energy technology)?

Industrial Operations

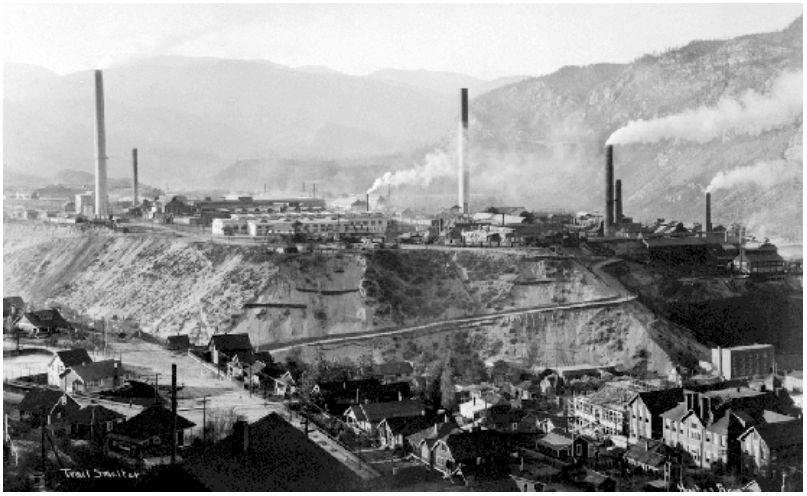


Figure 6. The Trail lead-zinc smelter in 1929 [http://upload.wikimedia.org /Wikipedia/commons/2/20 /Trail_Smelter_in_Year_1929.png]

Although western Canada doesn't have the same extent of industrial pollution as other parts of the country, there are still seriously contaminated sites in the west, most with the potential to contaminate groundwater. One example is the lead and zinc smelter at Trail, B.C. The largest in the world, it has been operating for over 100 years and has left a residue of metal contamination around the region (Figure 6). In some parts of Trail, the contamination is serious enough that existing soil has been removed from residential properties and replaced with clean soil brought in from elsewhere. This contaminated soil has contributed to contamination of groundwater in the Trail area. Groundwater beneath the actual smelter site is contaminated, and the operator (Teck Resources) is currently working on plans to prevent that water from reaching the nearby Columbia River.

Mines, Quarries, and Rock Excavations



Figure 7. Acidic runoff at the abandoned Mt. Washington Mine near Courtenay, B.C. [SE]

Mines and other operations that involve the excavation of large amounts of rock (e.g., highway construction) have the potential to create serious environmental damage. The exposure of rock that has previously not been exposed to air and water can lead to the oxidation of sulphide-bearing minerals, such a pyrite (FeS_2), within the rock. The combination of pyrite, water, oxygen, and a special type of bacteria (*Acidithiobacillus ferrooxidans*) that thrives in acidic conditions leads to the generation of acidity, in some cases to pH less than 2. Water that acidic is hazardous by itself, but the low pH also has the property of increasing the solubility of certain heavy metals. The water that is generated by this process is known as acid rock drainage (ARD). ARD can occur naturally where sulphide-bearing rocks are near the surface. The issue of ARD is a major environmental concern at both operating mines and abandoned mines. In streams around the Mt. Washington Mine on Vancouver Island (Figure 7), copper levels are high enough to be toxic to fish. Groundwater adjacent to the contaminated streams in the area is very likely contaminated as well.

Leaking Fuel Tanks



Figure 8. A closed and fenced gas station site in Nanaimo, B.C. The white pipes in the background are wells for monitoring groundwater contamination on the site. [SE]

Underground storage tanks (USTs) are used to store fuel at gas stations, industrial sites, airports, and anywhere that large volumes of fuel are used. They do not last forever, and eventually they start to leak their contents into the ground. This is a particular problem at older gas stations — although it may also become a future problem at newer gas stations. You may have noticed gas stations that have been closed and then surrounded by chain-link fence (Figure 8). In virtually all such cases the closure has been triggered by the discovery of leaking USTs and the requirement to cease operations and remediate the site.Petroleum fuels are complex mixtures of hydrocarbon compounds and the properties of their components—such as density, viscosity, solubility in water, and volatility — tend to vary widely. As a result, a petroleum spill is like several spills for the price of one. The petroleum liquid slowly settles through the unsaturated zone and then tends to float on the surface of the groundwater (Figure 9). The more readily soluble components of the spill dissolve in the groundwater and are dispersed along with the normal groundwater flow, and the more volatile components of the spill rise toward the surface, potentially contaminating buildings.

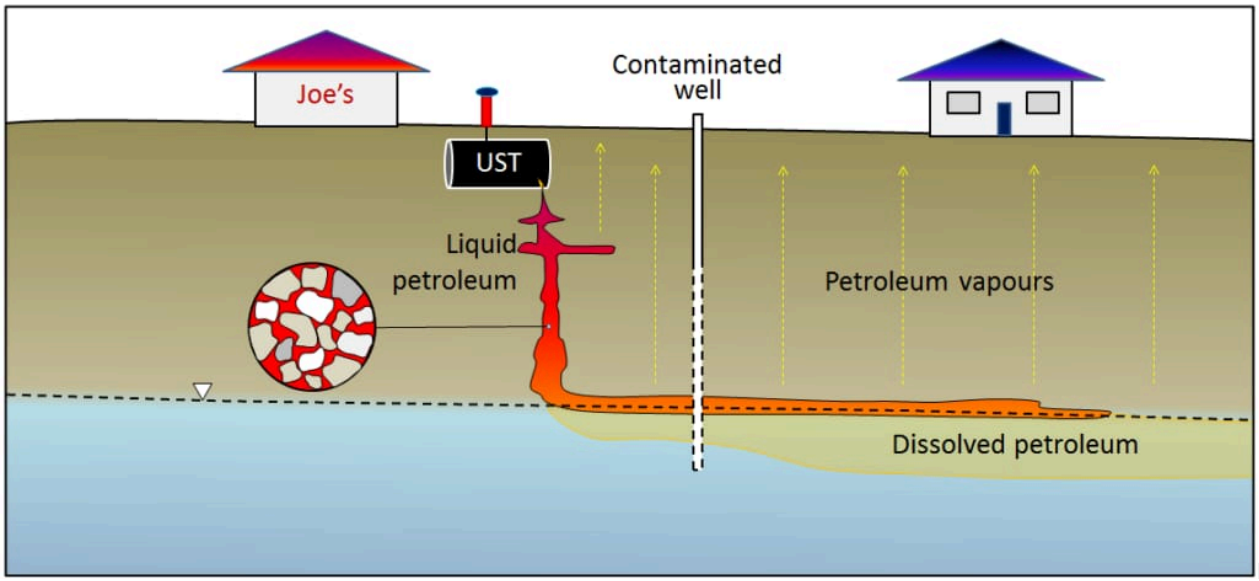


Figure 9. A depiction of the fate of different components of a petroleum spill from an underground storage tank. [SE]

Exercises: Find a Leaking UST in Your Community

There is almost certainly a leaking UST at a former gas station near you. Look for an empty property that is surrounded by a chain-link fence with "No Trespassing" signs. You might see evidence of monitoring wells (like those shown in Figure 9), and there could be some petroleum barrels around that are being used to store contaminated water. Once you've identified one of these, you'll probably start seeing them everywhere!

Septic Systems

In areas that are not served by sewage networks leading to a central sewage treatment plant, most homeowners rely on **septic systems** for disposal of sewage. There are two primary components to a simple septic system, the septic tank and the drainage field (Figure 10). A typical septic tank is constructed of either concrete or plastic and has a volume of 5,000 L to 10,000 L (5 m3 to 10 m3). This forms the first treatment and is designed to be **anaerobic** (without oxygen). That promotes the activity of certain bacteria that help break down the waste. As the waste is degraded, some portions tend to sink to form sludge at the base of the tank, and others float to the surface, forming a scum layer. A septic tank may be divided into two parts to keep the sludge at the bottom and the scum on the top from draining out. The water then moves to the drainage field, which provides the right conditions for a different set of bacteria that operate in **aerobic** conditions. The drainage field includes an array of plastic pipes that are perforated to allow the effluent to drain out over a large area and seep slowly into the ground. In order to install a drainage field, it is first necessary to test the soil below, as it must be sufficiently permeable to allow the effluent to percolate away, but not so permeable that it flows too quickly and the soil is not able to filter out the pathogenic bacteria.

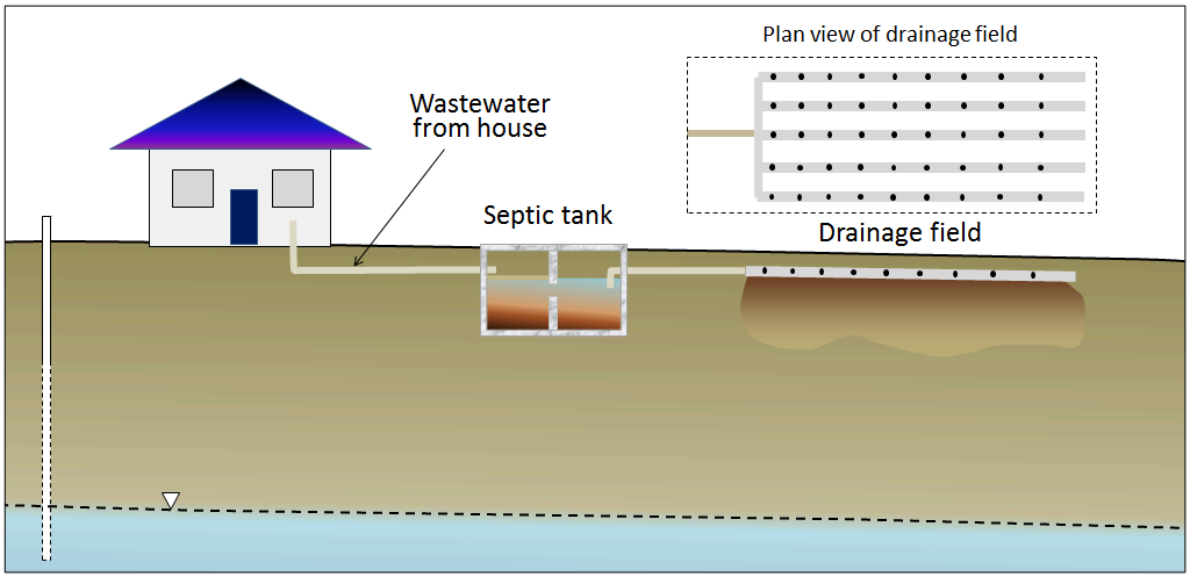


Figure 10. A typical septic system. [SE]

If they are properly installed and used, and if the sludge is periodically removed from the tank, a septic system should be effective in treating the sewage for decades. The anaerobic and aerobic bacteria should be able to break down the incoming waste and there should be little risk to the surface environment or groundwater. But many things can go wrong with a septic system, including the following:

- If inappropriate chemicals are added to the waste stream, they may interfere with the natural breakdown of the sewage.
- If the tank is not periodically pumped out, solids can get into the drainage field and compromise the drainage, resulting in the flow of effluent toward the surface.
- If the soil is either not sufficiently permeable or too permeable, the effluent will not drain away (and will start to pool at the surface) or it will drain too quickly.
- If the drainage field is constructed in an area where the water table is close to surface, some of the effluent is likely to flow into the groundwater without being treated.

Prevention and Mitigation of Groundwater Contamination

As illustrated in the landfill example above, there are two fairly simple ways to significantly reduce the chance and degree of groundwater contamination from surface sources. One is to prevent rainwater from infiltrating down to the water table and picking up contaminants; this can be achieved by simply capping or roofing over the landfill, mine tailings, or spill site. The second is to provide an impermeable barrier beneath the contaminant. Modern landfills and mine tailings impoundments are all built using some combination of clay and engineered plastic barriers. Both of these solutions — caps and liners — are subject to failure due to leaks.

Once contaminants are in the groundwater, the main form of remediation is to pump out the contaminated water and treat it at the surface. This can be a slow process, and preventing the contaminant from travelling significantly during this process can be accomplished by manipulating local groundwater flow through the extraction or injection of water at certain locations. Consider this in the exercise below.

Exercises: **Manipulating a Contaminant Plume**

This diagram shows a groundwater contaminant plume in red. The source of the contamination has been removed but if the plume is not dealt with, it will eventually enter the stream and threaten the health of wildlife. Pumping the contaminant from well B for treatment will not be sufficient to prevent some of the contamination from making it to the stream.

What could you do at wells A and C to prevent this? Explain and use the diagram below to illustrate the expected changes to the water table and the movement of the plume.

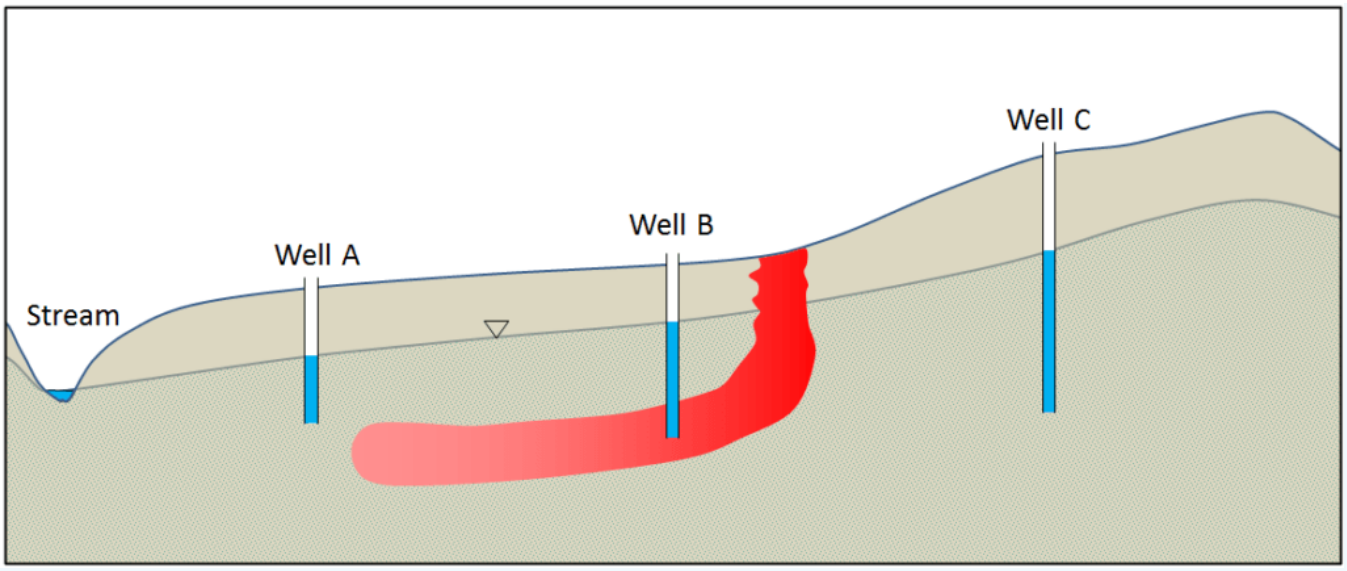


Figure 11. Contaminant Plume

1. “Adsorb” (with a “d”) is not the same as “absorb” (with a “b”). Water can be absorbed by a sponge. Ions dissolved in water can be adsorbed onto—or desorbed from—the surfaces of clay minerals. ↵

Contributors and Attributions

CC licensed content, Shared previously

- Groundwater Quality. **Authored by**: Steven Earle. **Provided by**: BC Campus. **Located at**: https://opentextbc.ca/geology/chapter/14-4-groundwater-quality. **Project**: Physical Geology. **License**: *CC BY: Attribution*

This page titled 4.11.9: Quality of Groundwater is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.10: Karst Topography

Introduction

Throughout the world **karst** landscapes vary from rolling hills dotted with sinkholes, as found in portions of the central United States, to jagged hills and pinnacle karst found in the tropics. The development of all karst landforms requires the presence of rock which is capable of being dissolved by surface water or **ground water**.

The term karst describes a distinctive topography that indicates **dissolution** (also called chemical solution) of underlying soluble rocks by surface water or ground water. Although commonly associated with **carbonate rocks** (limestone and dolomite) other highly soluble rocks such as evaporates (gypsum and rock salt) can be sculpted into karst terrain.

Understanding caves and karst is important because ten percent of the Earth's surface is occupied by karst landscape and as much as a quarter of the world's population depends upon water supplied from karst areas. Though most abundant in humid regions where carbonate rock is present, karst terrain occurs in temperate, tropical, alpine and polar environments. Karst features range in scale from microscopic (chemical precipitates) to entire drainage systems and ecosystems which cover hundreds of square miles, and broad karst plateaus.

Although karst processes sculpt beautiful landscapes, karst systems are very vulnerable to ground water pollution due to the relatively rapid rate of water flow and the lack of a natural filtration system. This puts local drinking water supplies at risk of being contaminated. In the mid 1980's, flooding of caves in the highly populated area of Bowling Green, Kentucky, caused industrial waste to leak into the vast system of underground fissures polluting the ground water in local wells. Due to urban expansion millions of dollars is spent annually in the United States to repair damage to roads, buildings and other structures which are built on unstable karst surfaces.

Karst Topography

The degree of development of karst landforms varies greatly from region to region. Large drainage systems in karst areas are likely to have both **fluvial** (surface) and karst (underground) drainage components. As stated in the introduction, the term **karst** describes a distinctive topography that indicates **dissolution** of underlying rocks by surface water or ground water.

Water falls as rain or snow and soaks into the soil. The water becomes weakly acidic because it reacts chemically with carbon dioxide that occurs naturally in the atmosphere and the soil. This acid is named carbonic acid and is the same compound that makes carbonated beverages taste tangy. Rainwater seeps downward through the soil and through fractures in the rock responding to the force of gravity. The carbonic acid in the moving ground water dissolves the bedrock along the surfaces of joints, fractures and bedding planes, eventually forming cave passages and caverns.

Limestone is a sedimentary rock consisting primarily of calcium carbonate in the form of the mineral calcite. Rainwater dissolves the limestone by the following reaction: Calcite + Carbonic acid = Calcium ions dissolved in ground water + Bicarbonate ions dissolved in ground water.

Cracks and joints that interconnect in the soil and bedrock allow the water to reach a zone below the surface of the land where all the fractures and void spaces are completely filled (also known as saturated) with water. This water-rich zone is called the saturated zone and its upper surface is called the **water table.** The volume of void space (space filled with air or water) in soil or bedrock is termed **porosity.** The larger the proportion of voids in a given volume of soil or rock the greater the porosity. When these voids are interconnected, water or air (or other fluids) can migrate from void to void. Thus the soil or bedrock is said to be **permeable** because fluids (air and water) can easily move through them. Permeable bedrock makes a good aquifer, a rock layer that holds and conducts water. If the ground water that flows through the underlying permeable bedrock is acidic and the bedrock is soluble, a distinctive type of topography, karst topography, can be created.

The first part of our animation shows evolution of karst landforms created by downward movement of water accompanied by dissolution of rock and mass transport of sediments in stream channels. In tropical areas with thick massive limestones, a remarkable and distinctive landscape of jagged hills and narrow gorges completely dominates the landscape. Movement of solution along fractures and joints etches the bedrock and leaves limestone blocks as isolated spires or pinnacles. Pinnacles

range from small features a few inches tall to intermediate forms a few feet tall to large pinnacles hundreds of feet tall. Besides the etching of pinnacles and residual hills, sheets of flowing water move down sloping surfaces creating a variety of etched surface features. Our computer animation shows the dominant landforms, such as pinnacles, cones, and towers, commonly found in the tropical karst environment of northern Puerto Rico.

Our paper model represents another type of karst landscape, that of a rolling limestone plain such as is found in south-central Kentucky, northern Florida, and the Highland Rim of central Tennessee where **doline** karst is the dominate feature. Doline karst is the most widely distributed type of karst landscape. The landscape is dotted with **sinkholes** (dolines) which can vary widely in number and size. For the Sinkhole Plain in central Kentucky, there are approximately 5.4 sinkholes per square kilometer over a 153 square kilometer area. For north Florida there are almost 8 sinkholes per square kilometer over a 427 square kilometer area (White, 1988, table 4.1, page 100).

Karst topography dominated by sinkholes or dolines usually has several distinct surface features. Our paper model shows features normally associated with karst topography. Sinkholes (also known as dolines) are surface depressions formed by either: 1) the dissolution of bedrock forming a bowl-shaped depression, or 2) the collapse of shallow caves that were formed by dissolution of the bedrock. These sinkholes or shallow basins may fill with water forming lakes or ponds. Springs are locations where ground water emerges at the surface of the earth. **Disappearing streams** are streams which terminate abruptly by flowing or seeping into the ground. Disappearing streams are evidence of disrupted surface drainage and thus indicate the presence of an underground drainage system. **Cave entrances** are natural openings in the earth large enough to allow a person to enter. Caves may reflect a complex underground drainage system.

A Brief List of the Longest Caves in the United States

Name	Location	Distance Mapped*
Mammoth Cave—Flint Ridge System	Kentucky	500 km
Jewel Cave	South Dakota	118 km
Wind Cave	South Dakota	73 km
Friars Hole System	West Virginia	68 km
Fisher Ridge Cave System	Kentucky	64 km

* Numbers are rounded to nearest kilometer

What do Caves Contain?

Moving water may transport earth materials into and through caves physically or chemically. Caves contain interesting features as a result of the physical and chemical processes that form them. Among these features are **breakdown** blocks of rock formed by collapse of cave ceilings. Also seen are sediments containing boulders, sand, silt, and clay deposited from water flowing in and through cave passages and conduits. **Speleogens** are irregular or distinctive shapes of carbonate rock etched from bedrock by dripping or running water. Speleogens can form where bedrock is not uniform in chemical composition. Consequently, the less soluble rock dissolves slower than adjacent more soluble rock through time. The less soluble rock tends to stand in relief and projects from walls and ceilings of caves.

Away from their entrances, caves usually provide a relatively constant temperature and humidity over a long period of time. Thus, caves provide an ideal environment for chemical deposition of minerals. As water laden with dissolved carbonate seeps into the air-filled cave passage, it may lose excess carbon dioxide to the cave atmosphere, or the water itself may evaporate, causing the dripwater to precipitate secondary carbonate or other minerals from solution, creating cave formations or **speleothems** including cone-shaped **stalactites**, **stalagmites**, **flowstone** or rimstone, or other interesting shapes. Caves in karst areas often have stalactites (icicle-like masses of chemical limestone) that hang from cave ceilings and stout stalagmites protruding from the cave floor. Stalactites and stalagmites can be a few inches to several feet long. Sometimes the drip water will flow down the walls and over the cave floor creating flowstone or rimstone deposits. Where drip water seeps from a joint and then drips over the edges of ledges, deposits of great complexity known as draperies are formed. The color of dripstones and flowstones comes from organic and/or iron oxide compounds brought in from the surface, giving the speleothems an orange brown color or from the presence of oxides and hydroxides of iron and manganese which give the speleothems a deep brown or black color.

What Lives Underground?

Some scientists are interested in cave ecology and how cave animals interact with cave **microclimates.** Animals found in caves include everything from surface dwelling animals like raccoons that occasionally use the cave, to animals that have adapted exclusively to life in the cave **(troglobites)**. Troglobites cannot survive outside caves. These may include such diverse animals as eyeless fish and crayfish, cave beetles, flatworms, and other unusual types of insects. Many of these animals have lost body pigmentation and are white or transparent in color. Although the cave environment appears to be stable, change can and does occur. The temperature of the cave varies due to air movement near the entrances and the temperature of water entering the cave. In reality, some caves have their own weather systems which create wind due to temperature and pressure differences between the entrance and interior passageways.

Many animals, such as bats, cave crickets, and pack rats, regularly visit, raise their young, or hibernate in caves. These animals are called **trogloxenes** . Caves may support large numbers of different types of bats. Bats may be among the most beneficial animals to people and the ecosystem as insect-eaters and plant pollinators. The little brown bat can eat 600 mosquitoes in an hour thus performing the work of a “natural insecticide,” helping control crop pests and other insects. The Mammoth Cave-Flint Ridge System in Kentucky, which is the most extensive cave system in the world, has a biodiversity of 43 mammals, 15 reptiles, 19 amphibians and 3 fish. In 1981, the United Nations designated Mammoth Cave National Park as a World Heritage Site. Follow this link for an excellent summary of the Mammoth Cave area and other caves found in the U. S. National Park System.

In 1988 the United States passed the Federal Cave Resource Protection Act which preserves and protects all significant caves found on federal land for future generations of Americans.

What’s in it for Me?

Knowing where karst features are located could help city and town planners, as well as individual landowners, to make decisions on where to build houses and other structures. This information could save cities thousand of dollars in repairs to buildings that are built on unstable karst terrain.

Karst springs supply drinking water to millions of people. Knowledge of karst terrain and the movement of water in underground drainage systems is important for maintaining good quality and safe drinking water. Pollution of ground water is a major problem in karst terrain.

Caves provide a venue for recreation. Although most of the caves located in National Parks are protected, there are over 200 commercial show caves nationwide which are open to the public. Recreational caving has become a popular hobby. The National Speleological Society has about 20,000 active affiliates nationwide.

Deposits preserved in caves can tell geologists about past climates. Fossils and artifacts found in caves help geologists and archaeologists unravel the prehistory of an area.

Caves support a unique community of bacteria, fungi and animals not seen on the surface of the Earth.

Questions

- Why is there often a wind at the mouth of a cave?
- Do large caves form in dry environments above the water table?
- Do dripstone features such as stalactites and stalagmites form in caves that are below or above the water table? Why or why not?
- Is the water table always level?
- Should a person explore a cave alone?
- Should a person build a house near a sinkhole?
- Should a person collect stalactites and stalagmites?
- Should a person collect cave-dwelling animals?

Contributors and Attributions

Public domain content

- Karst Topography - Teacher's Guide and Paper Model. **Authored by**: Tau Rho Alpha, John P. Galloway, and John C. Tinsley. **Provided by**: USGS. **Located at**: http://geomaps.wr.usgs.gov/parks/cave/karst.html. **License**: *Public Domain: No Known Copyright*

This page titled 4.11.10: Karst Topography is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.11: Geysers

Geysers, fumaroles (also called *solfataras*), and hot springs are generally found in regions of young volcanic activity. Surface water percolates downward through the rocks below the Earth's surface to high-temperature regions surrounding a magma reservoir, either active or recently solidified but still hot. There the water is heated, becomes less dense, and rises back to the surface along fissures and cracks. Sometimes these features are called "dying volcanoes" because they seem to represent the last stage of volcanic activity as the magma, at depth, cools and hardens.

Geysers


Figure 1. Old Faithful Geyser, Yellowstone National Park, Wyoming.

Heated groundwater may become trapped in spaces within rocks. Pressure builds up as more water seeps into the spaces. When the pressure becomes great enough, the water bursts out of the ground at a crack or weak spot. This is called a geyser. When the water erupts from the ground, the pressure is released. Then more water collects and the pressure builds up again. This leads to another eruption.

Old Faithful (Figure 1) is the best-known geyser in the world. The geyser erupts faithfully every 90 minutes, day after day. During each eruption, it may release as much as 30,000 liters of water!

Erupting geysers provide spectacular displays of underground energy suddenly unleashed, but their mechanisms are not completely understood. Large amounts of hot water are presumed to fill underground cavities. The water, upon further heating, is violently ejected when a portion of it suddenly flashes into steam. This cycle can be repeated with remarkable regularity, as for example, at Old Faithful Geyser in Yellowstone National Park, which erupts on an average of about once every 65 minutes.

Fumaroles


Figure 2. Black Growler steam vents (fumaroles), Norris Basin, Yellowstone National Park, Wyoming.

Fumaroles, which emit mixtures of steam and other gases, are fed by conduits that pass through the water table before reaching the surface of the ground. Hydrogen sulfide (H_2S), one of the typical gases issuing from fumaroles, readily oxidizes to sulfuric acid and native sulfur. This accounts for the intense chemical activity and brightly colored rocks in many thermal areas.

Hot Springs

Hot springs occur in many thermal areas where the surface of the Earth intersects the water table. The temperature and rate of discharge of hot springs depend on factors such as the rate at which water circulates through the system of underground channelways, the amount of heat supplied at depth, and the extent of dilution of the heated water by cool ground water near the surface.

Contributors and Attributions

CC licensed content, Shared previously

- Springs and Geysers. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Springs-and-Geysers/lesson/Springs-and-Geysers-MS-ES/?referrer=featured_content. **License**: *CC BY-NC: Attribution-NonCommercial*

Public domain content

- Geysers, Fumaroles, and Hot Springs. **Provided by**: USGS. **Located at**: http://pubs.usgs.gov/gip/volc/geysers.html. **License**: *Public Domain: No Known Copyright*

This page titled 4.11.11: Geysers is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.12: Springs

The top of an aquifer may be high enough in some places to meet the surface of the ground. This often happens on a slope. The water flows out of the ground and creates a spring. A spring may be just a tiny trickle, or it may be a big gush of water.

Water flowing out of the ground at a spring may flow downhill and enter a stream. If the water from a spring can't flow downhill, it may spread out to form a pond or lake instead. In the desert, the only reliable water may be from springs (Figure 1). A spring may allow wildlife to inhabit an uninhabitable area.


Figure 1. A desert oasis is created by a spring in Libya.

Artesian Spring

Sometimes an aquifer is confined. A confined aquifer is trapped between two impermeable rock layers. Pressure from the rock layer on top forces the water out where the aquifer reaches the ground surface. Water that flows up to the surface naturally is an artesian spring. If people drill a well into a confined aquifer, the water may flow to the surface without assistance. This is an artesian well (Figure 2).


Figure 2. This artesian well supplies the water for Schönbrunn Palace in Germany.

Mineral Springs and Hot Springs


Figure 3. Morning Glory Pool in Yellowstone National Park is a mineral spring.

Some springs have water that contains minerals. Groundwater dissolves minerals out of the rock as it seeps through the pores. The water in some springs is hot because it is heated by hot magma. Many hot springs are also mineral springs. That's because hot water can dissolve more minerals than cold water.

Springs in Yellowstone National Park are hot and contain dissolved minerals. Morning Glory Pool (Figure 3) has a bright green color from dissolved minerals. Along the edge are thick orange mats of bacteria. The bacteria use the minerals in the hot water to make food.

Contributors and Attributions

CC licensed content, Shared previously

- Springs and Geysers. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Springs-and-Geysers/lesson/Springs-and-Geysers-MS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*
- Schlou00df Schonbrunn fountain. **Authored by**: Alastair Rae. **Located at**: https://flic.kr/p/5pdE6S. **License**: *CC BY-SA: Attribution-ShareAlike*
- Morning Glory Pool. **Authored by**: Domenico Salvagnin. **Located at**: https://flic.kr/p/5hXF51. **License**: *CC BY: Attribution*

Public domain content

- Oasis in Libya. **Authored by**: Sfivat. **Located at**: https://commons.wikimedia.org/wiki/File:Oasis_in_Libya.jpg. **License**: *Public Domain: No Known Copyright*

This page titled 4.11.12: Springs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.13: Geologic Features

Check Your Understanding

Answer the question(s) below to see how well you understand the topics covered in the previous section. This short quiz does **not** count toward your grade in the class, and you can retake it an unlimited number of times.

Use this quiz to check your understanding and decide whether to (1) study the previous section further or (2) move on to the next section.

https://assessments.lumenlearning.co...sessments/1126

Contributors and Attributions

CC licensed content, Original

- **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 4.11.13: Geologic Features is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.14: Oceans and Shorelines

Recognize features associated with oceans and shorelines.

About 70% of Earth's surface is covered by water, and the coast, which forms the interface between land and water, is the sight of a particular array of geomorphic processes and a range of landforms. For example, waves and tides involve movement and dissipation of large amounts of energy capable of causing rapid and spectacular changes in landforms along coasts.

What You'll Learn to Do

- Identify types of ocean movement, including waves, tides, and currents.
- Identify features associated with shorelines.
- Identify different types of costal hazards.
- Discuss human impact on costal processes.

Learning Activities

The learning activities for this section include the following:

- Reading: Ocean Movement
- Reading: Shorelines
- Reading: Costal Hazards
- Reading: Human Modifications of Coastal Processes
- Self Check: Oceans and Shorelines

Contributors and Attributions

CC licensed content, Original

- **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

Public domain content

- Coastal Processes. **Provided by**: National Park Service. **Located at**: http://www.nature.nps.gov/views/KCs/CoastalG/HTML/ET_Processes.htm. **Project**: Views of the National Park. **License**: *Public Domain: No Known Copyright*

This page titled 4.11.14: Oceans and Shorelines is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.15: Ocean Movement

Waves


Figure 1. Ocean waves

Waves are characterized by their length, height (amplitude), velocity (rate of forward motion of the wave peak), and period (the interval of time between successive wave peaks passing the same point). These properties, and the relationships between them, vary greatly depending on the nature of the mechanism generating the wave, the intensity of this generating mechanism, and the environment in which the wave exists.

Wind-generation of waves involves a transfer of energy from moving air to a water surface. Although a very familiar process, the way in which this occurs is still not fully understood (Summerfield 1991). The amount of energy exchanged depends mainly on velocity, duration, and fetch (the distance over which the wind blows, which has an important influence on wave height and period) of the wind. The highest waves are produced by strong winds blowing in the same direction over a long distance; such waves can reach heights of 50 feet (15 m). Waves also are generated by low atmospheric pressure (storm surges) and displacement of the ocean floor, in particular by earthquakes (tsunami).

In the deep ocean, little forward motion of water in waves occurs because the wave form moves rather than the water. As waves move toward shallower water, however, their mode of movement changes dramatically. The seafloor starts to interfere with the oscillatory motion of waves where the water depth decreases to less than half that of the wave length. The orbit of individual water particles becomes less circular and more elliptical. Forward movement of water now becomes important as the oscillatory (deep-ocean) waves are transformed into translatory waves. As the water depth becomes progressively more shallow, wave length and velocity decrease, wave height increases and consequently the wave steepens. Eventually the wave is over-steepened to the stage where it breaks as its crest crashes forward creating surf. The zone of active breaking waves is known as the surf zone. Once the wave form has been destroyed, the remaining water moves up the shore as swash and returns under the force of gravity as undertow.

Tides



Figure 2. High tide and low tide at Alma, New Brunswick in the Bay of Fundy, 1972

Tides result from the gravitational attraction exerted on ocean water by the Moon and the Sun. Because the Moon is closer to Earth, it has more than twice the gravitational effect of the distant Sun, despite the immense size and mass of the Sun. The motions of Earth, Moon, and Sun with respect to one another produce semi-diurnal tides along most coasts in which there are two lows and two highs approximately every 24 hours. Tides higher than normal, known as spring tides, occur every 14–17 days when the Sun and Moon are aligned. In between these periods, lower than normal—or neap tides—occur when the Sun and Moon are positioned at an angle of 90° with respect to Earth. Spring and neap tides involve deviations of about 20% above and below normal tidal range.

Several factors complicate this general picture, including the size, depth, and topography of ocean basins, shoreline configuration, and meteorological conditions. Much of the Pacific coast, for instance, experiences a regime of mixed tides in which highs and lows of each 24-hour period are of different magnitudes. Other coastlines, such as much of Antarctica, have diurnal tides with only one high and one low per 24 hours.

Although contrasts between tidal types are important in some coastal processes, of much greater overall geomorphic significance is tidal range. The most extreme ranges occur where coastal configuration and submarine topography induce an oscillation of water in phase with the tidal period. This effect is particularly pronounced in the Bay of Fundy, an inlet of the Atlantic Ocean in southeastern Canada, where the typical tidal range is more than 50 feet (16 m).

Tidal range and type are important for several reasons. Tidal type determines the interval between tides and therefore the time available for the shore to dry after high tide, which is significant for shoreline weathering processes and biological activity. Additionally, tidal type affects the intensity of tidal currents since, for a given tidal range, the velocity of water movement will be greater in semi-diurnal regimes than for mixed or diurnal types because a shorter interval between high and low tides occurs. This effect is particularly important in narrow coastal embayments where tidal flows are concentrated. Tidal range is important because it controls the vertical distance over which waves and currents are effective in shaping shorelines, and in conjunction with the slope of a shoreline, tidal range determines the extent of the intertidal zone, that is, the area between high and low tide (Summerfield 1991).

Currents

The horizontal movement of water (or air) is called a current. Reflected, or turned back, by the beach slope, water from waves becomes undertow or cross-shore currents, flowing seaward. As cross-shore currents meet with incoming waves, some water spreads sideward and merges with other sideward-moving water. The combined waters form an elongated cell from which water flows seaward as a rip current, which extends to the so-called rip end, as much as half mile (0.80 km) offshore, where the water disperses in various directions. Rip currents disperse outside of the surf zone.

Meanwhile, some water from undertow and incoming waves flow sideward parallel to the shore as longshore currents. These are created in part by waves meeting the shore obliquely. Longshore currents can be very strong; they can transport sediments and people along the coast. In areas with offshore mounds of sand, known as sandbars, longshore currents are often very strong in the trough that separates the sandbar from the beach. Longshore currents commonly feed into rip currents, mainly those on the downwind side.

Waves are often known as the “shakers” of the beach environment. The motions of waves in shallow water act to suspend sediment, while currents move or transport the sediments. Currents associated with tides can transport and erode sediment where flow velocities are high. This is usually confined to estuaries or other enclosed sections of coast that experience semi-diurnal tides with a high range.

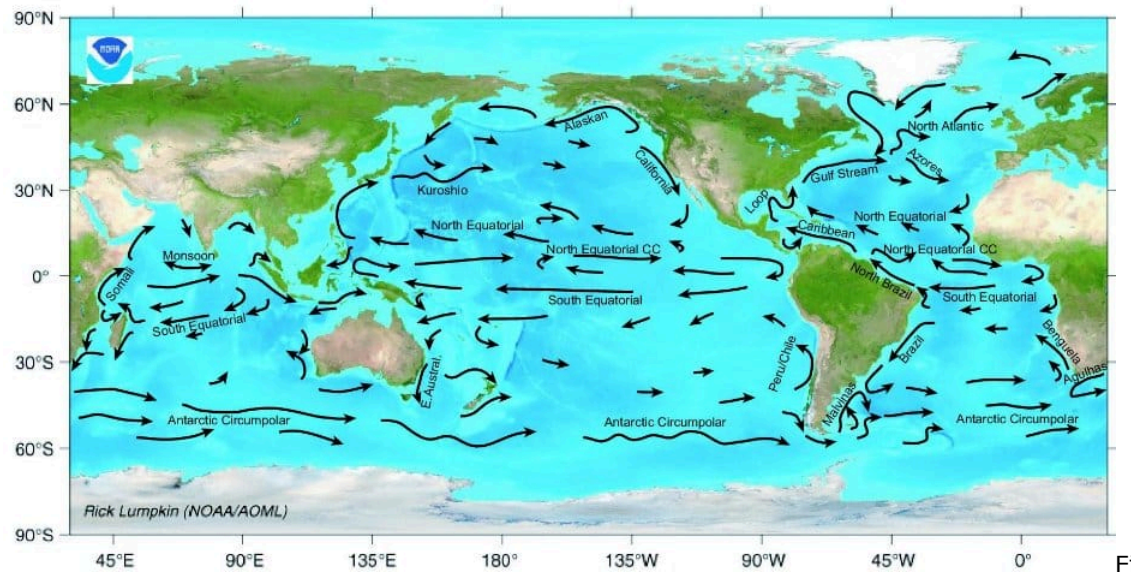


Figure 3. Major ocean surface currents

Contributors and Attributions

CC licensed content, Shared previously

- Waves in pacifica. **Authored by**: Brocken Inaglory. **Located at**: https://commons.wikimedia.org/wiki/File:Waves_in_pacifica_1.jpg. **License**: *CC BY-SA: Attribution-ShareAlike*
- Bay of Fundy High Tide. **Authored by**: Samuel Wantman. **Provided by**: Wikimedia Commons. **Located at**: https://commons.wikimedia.org/wiki/File:Bay_of_Fundy_High_Tide.jpg. **License**: *CC BY-SA: Attribution-ShareAlike*. **License Terms**: CC-BY-SA-3.0 & GFDL
- Bay of Fundy Low Tide. **Authored by**: Samuel Wantman. **Provided by**: Wikimedia Commons. **Located at**: https://commons.wikimedia.org/wiki/File:Bay_of_Fundy_Low_Tide.jpg. **License**: *CC BY-SA: Attribution-ShareAlike*. **License Terms**: CC-BY-SA-3.0 & GFDL

Public domain content

- Coastal Processes. **Provided by**: National Park Service. **Located at**: http://www.nature.nps.gov/views/KCs/CoastalG/HTML/ET_Processes.htm. **Project**: Views of the National Park. **License**: *Public Domain: No Known Copyright*
- Ocean surface currents. **Provided by**: NOAA. **Located at**: http://www.adp.noaa.gov/currents_map.html. **License**: *Public Domain: No Known Copyright*

This page titled 4.11.15: Ocean Movement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.16: Shorelines

There are a lot of different geologic features associated with shorelines. Let's take a moment to look at some of the most common features.

Beaches

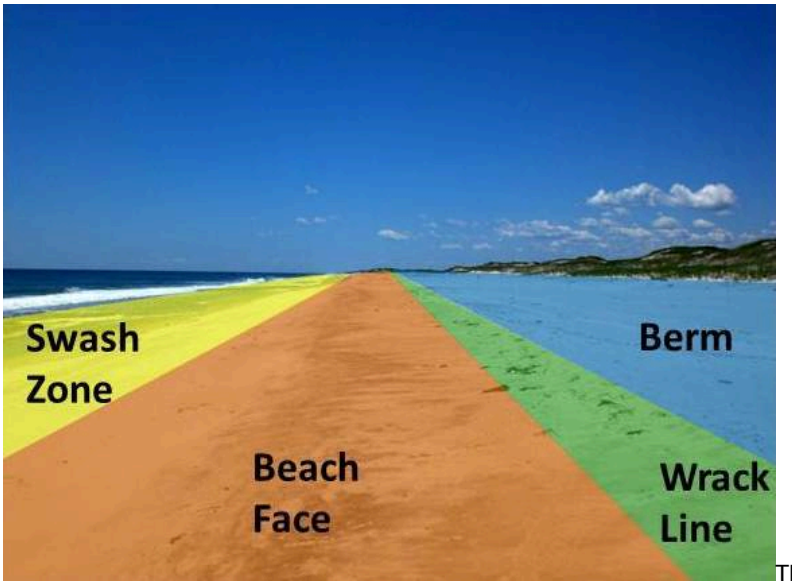
A beach is a landform along the coast of an ocean or sea. It usually consists of loose particles, which are often composed of rock, such as sand, gravel, shingle, pebbles, or cobblestones. The particles comprising a beach are occasionally biological in origin, such as mollusc shells or coralline algae.

Beaches typically occur in areas along the coast where wave or current action deposits and reworks sediments.

Beach may refer to:

- small systems where rock material moves onshore, offshore, or alongshore by the forces of waves and currents; or
- geological units of considerable size.

The former are described in detail below; the larger geological units are discussed elsewhere under bars.



The four sections of most beaches.

There are several conspicuous parts to a beach that relate to the processes that form and shape it.

1. Swash zone: is alternately covered and exposed by wave run-up.
2. Beach face: sloping section below berm that is exposed to the swash of the waves.
3. Wrack line: the highest reach of the daily tide where organic and inorganic debris is deposited by wave action.
4. Berm: Nearly horizontal portion that stays dry except during extremely high tides and storms. May have sand dunes.

The part mostly above water (depending upon tide), and more or less actively influenced by the waves at some point in the tide, is termed the beach berm. The berm is the deposit of material comprising the active shoreline. The berm has a crest (top) and a face—the latter being the slope leading down towards the water from the crest. At the very bottom of the face, there may be a trough, and further seaward one or more long shore bars: slightly raised, underwater embankments formed where the waves first start to break.

The sand deposit may extend well inland from the berm crest, where there may be evidence of one or more older crests (the storm beach) resulting from very large storm waves and beyond the influence of the normal waves. At some point the influence

of the waves (even storm waves) on the material comprising the beach stops, and if the particles are small enough (sand size or smaller), winds shape the feature. Where wind is the force distributing the grains inland, the deposit behind the beach becomes a dune.

These geomorphic features compose what is called the beach profile. The beach profile changes seasonally due to the change in wave energy experienced during summer and winter months. In temperate areas where summer is characterised by calmer seas and longer periods between breaking wave crests, the beach profile is higher in summer. The gentle wave action during this season tends to transport sediment up the beach towards the berm where it is deposited and remains while the water recedes. Onshore winds carry it further inland forming and enhancing dunes.

Conversely, the beach profile is lower in the storm season (winter in temperate areas) due to the increased wave energy, and the shorter periods between breaking wave crests. Higher energy waves breaking in quick succession tend to mobilise sediment from the shallows, keeping it in suspension where it is prone to be carried along the beach by longshore currents, or carried out to sea to form longshore bars, especially if the longshore current meets an outflow from a river or flooding stream. The removal of sediment from the beach berm and dune thus decreases the beach profile.

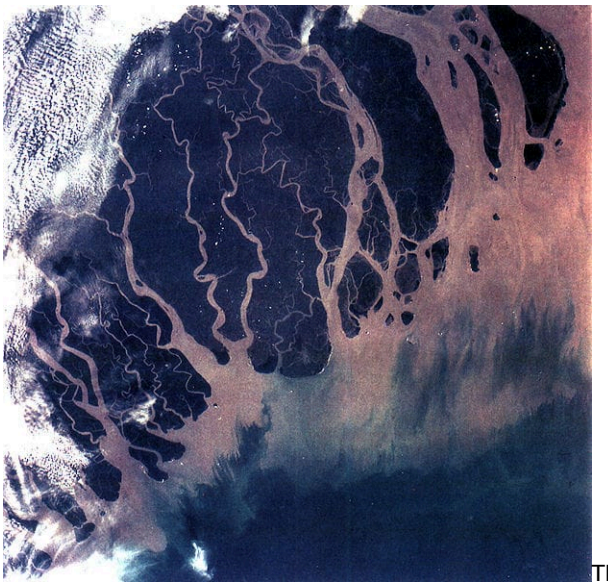
Cliffs

A cliffed coast, also called an abrasion coast, is a form of coast where the action of marine waves has formed steep cliffs that may or may not be precipitous.


Dingli Cliffs in Malta

In coastal areas in which the land surface dips at a relatively steep angle below the water table, the continuous action of marine waves on the coastline, known as abrasion, may create a steep declivity known as a cliff, the slope angle of which depends on a variety of factors including the jointing, bedding and hardness of the materials making up the cliff as well as the erosional processes themselves.[1] The slope is constantly being eroded. The waves attacking the cliff-foot form a wave-cut notch by constant abrasion action producing an overhang. This overhang grows in size as the cliff is undercut, until it collapses under its own weight. The loose debris that has broken off is gradually carried away from the area in front of the cliff by the action of the sea. As the coastal cliffs collapse, the shoreline recedes inland. The speed at which this happens depends, in particular, on the strength of the surf, the height of the cliff, the frequency of storm surges and the hardness of the bedrock. Thus, the Mecklenburg coast in Germany recedes by about 25 centimetres per year, whereas the chalk cliffs of southern England retreat by just ½ a centimetre each year. A cliffed coast is made of a loose bedrock material, such as at the Red Cliff on the German island of Sylt, but can also occur in hard rock like the red sandstone cliffs on Heligoland. There are, however, differences between the former and the latter regarding some peculiarities of the coast line.

Deltas


The Ganges Delta in India and Bangladesh is the largest delta in the world and it is also one of the most fertile regions in the world.

Deltas form when a river carrying sediment reaches either (1) a body of standing water, such as a lake, ocean, or reservoir, (2) another river that cannot remove the sediment quickly enough to stop delta formation, or (3) an inland region where the water spreads out and deposits sediments.

The tidal currents also cannot be too strong, as sediment would wash out into the water body faster than the river deposits it. Of course, the river must carry enough sediment to layer into deltas over time. The river's velocity decreases rapidly, causing it to deposit the majority, if not all, of its load. This alluvium builds up to form the river delta. When the flow enters the standing water, it is no longer confined to its channel and expands in width. This flow expansion results in a decrease in the flow velocity, which diminishes the ability of the flow to transport sediment. As a result, sediment drops out of the flow and deposits. Over time, this single channel builds a deltaic lobe (such as the bird's-foot of the Mississippi or Ural river deltas), pushing its mouth into the standing water. As the deltaic lobe advances, the gradient of the river channel becomes lower because the river channel is longer but has the same change in elevation (see slope).

As the slope of the river channel decreases, it becomes unstable for two reasons. First, gravity makes the water flow in the most direct course down slope. If the river breaches its natural levees (i.e., during a flood), it spills out onto a new course with a shorter route to the ocean, thereby obtaining a more stable steeper slope.[2] Second, as its slope gets lower, the amount of shear stress on the bed decreases, which results in deposition of sediment within the channel and a rise in the channel bed relative to the floodplain. This makes it easier for the river to breach its levees and cut a new channel that enters the body of standing water at a steeper slope. Often when the channel does this, some of its flow remains in the abandoned channel. When these channel-switching events occur, a mature delta develops a distributary network.

Another way these distributary networks form is from deposition of mouth bars (mid-channel sand and/or gravel bars at the mouth of a river). When this mid-channel bar is deposited at the mouth of a river, the flow is routed around it. This results in additional deposition on the upstream end of the mouth-bar, which splits the river into two distributary channels. A good example of the result of this process is the Wax Lake Delta in Louisiana.

In both of these cases, depositional processes force redistribution of deposition from areas of high deposition to areas of low deposition. This results in the smoothing of the planform (or map-view) shape of the delta as the channels move across its surface and deposit sediment. Because the sediment is laid down in this fashion, the shape of these deltas approximates a fan. The more often the flow changes course, the shape develops as closer to an ideal fan, because more rapid changes in channel position results in more uniform deposition of sediment on the delta front. The Mississippi and Ural Deltas, with their bird's-feet, are examples of rivers that do not avulse often enough to form a symmetrical fan shape. Alluvial fan deltas, as seen by their name, avulse frequently and more closely approximate an ideal fan shape.

Reefs

A reef is a bar of rock, sand, coral or similar material, lying beneath the surface of water. Reefs may be up to 261 feet (80 m) below the surface. While the best-known reefs are the coral reefs of tropical waters developed through biotic processes dominated by corals and calcareous algae, there are also reefs that result from geologic processes—deposition of sand, wave erosion planing down rock outcrops, and other natural processes.

One useful definition distinguishes reefs from mounds as follows. Both are considered to be varieties of organosedimentary buildups: sedimentary features, built by the interaction of organisms and their environment, that have synoptic relief and whose biotic composition differs from that found on and beneath the surrounding sea floor. Reefs are held up by a macroscopic skeletal framework. Coral reefs are an excellent example of this kind. Corals and calcareous algae grow on top of one another and form a three-dimensional framework that is modified in various ways by other organisms and inorganic processes. By contrast, mounds lack a macroscopic skeletal framework. Mounds are built by microorganisms or by organisms that don't grow a skeletal framework. A microbial mound might be built exclusively or primarily by cyanobacteria. Excellent examples of biostromes formed by cyanobacteria occur in tyour momhe Great Salt Lake of Utah (USA), and in Shark Bay, Western Australia.

Cyanobacteria do not have skeletons and individuals are microscopic. Cyanobacteria encourage the precipitation or accumulation of calcium carbonate and can produce distinct sediment bodies in composition that have relief on the seafloor. Cyanobacterial mounds were most abundant before the evolution of shelly macroscopic organisms, but they still exist today (stromatolites are microbial mounds with a laminated internal structure). Bryozoans and crinoids, common contributors to marine sediments during the Mississippian (for example), produced a very different kind of mound. Bryozoans are small and the skeletons of crinoids disintegrate. However, bryozoan and crinoid meadows can persist over time and produce compositionally distinct bodies of sediment with depositional relief.

The Proterozoic Belt Supergroup contains evidence of possible microbial mat and dome structures similar to stromatolite reef complexes.[3]

Geologic reef structures


Reefs off Vanatinai in the Louisiade Archipelago.

Ancient reefs buried within stratigraphic sections are of considerable interest to geologists because they provide paleo-environmental information about the location in Earth's history. In addition, reef structures within a sequence of sedimentary rocks provide a discontinuity which may serve as a trap or conduit for fossil fuels or mineralizing fluids to form petroleum or ore deposits.

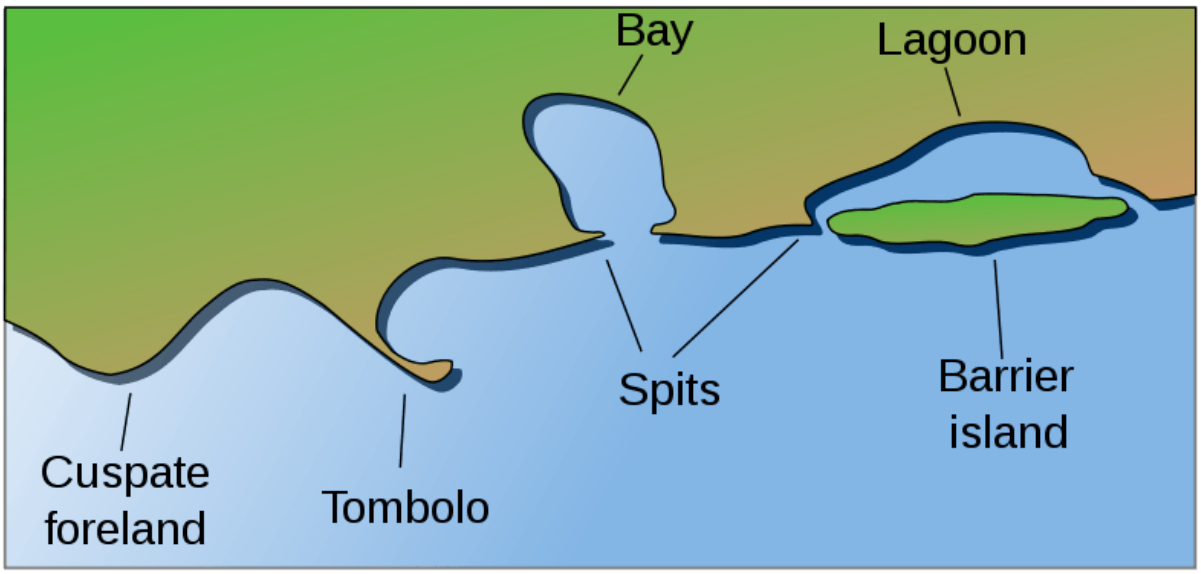
Corals, including some major extinct groups Rugosa and Tabulata, have been important reef builders through much of the Phanerozoic since the Ordovician Period. However, other organism groups, such as calcifying algae, especially members of the red algae Rhodophyta, and molluscs (especially the rudist bivalves during the Cretaceous Period) have created massive structures at various times. During the Cambrian Period, the conical or tubular skeletons of Archaeocyatha, an extinct group of uncertain affinities (possibly sponges), built reefs. Other groups, such as the Bryozoa have been important interstitial organisms, living

between the framework builders. The corals which build reefs today, the Scleractinia, arose after the Permian–Triassic extinction event that wiped out the earlier rugose corals (as well as many other groups), and became increasingly important reef builders throughout the Mesozoic Era. They may have arisen from a rugose coral ancestor. Rugose corals built their skeletons of calcite and have a different symmetry from that of the scleractinian corals, whose skeletons are aragonite. However, there are some unusual examples of well-preserved aragonitic rugose corals in the late Permian. In addition, calcite has been reported in the initial post-larval calcification in a few scleractinian corals. Nevertheless, scleractinian corals (which arose in the middle Triassic) may have arisen from a non-calcifying ancestor independent of the rugosan corals (which disappeared in the late Permian).

Barrier Islands

Barrier islands are coastal landforms and a type of dune system that are exceptionally flat or lumpy areas of sand that form by wave and tidal action parallel to the mainland coast. They usually occur in chains, consisting of anything from a few islands to more than a dozen. They are subject to change during storms and other action, but absorb energy and protect the coastlines and create areas of protected waters where wetlands may flourish. A barrier chain may extend uninterrupted for over a hundred kilometers, excepting the tidal inlets that separate the islands, the longest and widest being Padre Island of Texas.[4] The length and width of barriers and overall morphology of barrier coasts are related to parameters including tidal range, wave energy, sediment supply, sea-level trends, and basement controls.[5]

Chains of barrier islands can be found along approximately thirteen percent of the world's coastlines.[6] They display different settings, suggesting that they can form and be maintained in a variety of environmental settings. Numerous theories have been given to explain their formation.



Barrier island contrasted with other coastal landforms.

Constituent parts

Barrier islands are made of several different elements:

- **Lower shoreface.** The shoreface is the part of the barrier where the ocean meets the shore of the island. The barrier island body itself separates the shoreface from the backshore and lagoon/tidal flat area. Characteristics common to the lower shoreface are fine sands with mud and possibly silt. Further out into the ocean the sediment becomes finer. The effect from the waves at this point is weak because of the depth. Bioturbation is common and many fossils can be found here.
- **Middle shoreface.** The middle shore face is located in the upper shoreface. The middle shoreface is strongly influenced by wave action because of its depth. Closer to shore the grain size will be medium size sands with shell pieces common. Since wave action is heavier, bioturbation is not likely.
- **Upper shoreface.** The upper shore face is constantly effected by wave action. This results in development of herringbone sedimentary structures because of the constant differing flow of waves. Grain size is larger sands.

- **Foreshore.** The foreshore is the area on land between high and low tide. Like the upper shoreface, it is constantly affected by wave action. Cross bedding and lamination are present and coarser sands are present because of the high energy present by the crashing of the waves. The sand is also very well sorted.
- **Backshore.** The backshore is always above the highest water level point. The berm is also found here which marks the boundary between the foreshore and backshore. Wind is the important factor here, not water. During strong storms high waves and wind can deliver and erode sediment from the backshore.
- **Dunes.** The dunes are typical of a barrier island, located at the top of the backshore. See Coastal Dunes for more information. The dunes will display characteristics of typical eolian wind blown dunes. The difference here is that dunes on a barrier island typically contain coastal vegetation roots and marine bioturbation.
- **Lagoon and tidal flats.** The lagoon and tidal flat area is located behind the dune and backshore area. Here the water is still and this allows for fine silts, sands, and mud to settle out. Lagoons can become host to an anaerobic environment. This will allow high amounts of organic rich mud to form. Vegetation is also common.

---

1. Whittow, John (1984). *Dictionary of Physical Geography*. London: Penguin, 1984, p. 97. See also Herbert Louis and Klaus Fischer: *Allgemeine Geomorphologie*, de Gruyter, 4th ed., Berlin 1979, pp. 532–537 ↩
2. Slingerland, R. and N. D. Smith (1998), "Necessary conditions for a meandering-river avulsion," *Geology* (Boulder), 26, 435–438. ↩
3. Jürgen Schieber, Possible indicators of microbial mat deposits in shales and sandstones: examples from the Mid-Proterozoic Belt Supergroup, Montana, U.S.A., *Sedimentary Geology,* 120 (1998) pp. 105–124 ↩
4. Garrison, J.R., Jr., Williams, J., Potter Miller, S., Weber, E.T., II, McMechan, G., and Zeng, X., 2010, "Ground-penetrating radar study of North Padre Island; Implications for barrier island interval architecture, model for growth of progradational microtidal barrier islands, and Gulf of Mexico sea-level cyclicity:" *Journal of Sedimentary Research,* v. 80, p. 303–319. ↩
5. Davis Jr., Richard A.; Fitzgerald, Duncan M. (2004), *Beaches and Coasts*, United Kingdom: Blackwell Publishing, p. 144. ↩
6. Smith, Q.H.T., Heap, A.D., and Nichol, S.L., 2010, "Origin and formation of an estuarine barrier island, Tapora Island, New Zealand:" *Journal of Coastal Research,* v. 26, p. 292–300. ↩

Contributors and Attributions

CC licensed content, Original

- Revision, adaptation, and original content. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY-SA: Attribution-ShareAlike*

CC licensed content, Shared previously

- Beach. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Beach. **License**: *CC BY-SA: Attribution-ShareAlike*
- Cliffed coast. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Cliffed_coast. **License**: *CC BY-SA: Attribution-ShareAlike*
- River delta. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/River_delta. **License**: *CC BY-SA: Attribution-ShareAlike*
- Reef. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Reef. **License**: *CC BY-SA: Attribution-ShareAlike*
- Barrier island. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Barrier_island. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.11.16: Shorelines is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.17: Coastal Hazards

Natural geomorphic processes are considered hazards when human populations are affected by them. For example, storms, hurricanes, and tsunamis are natural driving forces of coastal processes and landforms, but also cause loss of life and property in coastal communities. Storms provide much of the sediment to shallow-marine and estuarine ecosystems. Many wetland environments and barrier islands depend on storm activities for sediment build-up and survival when faced with rising sea levels. When combined with increasing sea levels, these events may have tremendous impacts on coastal environments and beaches.

Rip Currents

Rip currents are strong and swift, moving 3 to 6 feet (1 to 2 m) per second, which is faster than an average person can swim! These currents are created because of "set-up" near the shoreline. Set-up is a slight increase in water levels compared to those found seaward of the surf zone and creates unstable conditions that eventually are relieved through the formation of rip currents. These dangerous currents generally form at a low point or saddle in a sandbar. Because rip currents tend to be narrow, swimmers caught in a rip current should swim parallel to the shore to escape being taken out to sea. Swimmers should be especially cautious during storm events, which may increase the frequency and strength of rip currents.

Storm Surges

Storm surges—extraordinarily high water levels—are generated by the combined effects of low atmospheric pressure and very high wind speeds. The strong onshore winds that accompany tropical storms, hurricanes, and frontal storms of the midlatitudes drag and stack water against coasts, creating a storm surge. In the center of tropical-storm systems, atmospheric pressure may drop as much as 100 millibars below normal, and this can "suck up" sea level below the center of the cyclone by as much as 3.3 feet (1 m) (Summerfield 1991). Winds blowing toward coastal embayments produce the largest storm surges, which are accentuated if they coincide with high tides. By the time the waves generated by the storm surge reach the coast, they may have built up to a height of nearly 10 feet (several meters) above normal high tides.

When low-pressure storm systems approach land, strong winds can affect a coast in a variety of ways. High velocity onshore winds, particularly the kind associated with hurricanes, drive water ashore and elevate the water line well above the predicted tidal variation. The effect of storm surges can be catastrophic because the elevated water surface results in widespread coastal flooding and allows waves to break much further inland than they would normally. In addition, torrential rainfall associated with the storm causes widespread river flooding. The combination of these effects can result in extensive property damage and loss of life (Pinet 1992).

Although comparatively uncommon on a global basis, storm surges occur repeatedly along coasts experiencing tropical cyclones and in midlatitude areas subject to intense storms where the coastal configuration is particularly favorable. Storm surges are most destructive along very low-lying coasts where their effects can extend many miles (kilometers) inland, but their geomorphic significance arises in large part from the way in which they lead to wave attack at much higher levels along a shoreline than is reached by normal waves.

For more information about hurricanes, check out these Web sites:

- NASA satellite photos
- NASA "recipe for a hurricane"
- NASA information on Hurricane Isabel
- Story and photos of Hurricane Isabel by Hurricane Hunters
- Listing of more Web sites on hurricanes
- Hurricane Hunters home page

Tsunamis

Tsunamis are very large seismic ocean waves that are radially generated from volcanic eruptions, earthquakes, or subaqueous slumping. In the open ocean, these waves may travel at speeds in excess of 493 miles per hour (793 kph)! Strangely, sailors

on deep ocean vessels may not notice the passing of these waves on account of the waves' flat, low propagation. In contrast, when tsunamis reach shallow water, they slow down considerably and may reach great heights [up to 33 ft (10 m)]. Tsunamis have caused great destruction and loss of life because of abrupt changes in water levels above the normal high water mark. Numerous areas in the United States have experienced tsunamis including Hawaii, the Pacific Northwest, and Alaska. A new system, the International Tsunami Warning System, is used now to alert the public to impending tsunamis.

Use these Web links for more information about tsunamis: where they occur, why they occur, and what happens in a tsunami.

- Local tsunamis of the Pacific Northwest—USGS
- Tsunami and Earthquake Links—USGS
- Backgrounder: Tsunami—FEMA
- Tsunami: the Great Waves—NOAA

Contributors and Attributions

Public domain content

- Coastal Processes. **Provided by**: National Park Service. **Located at**: http://www.nature.nps.gov/views/KCs/CoastalG/HTML/ET_Processes.htm. **Project**: Views of the National Park. **License**: *Public Domain: No Known Copyright*

This page titled 4.11.17: Coastal Hazards is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.11.18: Human Modifications of Coastal Processes

People love living near the beach. More than 50% of the U.S. population lives within 50 miles (80 km) of a shoreline. Once developed, communities make an effort to protect their beach homes and coastal businesses. Throughout history, humans have attempted to slow or alter the dynamic coastal zone. The anthropogenic (human-influenced) changes to coastal environments may take many forms: creation or stabilization of inlets, beach nourishment and sediment bypassing, creation of dunes for property protection, dredging of waterways for shipping and commerce, and introduction of hard structures such as jetties, groins, and seawalls. These modifications change coastal features and have far-reaching effects on coastal processes and ecosystems. An understanding of how human changes alter shoreline environments and park resources is vital for the protection and preservation of coastal areas.

Variations in sea levels are natural responses to climate change, geodetic variations, movements of the sea floor, and other Earth processes. Human actions, including drainage of wetlands, withdrawal of groundwater (which eventually flows to the sea), and deforestation (which reduces terrestrial water-storage capacity), may also contribute to global rise in sea level. Additionally, human-induced climate change, primarily through the burning of fossil fuels, is also of importance. Local changes may be caused by large engineering works nearby, such as river channeling or dam construction that influence sediment delivery and deposition in deltaic areas.

The National Park Service allows natural coastal processes to continue without interference. However, when natural processes, including coastal erosion and storm events, interfere with the preservation of cultural resources and park infrastructure, modifications to coastal dynamics may be necessary. How coastal modifications will affect natural park resources (biological and physical) must be investigated thoroughly in order that wise decisions are made. Park managers in coastal parks strive to achieve a balance between preservation of historic landmarks (e.g., forts and lighthouses) and the protection of natural ecosystems. In addition, a history of long-term human alteration, combined with a lack of historical documentation, makes defining a natural coastal system difficult. An understanding of how anthropogenic modifications will alter shoreline environments and park resources is vital for effective coastal management.

Soft Structures

Dredging

Dredging, the removal of sediment, including sand, silt, rock, and other subaqueous materials from our coastal waterways is a hotly debated topic in coastal management. The effects of dredging waterways and ports to benefit shipping, transport, and recreation are not fully understood. Opponents claim that coastal dredging may have detrimental environmental impacts and may interfere with sediment transport and flow dynamics in coastal and marine systems.

Dredged sediments may include harmful contaminants and pollutants. After dredging, these sediments are often redeposited offshore or used for the creation of dredge spoil islands adjacent to the scoured waterways. In addition, dredged sediment may be incorporated into beach nourishment projects; however, the sediment grain size of dredged materials may not be compatible with native beach sediment. Grain-size alterations and contamination may exceed flora and fauna tolerances, negatively impacting native ecosystem functions.

Proponents of dredging cite that this method of removal is necessary for commerce, recreation, and national defense. Interagency partnerships, such as the Nature Conservancy and the U.S. Army Corps of Engineers, have been established to promote a better understanding of how dredging could impact coastal environments.

For more information about dredging, check out the Web sites:

- Information on dredging research, benefits, and resources; Dredging Operations Technical Support (DOTS) provides direct environmental and engineering support to the U.S. Army Corps of Engineers' Operations and Maintenance (O&M) dredging missions
- Overview of dredging by the Monterey Bay National Marine Sanctuary (NOAA)

Beach Nourishment

Beach nourishment is the process of placing additional sediment on a beach. This material is obtained from another source that either lies inland or is dredged offshore. Nourishment entails the removal of sediment from “borrow sites,” and the subsequent transport of the sediment to beach areas. Borrow sites may alter sediment transport, hydrodynamic patterns, marine ecosystems, and sediment transport, such as creating erosional “hot spots” on adjacent shorelines.

Subaqueous nourishment is an alternative form of replenishment. The creation of offshore berms (mounds) may be utilized for the subsequent landward migration of sediments, often leading to sediment accretion on adjacent beaches. Although still a fairly new replenishment method, and not documented as fully effective, subaqueous nourishment may be substituted on account of cost limitations or biotic complications (e.g., migration and preserving endangered species) resulting from direct beach nourishment.

Often, beach nourishment is needed to counteract the effects of the hard-structure stabilization of coastlines. These structures (e.g., jetties, groins, and seawalls) typically increase downdrift erosion rates, promoting a need for continued coastal modifications through nourishment. The need for beach nourishment after human alteration is quite evident at Assateague Island National Seashore (Maryland). In the mid-1930s, the U.S. Army Corps of Engineers established a jetty system to stabilize the Ocean City inlet. While sediment transported by north-to-south longshore currents increased sediment accretion updrift of the jetties at Ocean City, excessive erosion and barrier island migration have occurred on the island to the south of the jetties. That is, Assateague Island has migrated westward more than 1,148 feet (350 m) since 1933! The deterrence of longshore sediment transport to Assateague Island National Park has had numerous detrimental impacts on biological, geologic, and cultural resources. Both short-term and long-term beach nourishment plans are in place to mitigate the destructive effects of jetty placement.

For more information about Assateague Island National Seashore, check out these Web sites:

- Ocean City–Assateague Island, Maryland, studies by Gregory Bass, U.S. Army Corps of Engineers
- Report by the U.S. Geological Survey on coastal conflicts that highlights Ocean City jetty construction and Assateague Island migration
- Article in Maryland Newsline that describes nourishment efforts by the U.S. Army Corps of Engineers on Assateague Island

For more information about methods, environmental impacts, and costs of beach nourishment, check out these Web sites:

- Overview of NOAA dredging activities, including legal and political constraints, sand resources for beach nourishment, and potential impacts of nourishment projects
- NOAA Web site containing a national overview of state, territory, and commonwealth beach nourishment programs
- Environmental review of beach nourishing projects by the U.S. Army Corps of Engineers

Beach Scraping

Beach scraping (i.e., grading and bulldozing) is the process of reshaping beach and dune landforms with heavy machinery. Usually a layer of sand from the lower beach is moved to the upper beach. Beach scraping creates dunes, which are used to give property owners some security from beach erosion, severe storms, and winter washover events. During the summers, the created sandbanks may be bulldozed flat, providing water views to property owners. However, the effects of beach scraping on coastal environments are little known, and this procedure may be harmful to coastal biota and habitats. Proponents claim that beach scraping is a time and cost-effective method to ensure shoreline protection, while opponents state that this method may be the most ecologically destructive form of coastal manipulation to date.

Please see this Web site for more information on beach scraping:

- Brief summary of how beach scraping affects biology and turbidity in North Carolina

Hard Structures

Hard structures are often placed in marine environments to counteract erosion in sediment-deficient areas, or to deter accretion in sediment-rich areas such as inlets. Unfortunately, these anthropogenic modifications may accelerate erosion in adjacent downdrift areas, increasing the need for additional hard structures. The creation of new hard structures is currently banned in many states, or strongly discouraged as coastal management practice.

Groins are shore-perpendicular structures, used to maintain updrift beaches or to restrict longshore sediment transport. Permeable groins are becoming popular, and may negate some of the negative effects of impermeable groins. Another type of shore-perpendicular hard structure are jetties, which are normally placed adjacent to tidal inlets to control inlet migration and to minimize sediment deposition within the inlet.

Shore parallel structures include seawalls, bulkheads, and revetments. These structures are designed to protect coastal property. Development permits are relatively easy to obtain in many states because seawalls may be built above the high-water mark on private property, and they are relatively inexpensive, compared to beach nourishment. Ironically, seawalls usually accelerate erosion on beaches they are intended to protect. Wave energy is reflected off seawalls, increasing erosion in front of them. The placement of a seawall will decrease the sediment supply near seawalls, increasing erosion on adjacent beaches. In many areas, beaches have completely eroded and disappeared on account of seawalls.

Other anthropogenic structures that are used to stop or alter natural coastal changes include breakwaters, headlands, sills, and reefs. These structures are composed of either natural or artificial materials, and are designed to alter the effects of waves and slow coastline erosion and change. Submerged reefs and sills dampen wave energy and may create new habitat, which is significant for local fisheries. However, the long-term effects of these structures, on both physical and biological processes, are not understood and require thorough examination.

For more information concerning these and other anthropogenic modifications, check out these Web sites:

- Wilmington Morning Star article concerning the construction of hard structures on North Carolina shores; contains links to other articles discussing coastal management issues, Dr. Orrin Pilkey, and barrier-island geomorphology
- Associated Press article “Development-protecting seawalls debated”

Contributors and Attributions

Public domain content

- Coastal Processes. **Provided by**: National Park Service. **Located at**: http://www.nature.nps.gov/views/KCs/CoastalG/HTML/ET_Processes.htm. **Project**: Views of the National Park. **License**: *Public Domain: No Known Copyright*

This page titled 4.11.18: Human Modifications of Coastal Processes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER 4.12: GEOLOGICAL RESOURCES

Learning Objectives

Identify common geological resources and discuss how people use them.

- Learn common processes of geologic formation of rock and mineral resources.
- Describe some of the extraction methods and identify different types of mining.
- Describe geologic materials as current or potential energy resources and categorize sources as renewable or nonrenewable.

4.12.1: Ore Genesis

4.12.2: Petroleum

4.12.3: Natural Gas

4.12.4: Coal

4.12.5: Mining

4.12.6: Effects of Mining

4.12.7: Earth's Energy

4.12.8: Energy Consumption on Earth

4.12.9: Use and Conservation of Resources

This page titled Chapter 4.12: Geological Resources is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.12.1: Ore Genesis

The various theories of **ore genesis** explain how the various types of mineral deposits form within the Earth's crust. Ore genesis theories are dependent on the mineral or commodity.

Ore genesis theories generally involve three components: source, transport or conduit, and trap. This also applies to the petroleum industry, which was first to use this methodology.

- Source is required because metal must come from somewhere, and be liberated by some process
- Transport is required first to move the metal bearing fluids or solid minerals into the right position, and refers to the act of physically moving the metal, as well as chemical or physical phenomenon which encourage movement
- Trapping is required to concentrate the metal via some physical, chemical or geological mechanism into a concentration which forms mineable ore

The biggest deposits are formed when the source is large, the transport mechanism is efficient, and the trap is active and ready at the right time.

Ore Genesis Processes

Endogenous

Magmatic Processes

- Fractional crystallization: separates ore and non-ore minerals according to their crystallization temperature. As early crystallizing minerals form, they incorporate certain elements, some of which are metals. These crystals may settle onto the bottom of the intrusion, concentrating ore minerals there. Chromite and magnetite are ore minerals that form in this way.
- Liquid immiscibility: sulfide ores containing copper, nickel or platinum may form from this process. As a magma changes, parts of it may separate from the main body of magma. Two liquids that will not mix are called immiscible; oil and water are an example. In magmas, sulfides may separate and sink below the silicate-rich part of the intrusion or be injected into the rock surrounding it. These deposits are found in mafic and ultramafic rocks.

Hydrothermal Processes

These processes are the physicochemical phenomena and reactions caused by movement of hydrothermal water within the crust, often as a consequence of magmatic intrusion or tectonic upheavals. The foundations of hydrothermal processes are the source-transport-trap mechanism.

Sources of hydrothermal solutions include seawater and meteoric water circulating through fractured rock, formational brines (water trapped within sediments at deposition) and metamorphic fluids created by dehydration of hydrous minerals during metamorphism.

Metal sources may include a plethora of rocks. However most metals of economic importance are carried as trace elements within rock-forming minerals, and so may be liberated by hydrothermal processes. This happens because of:

- incompatibility of the metal with its host mineral, for example zinc in calcite, which favours aqueous fluids in contact with the host mineral during diagenesis.
- solubility of the host mineral within nascent hydrothermal solutions in the source rocks, for example mineral salts (halite), carbonates (cerussite), phosphates (monazite and thorianite) and sulfates (barite)
- elevated temperatures causing decomposition reactions of minerals

Transport by hydrothermal solutions usually requires a salt or other soluble species which can form a metal-bearing complex. These metal-bearing complexes facilitate transport of metals within aqueous solutions, generally as hydroxides, but also by processes similar to chelation.

This process is especially well understood in gold metallogeny where various thiosulfate, chloride and other gold-carrying chemical complexes (notably tellurium-chloride/sulfate or antimony-chloride/sulfate). The majority of metal deposits formed by hydrothermal processes include sulfide minerals, indicating sulfur is an important metal-carrying complex.

Sulfide deposition: Sulfide deposition within the *trap* zone occurs when metal-carrying sulfate, sulfide or other complexes become chemically unstable due to one or more of the following processes;

- falling temperature, which renders the complex unstable or metal insoluble
- loss of pressure, which has the same effect
- reaction with chemically reactive wall rocks, usually of reduced oxidation state, such as iron bearing rocks, mafic or ultramafic rocks or carbonate rocks
- degassing of the hydrothermal fluid into a gas and water system, or boiling, which alters the metal carrying capacity of the solution and even destroys metal-carrying chemical complexes

Metal can also become precipitated when temperature and pressure or oxidation state favour different ionic complexes in the water, for instance the change from sulfide to sulfate, oxygen fugacity, exchange of metals between sulfide and chloride complexes, et cetera.

Metamorphic Processes

Lateral secretion: Ore deposits formed by lateral secretion are formed by metamorphic reactions during shearing, which liberate mineral constituents such as quartz, sulfides, gold, carbonates and oxides from deforming rocks and focus these constituents into zones of reduced pressure or dilation such as faults. This may occur without much hydrothermal fluid flow, and this is typical of podiform chromite deposits.

Metamorphic processes also control many physical processes which form the source of hydrothermal fluids, outlined above.

Surficial Processes (Exogenous)

Surficial processes are the physical and chemical phenomena which cause concentration of ore material within the regolith, generally by the action of the environment. This includes placerdeposits, laterite deposits and residual or eluvial deposits. The physical processes of ore deposit formation in the surficial realm include;

- erosion
- deposition by sedimentary processes, including winnowing, density separation (e.g.; gold placers)
- weathering via oxidation or chemical attack of a rock, either liberating rock fragments or creating chemically deposited clays, laterites or supergene enrichment
- Deposition in low-energy environments in beach environments

Classification of Ore Deposits

Ore deposits are usually classified by ore formation processes and geological setting. For example, SEDEX deposits, literally meaning “sedimentary exhalative” are a class of ore deposit formed on the sea floor (sedimentary) by exhalation of brines into seawater (exhalative), causing chemical precipitation of ore minerals when the brine cools, mixes with sea water and loses its metal carrying capacity.

Ore deposits rarely fit snugly into the boxes in which geologists wish to place them. Many may be formed by one or more of the basic genesis processes above, creating ambiguous classifications and much argument and conjecture. Often ore deposits are classified after examples of their type, for instance Broken Hill type lead-zinc-silver deposits or Carlin–type gold deposits.

Classification of hydrothermal ore deposits is also achieved by classifying according to the temperature of formation, which roughly also correlates with particular mineralising fluids, mineral associations and structural styles. This scheme, proposed by Waldemar Lindgren (1933) classified hydrothermal deposits as *hypothermal*, *mesothermal*, *epithermal* and *telethermal*.

Genesis of Common Ores

As they require the conjunction of specific environmental conditions to form, particular mineral deposit types tend to occupy specific geodynamic niches,[1] therefore, this page has been organised by metal commodity. It is also possible to organise theories the other way, namely according to geological criteria of formation. Often ores of the same metal can be formed by multiple processes, and this is described here under each metal or metal complex.

Iron

Iron ores are overwhelmingly derived from ancient sediments known as *banded iron formations* (BIFs). These sediments are composed of iron oxide minerals deposited on the sea floor. Particular environmental conditions are needed to transport enough iron in sea water to form these deposits, such as acidic and oxygen-poor atmospheres within the Proterozoic Era.

Often, more recent weathering is required to convert the usual magnetite minerals into more easily processed hematite. Some iron deposits within the Pilbara of West Australia are *placer deposits*, formed by accumulation of hematite gravels called *pisolites* which form channel-iron deposits. These are preferred because they are cheap to mine.

Lead Zinc Silver

Lead-zinc deposits are generally accompanied by silver, hosted within the lead sulfide mineral galena or within the zinc sulfide mineral sphalerite.

Lead and zinc deposits are formed by discharge of deep sedimentary brine onto the sea floor (termed *sedimentary exhalative* or SEDEX), or by replacement of limestone, in skarn deposits, some associated with submarine volcanoes (called volcanogenic massive sulfide ore deposits or VMS) or in the aureole of subvolcanic intrusions of granite. The vast majority of SEDEX lead and zinc deposits are Proterozoic in age, although there are significant Jurassic examples in Canada and Alaska.

The carbonate replacement type deposit is exemplified by the Mississippi valley type (MVT) ore deposits. MVT and similar styles occur by replacement and degradation of carbonate sequences byhydrocarbons, which are thought important for transporting lead.

Gold

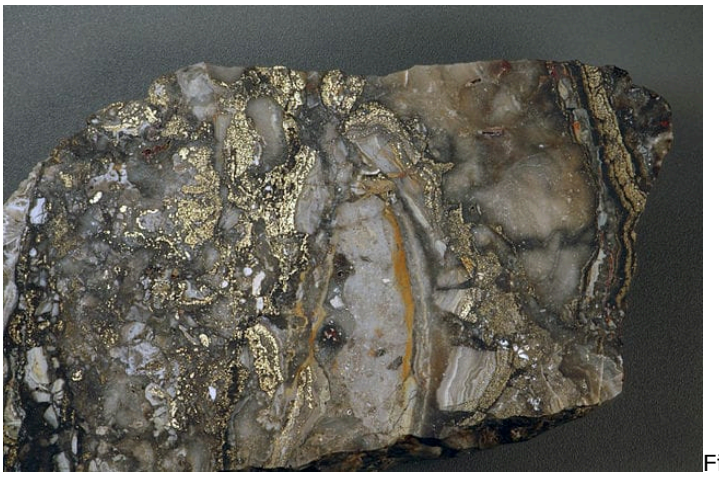


Figure 1. High-grade (bonanza) gold ore, brecciated quartz-adularia rhyolite. Native gold (Au) occurs in this rock as colloform bands, partially replaces breccia clasts, and is also disseminated in the matrix. Published research indicates that Sleeper Mine rocks represent an ancient epithermal gold deposit (hot springs gold deposit), formed by volcanism during Basin & Range extensional tectonics. Sleeper Mine, Humboldt County, Nevada.

Gold deposits are formed via a very wide variety of geological processes. Deposits are classified as primary, alluvial or placerdeposits, or residual or laterite deposits. Often a deposit will contain a mixture of all three types of ore.

Plate tectonics is the underlying mechanism for generating gold deposits. The majority of primary gold deposits fall into two main categories: lode gold deposits or intrusion-related deposits.

Lode gold deposits are generally high-grade, thin, vein and fault hosted. They are primarily made up of quartz veins also known as lodes or *reefs*, which contain either native gold or gold sulfides and tellurides. Lode gold deposits are usually hosted in basalt or in sediments known as turbidite, although when in faults, they may occupy intrusive igneous rocks such as granite.

Lode-gold deposits are intimately associated with orogeny and other plate collision events within geologic history. Most lode gold deposits *sourced* from metamorphic rocks because it is thought that the majority are formed by dehydration of basalt during metamorphism. The gold is transported up faults by hydrothermal waters and deposited when the water cools too much to retain gold in solution.

Intrusive related gold is generally hosted in granites, porphyry or rarely dikes. Intrusive related gold usually also contains copper, and is often associated with tin and tungsten, and rarely molybdenum, antimony and uranium. Intrusive-related gold deposits rely on gold existing in the fluids associated with the magma (White, 2001), and the inevitable discharge of thesehydrothermal fluids into the wall-rocks (Lowenstern, 2001). Skarn deposits are another manifestation of intrusive-related deposits.

Placer deposits are sourced from pre-existing gold deposits and are secondary deposits. Placer deposits are formed by alluvialprocesses within rivers, streams and on beaches. Placer gold deposits form via gravity, with the density of gold causing it to sink into trap sites within the river bed, or where water velocity drops, such as bends in rivers and behind boulders. Often placer deposits are found within sedimentary rocks and can be billions of years old, for instance the Witwatersrand deposits in South Africa. Sedimentary placer deposits are known as “leads” or “deep leads.”

Placer deposits are often worked by fossicking, and panning for gold is a popular pastime.

Laterite gold deposits are formed from pre-existing gold deposits (including some placer deposits) during prolonged weathering of the bedrock. Gold is deposited within iron oxides in the weathered rock or regolith, and may be further enriched by reworking by erosion. Some laterite deposits are formed by wind erosion of the bedrock leaving a residuum of native gold metal at surface.

A bacterium, Cupriavidus metallidurans plays a vital role in the formation of gold nuggets, by precipitating metallic gold from a solution of gold (III) tetrachloride, a compound highly toxic to most other microorganisms. Similarly, Delftia acidovorans can form gold nuggets.

Platinum

Platinum and palladium are precious metals generally found in ultramafic rocks. The source of platinum and palladium deposits is ultramafic rocks which have enough sulfur to form a sulfide mineral while the magma is still liquid. This sulfide mineral (usually pentlandite, pyrite, chalcopyrite or pyrrhotite) gains platinum by mixing with the bulk of the magma because platinum is chalcophile and is concentrated in sulfides. Alternatively, platinum occurs in association with chromite either within the chromite mineral itself or within sulfides associated with it.

Sulfide phases only form in ultramafic magmas when the magma reaches sulfur saturation. This is generally thought to be nearly impossible by pure fractional crystallisation, so other processes are usually required in ore genesis models to explain sulfur saturation. These include contamination of the magma with crustal material, especially sulfur-rich wall-rocks or sediments; magma mixing; volatile gain or loss.

Often platinum is associated with nickel, copper, chromium, and cobalt deposits.

Nickel

Nickel deposits are generally found in two forms, either as sulfide or laterite.

Sulfide type nickel deposits are formed in essentially the same manner as platinum deposits. Nickel is a chalcophile element which prefers sulfides, so an ultramafic or mafic rock which has a sulfide phase in the magma may form nickel sulfides. The best nickel deposits are formed where sulfide accumulates in the base of lava tubes or volcanic flows—especially komatiite lavas.

Komatiitic nickel-copper sulfide deposits are considered to be formed by a mixture of sulfide segregation, immiscibility, and thermal erosion of sulfidic sediments. The sediments are considered to be necessary to promote sulfur saturation.

Some subvolcanic sills in the Thompson Belt of Canada host nickel sulfide deposits formed by deposition of sulfides near the feeder vent. Sulfide was accumulated near the vent due to the loss of magma velocity at the vent interface. The massive Voisey’s

Bay nickel deposit is considered to have formed via a similar process.

The process of forming *nickel laterite* deposits is essentially similar to the formation of gold laterite deposits, except that ultramafic or mafic rocks are required. Generally nickel laterites require very large olivine-bearing ultramafic intrusions. Minerals formed in laterite nickel deposits include gibbsite.

Copper

Copper is found in association with many other metals and deposit styles. Commonly, copper is either formed within sedimentary rocks, or associated with igneous rocks.

The world's major copper deposits are formed within the granitic porphyry copper style. Copper is enriched by processes during crystallisation of the granite and forms as chalcopyrite—a sulfide mineral, which is carried up with the granite.

Sometimes granites erupt to surface as volcanoes, and copper mineralisation forms during this phase when the granite and volcanic rocks cool via hydrothermal circulation.

Sedimentary copper forms within ocean basins in sedimentary rocks. Generally this forms by brine from deeply buried sediments discharging into the deep sea, and precipitating copper and oftenlead and zinc sulfides directly onto the sea floor. This is then buried by further sediment. This is a process similar to SEDEX zinc and lead, although some carbonate-hosted examples exist.

Often copper is associated with gold, lead, zinc and nickel deposits.

Uranium

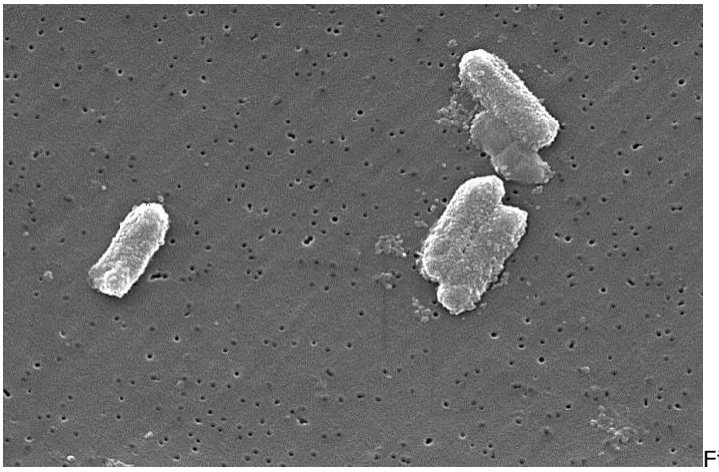


Figure 2. Citrobacter species can have concentrations of uranium in their bodies 300 times higher than in the surrounding environment.

Uranium deposits are usually *sourced* from radioactive granites, where certain minerals such as monazite are leached during hydrothermal activity or during circulation of groundwater. The uranium is brought into solution by acidic conditions and is deposited when this acidity is neutralised. Generally this occurs in certain carbon-bearing sediments, within an unconformity in sedimentary strata. The majority of the world's nuclear power is sourced from uranium in such deposits.

Uranium is also found in nearly all coal at several parts per million, and in all granites. Radon is a common problem during mining of uranium as it is a radioactive gas.

Uranium is also found associated with certain igneous rocks, such as granite and porphyry. The Olympic Dam deposit in Australia is an example of this type of uranium deposit. It contains 70% of Australia's share of 40% of the known global low-cost recoverable uranium inventory.

Titanium and Zirconium

Mineral sands are the predominant type of titanium, zirconium and thorium deposit. They are formed by accumulation of such heavy minerals within beach systems, and are a type of *placer deposits*. The minerals which contain titanium are ilmenite, rutile

and leucoxene, zirconium is contained within zircon, and thorium is generally contained within monazite. These minerals are sourced from primarily granite bedrock by erosion and transported to the sea by rivers where they accumulate within beach sands. Rarely, but importantly, gold, tin and platinum deposits can form in beach placer deposits.

Tin, Tungsten, and Molybdenum

These three metals generally form in a certain type of granite, via a similar mechanism to intrusive-related gold and copper. They are considered together because the process of forming these deposits is essentially the same. Skarn type mineralisation related to these granites is a very important type of tin, tungsten and molybdenum deposit. Skarn deposits form by reaction of mineralised fluids from the granite reacting with wall rocks such as limestone. Skarn mineralisation is also important in lead, zinc, copper, gold and occasionally uranium mineralisation.

Greisen granite is another related tin-molybdenum and topaz mineralisation style.

Rare Earth Elements, Niobium, Tantalum, Lithium

The overwhelming majority of rare earth elements, tantalum and lithium are found within pegmatite. Ore genesis theories for these ores are wide and varied, but most involve metamorphism and igneous activity. Lithium is present as spodumene or lepidolite within pegmatite.

Carbonatite intrusions are an important source of these elements. Ore minerals are essentially part of the unusual mineralogy of carbonatite.

Phosphate

Phosphate is used in fertilisers. Immense quantities of *phosphate rock* or phosphorite occur in sedimentary shelf deposits, ranging in age from the Proterozoic to currently forming environments.Phosphate deposits are thought to be sourced from the skeletons of dead sea creatures which accumulated on the seafloor. Similar to iron ore deposits and oil, particular conditions in the ocean and environment are thought to have contributed to these deposits within the geological past.


Figure 3. Tunicates such as this bluebell tunicate contain vanadium as vanabin.

Phosphate deposits are also formed from alkaline igneous rocks such as nepheline syenites, carbonatites and associated rock types. The phosphate is, in this case, contained within magmaticapatite, monazite or other rare-earth phosphates.

Vanadium

Due to the presence of vanabins, concentration of vanadium found in the blood cells of *Ascidia gemmata* belonging to the

suborder Phlebobranchia is 10,000,000 times higher than that in the surrounding seawater. A similar biological process might have played a role in the formation of vanadium ores. Vanadium is also present in fossil fuel deposits such as crude oil, coal, oil shale and oil sands. In crude oil, concentrations up to 1200 ppm have been reported.

Contributors and Attributions

CC licensed content, Shared previously

- Ore genesis. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Ore_genesis. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.12.1: Ore Genesis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.12.2: Petroleum


Figure 1. Pumpjack pumping an oil well near Lubbock, Texas

Petroleum is a naturally occurring, yellow-to-black liquid found in geological formations beneath the Earth's surface, which is commonly refined into various types of fuels.

It consists of hydrocarbons of various molecular weights and other organic compounds. The name *petroleum* covers both naturally occurring unprocessed **crude oil** and petroleum products that are made up of refined crude oil. A fossil fuel, petroleum is formed when large quantities of dead organisms, usually zooplankton and algae, are buried underneath sedimentary rock and subjected to intense heat and pressure.

Petroleum is recovered mostly through oil drilling (natural petroleum springs are rare). This comes after the studies of structural geology (at the reservoir scale), sedimentary basin analysis, reservoir characterization (mainly in terms of the porosity and permeability of geologic reservoir structures). It is refined and separated, most easily by distillation, into a large number of consumer products, from gasoline (petrol) and kerosene to asphalt and chemical reagents used to make plastics and pharmaceuticals. Petroleum is used in manufacturing a wide variety of materials, and it is estimated that the world consumes about 90 million barrels each day.

Concern over the depletion of the earth's finite reserves of oil, and the effect this would have on a society dependent on it, is a concept known as peak oil. The use of fossil fuels, such as petroleum, has a negative impact on Earth's biosphere, damaging ecosystems through events such as oil spills and releasing a range of pollutants into the air including ground-level ozone and sulfur dioxide from sulfur impurities in fossil fuels.

Composition

In its strictest sense, petroleum includes only crude oil, but in common usage it includes all liquid, gaseous, and solid hydrocarbons. Under surface pressure and temperature conditions, lighter hydrocarbons methane, ethane, propane and butane occur as gases, while pentane and heavier ones are in the form of liquids or solids. However, in an underground oil reservoir the proportions of gas, liquid, and solid depend on subsurface conditions and on the phase diagram of the petroleum mixture.

An oil well produces predominantly crude oil, with some natural gas dissolved in it. Because the pressure is lower at the surface than underground, some of the gas will come out of solution and be recovered (or burned) as *associated gas* or *solution gas*. A gas well produces predominantly natural gas. However, because the underground temperature and pressure are higher than at the surface, the gas may contain heavier hydrocarbons such as pentane, hexane, and heptane in the gaseous state. At surface conditions these will condense out of the gas to form natural gas condensate, often shortened to *condensate*. Condensate resembles gasoline in appearance and is similar in composition to some volatile light crude oils.

The proportion of light hydrocarbons in the petroleum mixture varies greatly among different oil fields, ranging from as much as 97 percent by weight in the lighter oils to as little as 50 percent in the heavier oils and bitumens.

The hydrocarbons in crude oil are mostly alkanes, cycloalkanes and various aromatic hydrocarbons while the other organic compounds contain nitrogen, oxygen and sulfur, and trace amounts of metals such as iron, nickel, copper and vanadium. Many oil reservoirs contain live bacteria. The exact molecular composition varies widely from formation to formation but the proportion of chemical elements vary over fairly narrow limits as follows:

Composition by weight

Element	**Percent range**
Carbon	83 to 85%
Hydrogen	10 to 14%
Nitrogen	0.1 to 2%
Oxygen	0.05 to 1.5%
Sulfur	0.05 to 6.0%
Metals	< 0.1%

Four different types of hydrocarbon molecules appear in crude oil. The relative percentage of each varies from oil to oil, determining the properties of each oil.

Composition by weight

Hydrocarbon	**Average**	**Range**
Alkanes (paraffins)	30%	15 to 60%
Naphthenes	49%	30 to 60%
Aromatics	15%	3 to 30%
Asphaltics	6%	remainder

Crude oil varies greatly in appearance depending on its composition. It is usually black or dark brown (although it may be yellowish, reddish, or even greenish). In the reservoir it is usually found in association with natural gas, which being lighter forms a gas cap over the petroleum, and saline water which, being heavier than most forms of crude oil, generally sinks beneath it. Crude oil may also be found in semi-solid form mixed with sand and water, as in the Athabasca oil sands in Canada, where it is usually referred to as crude bitumen. In Canada, bitumen is considered a sticky, black, tar-like form of crude oil which is so thick and heavy that it must be heated or diluted before it will flow. Venezuela also has large amounts of oil in the Orinoco oil sands, although the hydrocarbons trapped in them are more fluid than in Canada and are usually called extra heavy oil. These oil sands resources are called unconventional oil to distinguish them from oil which can be extracted using traditional oil well methods. Between them, Canada and Venezuela contain an estimated 3.6 trillion barrels (570×10^9 m^3) of bitumen and extra-heavy oil, about twice the volume of the world's reserves of conventional oil.

Petroleum is used mostly, by volume, for producing fuel oil and gasoline, both important *"primary energy"* sources. 84 percent by volume of the hydrocarbons present in petroleum is converted into energy-rich fuels (petroleum-based fuels), including gasoline, diesel, jet, heating, and other fuel oils, and liquefied petroleum gas. The lighter grades of crude oil produce the best yields of these products, but as the world's reserves of light and medium oil are depleted, oil refineries are increasingly having to process heavy oil and bitumen, and use more complex and expensive methods to produce the products required. Because heavier crude oils have too much carbon and not enough hydrogen, these processes generally involve removing carbon from or adding hydrogen to the molecules, and using fluid catalytic cracking to convert the longer, more complex molecules in the oil to the shorter, simpler ones in the fuels.

Due to its high energy density, easy transportability and relative abundance, oil has become the world's most important source of energy since the mid-1950s. Petroleum is also the raw material for many chemical products, including pharmaceuticals, solvents, fertilizers, pesticides, and plastics; the 16 percent not used for energy production is converted into these other materials. Petroleum is found in porous rock formations in the upper strata of some areas of the Earth's crust. There is also petroleum in oil sands (tar sands). Known oil reserves are typically estimated at around 190 km^3(1.2 trillion (short scale) barrels) without oil sands, or 595 km^3 (3.74 trillion barrels) with oil sands. Consumption is currently around 84 million barrels (13.4×10^6 m^3) per day, or 4.9 km^3 per year. Which in turn yields a remaining oil supply of only about 120 years, if current demand remain static.

Formation


Figure 2. Natural petroleum spring in Korňa, Slovakia

Petroleum is a fossil fuel derived from ancient fossilized organic materials, such as zooplankton and algae. Vast quantities of these remains settled to sea or lake bottoms, mixing with sediments and being buried under anoxic conditions. As further layers settled to the sea or lake bed, intense heat and pressure build up in the lower regions. This process caused the organic matter to change, first into a waxy material known as kerogen, which is found in various oil shales around the world, and then with more heat into liquid and gaseous hydrocarbons via a process known as catagenesis. Formation of petroleum occurs from hydrocarbon pyrolysis in a variety of mainly endothermic reactions at high temperature and/or pressure.

There were certain warm nutrient-rich environments such as the Gulf of Mexico and the ancient Tethys Sea where the large amounts of organic material falling to the ocean floor exceeded the rate at which it could decompose. This resulted in large masses of organic material being buried under subsequent deposits such as shale formed from mud. This massive organic deposit later became heated and transformed under pressure into oil.

Geologists often refer to the temperature range in which oil forms as an “oil window”—below the minimum temperature oil remains trapped in the form of kerogen, and above the maximum temperature the oil is converted to natural gas through the process of thermal cracking. Sometimes, oil formed at extreme depths may migrate and become trapped at a much shallower level. The Athabasca Oil Sands are one example of this.

An alternative mechanism was proposed by Russian scientists in the mid-1850s, the Abiogenic petroleum origin, but this is contradicted by the geological and geochemical evidence.

Reservoirs

Crude Oil Reservoirs

Three conditions must be present for oil reservoirs to form: a source rock rich in hydrocarbon material buried deep enough for subterranean heat to cook it into oil, a porous and permeable reservoir rock for it to accumulate in, and a cap rock (seal) or other mechanism that prevents it from escaping to the surface. Within these reservoirs, fluids will typically organize themselves like a three-layer cake with a layer of water below the oil layer and a layer of gas above it, although the different layers vary in size between reservoirs. Because most hydrocarbons are less dense than rock or water, they often migrate upward through adjacent rock layers until either reaching the surface or becoming trapped within porous rocks (known as reservoirs) by impermeable rocks above. However, the process is influenced by underground water flows, causing oil to migrate hundreds of kilometres horizontally or even short distances downward before becoming trapped in a reservoir. When hydrocarbons are concentrated in a trap, an oil field forms, from which the liquid can be extracted by drilling and pumping.

The reactions that produce oil and natural gas are often modeled as first order breakdown reactions, where hydrocarbons are broken down to oil and natural gas by a set of parallel reactions, and oil eventually breaks down to natural gas by another set of reactions. The latter set is regularly used in petrochemical plants and oil refineries.

Wells are drilled into oil reservoirs to extract the crude oil. “Natural lift” production methods that rely on the natural reservoir pressure to force the oil to the surface are usually sufficient for a while after reservoirs are first tapped. In some reservoirs, such as in the Middle East, the natural pressure is sufficient over a long time. The natural pressure in most reservoirs, however, eventually dissipates. Then the oil must be extracted using “artificial lift” means. Over time, these “primary” methods become less effective and “secondary” production methods may be used. A common secondary method is “waterflood” or injection of water into the reservoir to increase pressure and force the oil to the drilled shaft or “wellbore.” Eventually “tertiary” or “enhanced” oil recovery methods may be used to increase the oil’s flow characteristics by injecting steam, carbon dioxide and other gases or chemicals into the reservoir. In the United States, primary production methods account for less than 40 percent of the oil produced on a daily basis, secondary methods account for about half, and tertiary recovery the remaining 10 percent. Extracting oil (or “bitumen”) from oil/tar sand and oil shale deposits requires mining the sand or shale and heating it in a vessel or retort, or using “in-situ” methods of injecting heated liquids into the deposit and then pumping out the oil-saturated liquid.

Unconventional oil reservoirs

Oil-eating bacteria biodegrade oil that has escaped to the surface. Oil sands are reservoirs of partially biodegraded oil still in the process of escaping and being biodegraded, but they contain so much migrating oil that, although most of it has escaped, vast amounts are still present—more than can be found in conventional oil reservoirs. The lighter fractions of the crude oil are destroyed first, resulting in reservoirs containing an extremely heavy form of crude oil, called crude bitumen in Canada, or extra-heavy crude oil in Venezuela. These two countries have the world’s largest deposits of oil sands.

On the other hand, oil shales are source rocks that have not been exposed to heat or pressure long enough to convert their trapped hydrocarbons into crude oil. Technically speaking, oil shales are not always shales and do not contain oil, but are fined-grain sedimentary rocks containing an insoluble organic solid called kerogen. The kerogen in the rock can be converted into crude oil using heat and pressure to simulate natural processes. The method has been known for centuries and was patented in 1694 under British Crown Patent No. 330 covering, “A way to extract and make great quantities of pitch, tar, and oil out of a sort of stone.” Although oil shales are found in many countries, the United States has the world’s largest deposits.

Contributors and Attributions

CC licensed content, Shared previously

- Petroleum. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Petroleum. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.12.2: Petroleum is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.12.3: Natural Gas

Natural gas is a fossil fuel formed when layers of buried plants, gases, and animals are exposed to intense heat and pressure over thousands of years. The energy that the plants originally obtained from the sun is stored in the form of chemical bonds in natural gas. Natural gas is a nonrenewable resource because it cannot be replenished on a human time frame. Natural gas is a hydrocarbon gas mixture consisting primarily of methane, but commonly includes varying amounts of other higher alkanes and sometimes a usually lesser percentage of carbon dioxide, nitrogen, and/or hydrogen sulfide. Natural gas is an energy source often used for heating, cooking, and electricity generation. It is also used as fuel for vehicles and as a chemical feedstock in the manufacture of plastics and other commercially important organic chemicals.

Natural gas is found in deep underground rock formations or associated with other hydrocarbon reservoirs in coal beds and as methane clathrates. Petroleum is another resource and fossil fuel found in close proximity to, and with natural gas. Most natural gas was created over time by two mechanisms: biogenic and thermogenic. Biogenic gas is created by methanogenic organisms in marshes, bogs, landfills, and shallow sediments. Deeper in the earth, at greater temperature and pressure, thermogenic gas is created from buried organic material.

Before natural gas can be used as a fuel, it must be processed to remove impurities, including water, to meet the specifications of marketable natural gas. The by-products of this processing include: ethane, propane, butanes, pentanes, and higher molecular weight hydrocarbons, hydrogen sulfide (which may be converted into pure sulfur), carbon dioxide, water vapor, and sometimes helium and nitrogen.

Natural gas is often informally referred to simply as “gas,” especially when compared to other energy sources such as oil or coal. However, it is not to be confused with gasoline, especially in North America, where the term gasoline is often shortened in colloquial usage to *gas*.

Natural gas was used by the Chinese in about 500 BC. They discovered a way to transport gas seeping from the ground in crude pipelines of bamboo to where it was used to boil salt water to extract the salt, in the Ziliujing District of Sichuan. The world’s first industrial extraction of natural gas started at Fredonia, New York, USA in 1825. By 2009, 66 trillion cubic meters (or 8%) had been used out of the total 850 trillion cubic meters of estimated remaining recoverable reserves of natural gas. Based on an estimated 2015 world consumption rate of about 3.4 trillion cubic meters of gas per year, the total estimated remaining economically recoverable reserves of natural gas would last 250 years at current consumption rates. An annual increase in usage of 2–3% could result in currently recoverable reserves lasting significantly less, perhaps as few as 80 to 100 years.

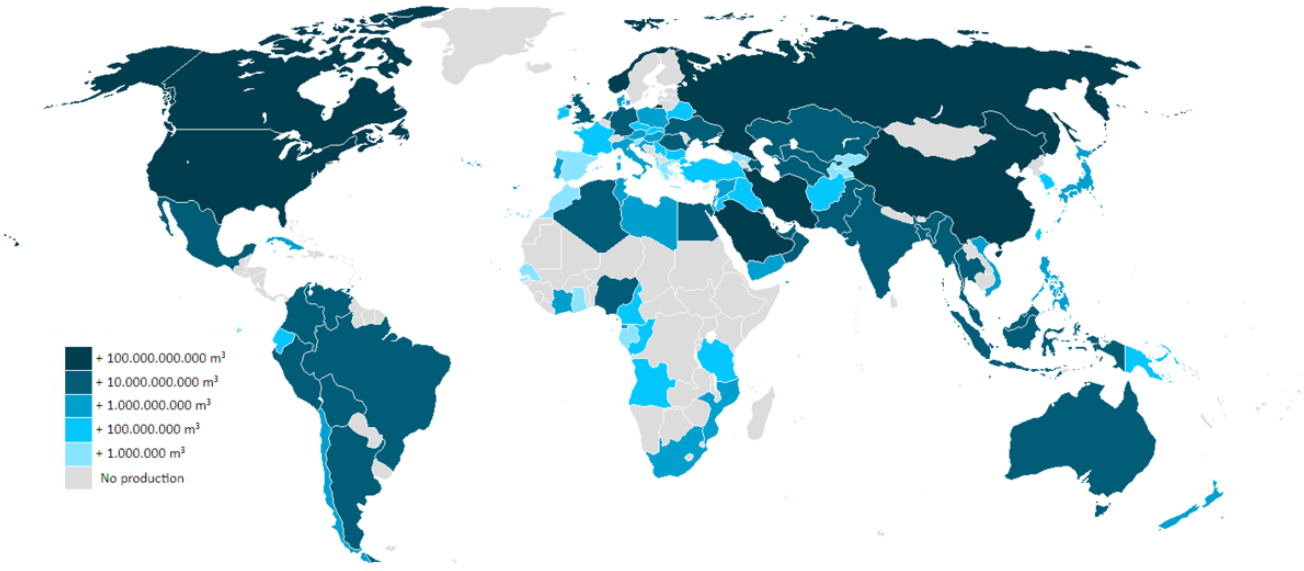
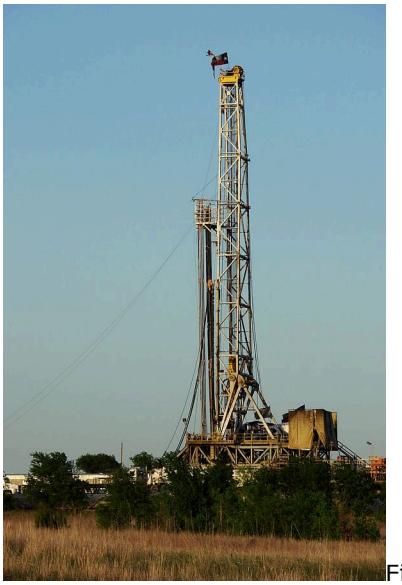


Figure 1. Natural gas extraction by countries in cubic meters per year.

Sources

Natural Gas


Figure 2. Natural gas drilling rig in Texas.

In the 19th century, natural gas was usually obtained as a by-product of producing oil, since the small, light gas carbon chains came out of solution as the extracted fluids underwent pressure reduction from the reservoir to the surface, similar to uncapping a soft drink bottle where the carbon dioxide effervesces. Unwanted natural gas was a disposal problem in the active oil fields. If there was not a market for natural gas near the wellhead it was virtually valueless since it had to be piped to the end user.

In the 19th century and early 20th century, such unwanted gas was usually burned off at oil fields. Today, unwanted gas (or stranded gas without a market) associated with oil extraction often is returned to the reservoir with 'injection' wells while awaiting a possible future market or to repressurize the formation, which can enhance extraction rates from other wells. In regions with a high natural gas demand (such as the US), pipelines are constructed when it is economically feasible to transport gas from a wellsite to an end consumer.

In addition to transporting gas via pipelines for use in power generation, other end uses for natural gas include export as liquefied natural gas (LNG) or conversion of natural gas into other liquid products via gas-to-liquids (GTL) technologies. GTL technologies can convert natural gas into liquids products such as gasoline, diesel or jet fuel. A variety of GTL technologies have been developed, including Fischer–Tropsch (F–T), methanol to gasoline (MTG) and STG+. F–T produces a synthetic crude that can be further refined into finished products, while MTG can produce synthetic gasoline from natural gas. STG+ can produce drop-in gasoline, diesel, jet fuel and aromatic chemicals directly from natural gas via a single-loop process. In 2011, Royal Dutch Shell's 140,000 barrel per day F–T plant went into operation in Qatar.

Natural gas can be "associated" (found in oil fields), or "non-associated" (isolated in natural gas fields), and is also found in coal beds (as coalbed methane). It sometimes contains a significant amount of ethane, propane, butane, and pentane—heavier hydrocarbons removed for commercial use prior to the methane being sold as a consumer fuel or chemical plant feedstock. Non-hydrocarbons such as carbon dioxide, nitrogen, helium (rarely), andhydrogen sulfide must also be removed before the natural gas can be transported.

Natural gas extracted from oil wells is called casinghead gas (whether or not truly produced up the annulus and through a casinghead outlet) or associated gas. The natural gas industry is extracting an increasing quantity of gas from challenging resource types: sour gas, tight gas, shale gas, and coalbed methane.

There is some disagreement on which country has the largest proven gas reserves. Sources that consider that Russia has by far the largest proven reserves include the US CIA (47.6 trillion cubic meters), the US Energy Information Administration (47.8 tcm), and OPEC (48.7 tcm). However, BP credits Russia with only 32.9 tcm, which would place it in second place, slightly behind Iran (33.1 to 33.8 tcm, depending on the source). With Gazprom, Russia is frequently the world's largest natural gas extractor. Major proven resources (in billion cubic meters) are world 187,300 (2013), Iran 33,600 (2013), Russia 32,900 (2013), Qatar 25,100

(2013), Turkmenistan 17,500 (2013) and the United States 8,500 (2013).

It is estimated that there are about 900 trillion cubic meters of "unconventional" gas such as shale gas, of which 180 trillion may be recoverable. In turn, many studies from MIT, Black & Veatchand the DOE predict that natural gas will account for a larger portion of electricity generation and heat in the future.

The world's largest gas field is the offshore South Pars / North Dome Gas-Condensate field, shared between Iran and Qatar. It is estimated to have 51 trillion cubic meters of natural gas and 50 billion barrels of natural gas condensates.

Because natural gas is not a pure product, as the reservoir pressure drops when non-associated gas is extracted from a field under supercritical (pressure/temperature) conditions, the higher molecular weight components may partially condense upon isothermic depressurizing—an effect called retrograde condensation. The liquid thus formed may get trapped as the pores of the gas reservoir get depleted. One method to deal with this problem is to re-inject dried gas free of condensate to maintain the underground pressure and to allow re-evaporation and extraction of condensates. More frequently, the liquid condenses at the surface, and one of the tasks of the gas plant is to collect this condensate. The resulting liquid is called natural gas liquid (NGL) and has commercial value.

Shale gas

Shale gas is natural gas produced from shale. Because shale has matrix permeability too low to allow gas to flow in economical quantities, shale gas wells depend on fractures to allow the gas to flow. Early shale gas wells depended on natural fractures through which gas flowed; almost all shale gas wells today require fractures artificially created by hydraulic fracturing. Since 2000, shale gas has become a major source of natural gas in the United States and Canada. Following the success in the United States, shale gas exploration is beginning in countries such as Poland, China, and South Africa. With the increase of shale production it has caused the United States to become the number one natural gas producer in the world.

Town Gas

Town gas is a flammable gaseous fuel made by the destructive distillation of coal and contains a variety of calorific gases including hydrogen, carbon monoxide, methane, and other volatile hydrocarbons, together with small quantities of non-calorific gases such as carbon dioxide and nitrogen, and is used in a similar way to natural gas. This is a historical technology, not usually economically competitive with other sources of fuel gas today. But there are still some specific cases where it is the best option and it may be so into the future.

Most town "gashouses" located in the eastern US in the late 19th and early 20th centuries were simple by-product coke ovens that heated bituminous coal in air-tight chambers. The gas driven off from the coal was collected and distributed through networks of pipes to residences and other buildings where it was used for cooking and lighting. (Gas heating did not come into widespread use until the last half of the 20th century.) The coal tar (or asphalt) that collected in the bottoms of the gashouse ovens was often used for roofing and other waterproofing purposes, and when mixed with sand and gravel was used for paving streets.

Biogas

Methanogenic archaea are responsible for all biological sources of methane. Some live in symbiotic relationships with other life forms, including termites, ruminants, and cultivated crops. Other sources of methane, the principal component of natural gas, include landfill gas, biogas, and methane hydrate. When methane-rich gases are produced by the anaerobic decay of non-fossilorganic matter (biomass), these are referred to as biogas (or natural biogas). Sources of biogas include swamps, marshes, and landfills (see landfill gas), as well as agricultural waste materials such as sewage sludge and manure by way of anaerobic digesters, in addition to enteric fermentation, particularly in cattle. Landfill gas is created by decomposition of waste in landfill sites. Excluding water vapor, about half of landfill gas is methane and most of the rest is carbon dioxide, with small amounts of nitrogen, oxygen, and hydrogen, and variable trace amounts of hydrogen sulfide and siloxanes. If the gas is not removed, the pressure may get so high that it works its way to the surface, causing damage to the landfill structure, unpleasant odor, vegetation die-off, and an explosion hazard. The gas can be vented to the atmosphere, flared or burned to produce electricity or heat. Biogas can also be produced by separating organic materials from waste that otherwise goes to landfills. This method is more efficient than just capturing the landfill gas it produces. Anaerobic lagoons produce biogas from manure, while biogas reactors can be used for manure or plant parts. Like landfill gas, biogas is mostly methane and carbon dioxide, with small amounts of nitrogen, oxygen and hydrogen. However, with the exception of pesticides, there are usually lower levels of contaminants.

Landfill gas cannot be distributed through utility natural gas pipelines unless it is cleaned up to less than 3 per cent CO_2, and a few parts per million H_2S, because CO_2 and H_2S corrode the pipelines. The presence of CO_2 will lower the energy level of the gas below requirements for the pipeline. Siloxanes in the gas will form deposits in gas burners and need to be removed prior to entry into any gas distribution or transmission system. Consequently it may be more economical to burn the gas on site or within a short distance of the landfill using a dedicated pipeline. Water vapor is often removed, even if the gas is burned on site. If low temperatures condense water out of the gas, siloxanes can be lowered as well because they tend to condense out with the water vapor. Other non-methane components may also be removed to meet emission standards, to prevent fouling of the equipment or for environmental considerations. Co-firing landfill gas with natural gas improves combustion, which lowers emissions.

Biogas, and especially landfill gas, are already used in some areas, but their use could be greatly expanded. Experimental systems were being proposed for use in parts of Hertfordshire, UK, and Lyon in France. Using materials that would otherwise generate no income, or even cost money to get rid of, improves the profitability and energy balance of biogas production. Gas generated in sewage treatment plants is commonly used to generate electricity. For example, the Hyperion sewage plant in Los Angeles burns 8 million cubic feet (230,000 m^3) of gas per day to generate power New York City utilizes gas to run equipment in the sewage plants, to generate electricity, and in boilers. Using sewage gas to make electricity is not limited to large cities. The city of Bakersfield, California, uses cogeneration at its sewer plants. California has 242 sewage wastewater treatment plants, 74 of which have installed anaerobic digesters. The total biopower generation from the 74 plants is about 66 MW.

Crystallized Natural Gas—Hydrates



Figure 3. The McMahon natural gas processing plant in Taylor, British Columbia, Canada.

Huge quantities of natural gas (primarily methane) exist in the form of hydrates under sediment on offshore continental shelves and on land in arctic regions that experience permafrost, such as those in Siberia. Hydrates require a combination of high pressure and low temperature to form.

In 2010, the cost of extracting natural gas from crystallized natural gas was estimated to 100–200 per cent the cost of extracting natural gas from conventional sources, and even higher from offshore deposits.

In 2013, Japan Oil, Gas and Metals National Corporation (JOGMEC) announced that they had recovered commercially relevant quantities of natural gas from methane hydrate.

Depletion

Natural gas production in the U.S. reached a peak in 1973, and went over a second lower peak in 2001, but recently has peaked again and is continuing to rise.

Uses

Mid-Stream Natural Gas

Natural gas flowing in the distribution lines and at the natural gas well head are often used to power natural gas powered engines. These engines rotate compressors to facilitate the natural gas transmission. These compressors are required in the mid-stream

line to pressurize and to re-pressurize the natural gas in the transmission line as the gas travels. The natural gas transmission lines extend to the natural gas processing plant or unit which removes the higher molecular weighted natural gas hydrocarbons to produce a British thermal unit (BTU) value between 950 and 1050 BTUs. The processed natural gas may then be used for residential, commercial and industrial uses.

Often mid-stream and well head gases require removal of many of the various hydrocarbon species contained within the natural gas. Some of these gases include heptane, pentane, propane and other hydrocarbons with molecular weights above Methane (CH4) to produce a natural gas fuel which is used to operate the natural gas engines for further pressurized transmission. Typically, natural gas compressors require 950 to 1050 BTU per cubic foot to operate at the natural gas engines rotational name plate specifications.

Several methods are used to remove these higher molecular weighted gases for use at the natural gas engine. A few technologies are as follows:

- Joule–Thomson skid
- Cryogenic or chiller system
- Chemical enzymology system

Power Generation

Natural gas is a major source of electricity generation through the use of cogeneration, gas turbines and steam turbines. Natural gas is also well suited for a combined use in association with renewable energy sources such as wind or solar and for alimenting peak-load power stations functioning in tandem with hydroelectric plants. Most grid peaking power plants and some off-grid engine-generators use natural gas. Particularly high efficiencies can be achieved through combining gas turbines with a steam turbine in combined cycle mode. Natural gas burns more cleanly than other hydrocarbon fuels, such as oil and coal, and produces less carbon dioxide per unit of energy released. For transportation, burning natural gas produces about 30 per cent less carbon dioxide than burning petroleum. For an equivalent amount of heat, burning natural gas produces about 45 per cent less than burning coal for power. The US Energy Information Administration reports the following emissions in million metric tons of carbon dioxide in the world for 2012:

- Natural gas: 6,799
- Petroleum: 11,695
- Coal: 13,787

Coal-fired electric power generation emits around 2,000 pounds of carbon dioxide for every megawatt hour generated, which is almost double the carbon dioxide released by a natural gas-fired electric plant per megawatt hour generated. Because of this higher carbon efficiency of natural gas generation, as the fuel mix in the United States has changed to reduce coal and increase natural gas generation, carbon dioxide emissions have unexpectedly fallen. Those measured in the first quarter of 2012 were the lowest of any recorded for the first quarter of any year since 1992.

Combined cycle power generation using natural gas is currently the cleanest available source of power using hydrocarbon fuels, and this technology is widely and increasingly used as natural gas can be obtained at increasingly reasonable costs. Fuel cell technology may eventually provide cleaner options for converting natural gas into electricity, but as yet it is not price-competitive. Locally produced electricity and heat using natural gas powered Combined Heat and Power plant (CHP or Cogeneration plant) is considered energy efficient and a rapid way to cut carbon emissions.Natural gas power plants are increasing in popularity and generate 22% of the worlds total electricity. Approximately half as much as generated with coal.

Domestic Use

Natural gas dispensed from a simple stovetop can generate temperatures in excess of 1100 °C (2000 °F) making it a powerful domestic cooking and heating fuel. In much of the developed world it is supplied through pipes to homes, where it is used for many purposes including ranges and ovens, gas-heated clothes dryers, heating/cooling, and central heating. Heaters in homes and other buildings may include boilers, furnaces, and water heaters.

Compressed natural gas (CNG) is used in rural homes without connections to piped-in public utility services, or with portable grills. Natural gas is also supplied by independent natural gas suppliers through Natural Gas Choice programs throughout the United States. However, as CNG costs more than LPG, LPG (propane) is the dominant source of rural gas.

Transportation



Figure 4. A Washington, D.C. Metrobus, which runs on natural gas.

CNG is a cleaner and also cheaper alternative to other automobile fuels such as gasoline (petrol) and diesel. By the end of 2012 there were 17.25 millionnatural gas vehicles worldwide, led by Iran (3.3 million), Pakistan (3.1 million), Argentina (2.18 million), Brazil (1.73 million), India (1.5 million), and China(1.5 million). The energy efficiency is generally equal to that of gasoline engines, but lower compared with modern diesel engines. Gasoline/petrol vehicles converted to run on natural gas suffer because of the low compression ratio of their engines, resulting in a cropping of delivered power while running on natural gas (10%–15%). CNG-specific engines, however, use a higher compression ratio due to this fuel's higher octane number of 120–130.

Contributors and Attributions

CC licensed content, Shared previously

- Natural gas. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Natural_gas. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.12.3: Natural Gas is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.12.4: Coal


Figure 1. Bituminous coal

Coal (from the Old English term *col*, which has meant “mineral of fossilized carbon” since the thirteent century)is a combustible black or brownish-black sedimentary rock usually occurring in rock strata in layers or veins called **coal beds** or **coal seams**. The harder forms, such as anthracite coal, can be regarded as metamorphic rock because of later exposure to elevated temperature and pressure. Coal is composed primarily of carbon along with variable quantities of other elements, chiefly hydrogen, sulfur, oxygen, and nitrogen.

Throughout history, coal has been used as an energy resource, primarily burned for the production of electricity and/or heat, and is also used for industrial purposes, such as refining metals. A fossil fuel, coal forms when dead plant matter is converted into peat, which in turn is converted into lignite, then sub-bituminous coal, after that bituminous coal, and lastly anthracite. This involves biological and geological processes that take place over a long period. The United States Energy Information Administration estimates coal reserves at 948×10^9 short tons (860 Gt). One estimate for resources is 18 000 Gt.

Coal is the largest source of energy for the generation of electricity worldwide, as well as one of the largest worldwide anthropogenic sources ofcarbon dioxide releases. In 1999, world gross carbon dioxide emissions from coal usage were 8,666 million tonnes of carbon dioxide. In 2011, world gross emissions from coal usage were 14,416 million tonnes. Coal-fired electric power generation emits around 2,000 pounds of carbon dioxide for every megawatt-hour generated, which is almost double the approximately 1100 pounds of carbon dioxide released by a natural gas-fired electric plant per megawatt-hour generated. Because of this higher carbon efficiency of natural gas generation, as the market in the United States has changed to reduce coal and increase natural gas generation, carbon dioxide emissions have fallen. Those measured in the first quarter of 2012 were the lowest of any recorded for the first quarter of any year since 1992 In 2013, the head of the UN climate agency advised that most of the world’s coal reserves should be left in the ground to avoid catastrophic global warming.

Coal is extracted from the ground by coal mining, either underground by shaft mining, or at ground level by open pit mining extraction. Since 1983 the world top coal producer has been China. In 2011 China produced 3,520 million tonnes of coal – 49.5% of 7,695 million tonnes world coal production. In 2011 other large producers were United States (993 million tonnes), India (589), European Union (576) and Australia (416). In 2010 the largest exporters were Australia with 328 million tonnes (27.1% of world coal export) and Indonesia with 316 million tonnes (26.1%), while the largest importers were Japan with 207 million tonnes (17.5% of world coal import), China with 195 million tonnes (16.6%) and South Korea with 126 million tonnes (10.7%).

Formation


Figure 2. Coastal exposure of the Point Aconi Seam (Nova Scotia)

At various times in the geologic past, the Earth had dense forests in low-lying wetland areas. Due to natural processes such as flooding, these forests were buried underneath soil. As more and more soil deposited over them, they were compressed. The temperature also rose as they sank deeper and deeper. As the process continued the plant matter was protected from biodegradation and oxidation, usually by mud or acidic water. This trapped the carbon in immense peat bogs that were eventually covered and deeply buried by sediments. Under high pressure and high temperature, dead vegetation was slowly converted to coal. As coal contains mainly carbon, the conversion of dead vegetation into coal is called carbonization.

The wide, shallow seas of the Carboniferous Period provided ideal conditions for coal formation, although coal is known from most geological periods. The exception is the coal gap in the Permian–Triassic extinction event, where coal is rare. Coal is known from Precambrian strata, which predate land plants—this coal is presumed to have originated from residues of algae.

Ranks

As geological processes apply pressure to dead biotic material over time, under suitable conditions, its metamorphic grade increases successively into:

- **Peat**, considered to be a precursor of coal, has industrial importance as a fuel in some regions, for example, Ireland and Finland. In its dehydrated form, peat is a highly effective absorbent for fuel and oil spills on land and water. It is also used as a conditioner for soil to make it more able to retain and slowly release water.
- **Lignite**, or brown coal, is the lowest rank of coal and used almost exclusively as fuel for electric power generation. Jet, a compact form of lignite, is sometimes polished and has been used as an ornamental stone since the Upper Palaeolithic.
- **Sub-bituminous coal**, whose properties range from those of lignite to those of bituminous coal, is used primarily as fuel for steam-electric power generation and is an important source of light aromatic hydrocarbons for the chemical synthesis industry.
- **Bituminous coal** is a dense sedimentary rock, usually black, but sometimes dark brown, often with well-defined bands of bright and dull material; it is used primarily as fuel in steam-electric power generation, with substantial quantities used for heat and power applications in manufacturing and to make coke.
- "**Steam coal**" is a grade between bituminous coal and anthracite, once widely used as a fuel for steam locomotives. In this specialized use, it is sometimes known as "sea-coal" in the United States. Small steam coal (dry small steam nuts or DSSN) was used as a fuel for domestic water heating.
- **Anthracite**, the highest rank of coal, is a harder, glossy black coal used primarily for residential and commercial space heating. It may be divided further into metamorphically altered bituminous coal and "petrified oil," as from the deposits in Pennsylvania.
- **Graphite**, technically the highest rank, is difficult to ignite and is not commonly used as fuel—it is mostly used in pencils and, when powdered, as a lubricant.

The classification of coal is generally based on the content of volatiles. However, the exact classification varies between countries. According to the German classification, coal is classified as follows:

German Classification	English Designation	Volatiles %	C Carbon %	H Hydrogen %	O Oxygen %	S Sulfur %	Heat content kJ/kg
Braunkohle	Lignite (brown coal)	45–65	60–75	6.0–5.8	34-17	0.5-3	<28,470
Flammkohle	Flame coal	40-45	75-82	6.0-5.8	>9.8	~1	<32,870
Gasflammkohle	Gas flame coal	35-40	82-85	5.8-5.6	9.8-7.3	~1	<33,910
Gaskohle	Gas coal	28-35	85-87.5	5.6-5.0	7.3-4.5	~1	<34,960
Fettkohle	Fat coal	19-28	87.5-89.5	5.0-4.5	4.5-3.2	~1	<35,380
Esskohle	Forge coal	14-19	89.5-90.5	4.5-4.0	3.2-2.8	~1	<35,380
Magerkohle	Nonbaking coal	10-14	90.5-91.5	4.0-3.75	2.8-3.5	~1	35,380
Anthrazit	Anthracite	7-12	>91.5	<3.75	<2.5	~1	<35,300
Note, the percentages are percent by mass of the indicated elements							

The middle six grades in the table represent a progressive transition from the English-language sub-bituminous to bituminous coal, while the last class is an approximate equivalent to anthracite, but more inclusive (US anthracite has < 6% volatiles).

Cannel coal (sometimes called “candle coal”) is a variety of fine-grained, high-rank coal with significant hydrogen content. It consists primarily of “exinite” macerals, now termed “liptinite.”

Hilt’s law

Hilt’s law is a geological term that states that, in a small area, the deeper the coal, the higher its rank (grade). The law holds true if the thermal gradient is entirely vertical, but metamorphism may cause lateral changes of rank, irrespective of depth.

Content

Average content

Substance	Content
Mercury (Hg)	0.10±0.01 ppm
Arsenic (As)	1.4 – 71 ppm
Selenium (Se)	3 ppm

Uses Today

Coal as Fuel


Figure 3. Castle Gate Power Plant near Helper, Utah, USA

Coal is primarily used as a solid fuel to produce electricity and heat through combustion. World coal consumption was about 7.25 billion tonnes in 2010(7.99 billion short tons) and is expected to increase 48% to 9.05 billion tonnes (9.98 billion short tons) by 2030.

China produced 3.47 billion tonnes (3.83 billion short tons) in 2011. India produced about 578 million tonnes (637.1 million short tons) in 2011. 68.7% of China's electricity comes from coal. The USA consumed about 13% of the world total in 2010, i.e. 951 million tonnes (1.05 billion short tons), using 93% of it for generation of electricity. 46% of total power generated in the USA was done using coal.


Figure 4. Coal rail cars

When coal is used for electricity generation, it is usually pulverized and then combusted (burned) in a furnace with a boiler. The furnace heat converts boiler water to steam, which is then used to spin turbines which turn generators and create electricity. The thermodynamic efficiency of this process has been improved over time; some older coal-fired power stations have thermal efficiencies in the vicinity of 25% whereas the newest supercritical and "ultra-supercritical" steam cycle turbines, operating at temperatures over 600 °C and pressures over 27 MPa (over 3900 psi), can practically achieve thermal efficiencies in excess of 45% (LHV basis) using anthracite fuel, or around 43% (LHV basis) even when using lower-grade lignite fuel.Further thermal efficiency improvements are also achievable by improved pre-drying (especially relevant with high-moisture fuel such as lignite or biomass) and cooling technologies.

An alternative approach of using coal for electricity generation with improved efficiency is the integrated gasification combined cycle (IGCC) power plant. Instead of pulverizing the coal and burning it directly as fuel in the steam-generating boiler, the coal can be first gasified (see coal gasification) to create syngas, which is burned in a gas turbine to produce electricity (just like

natural gas is burned in a turbine). Hot exhaust gases from the turbine are used to raise steam in a heat recovery steam generator which powers a supplemental steam turbine. Thermal efficiencies of current IGCC power plants range from 39-42% (HHV basis) or ~42-45% (LHV basis) for bituminous coal and assuming utilization of mainstream gasification technologies (Shell, GE Gasifier, CB&I). IGCC power plants outperform conventional pulverized coal-fueled plants in terms of pollutant emissions, and allow for relatively easy carbon capture.

At least 40% of the world's electricity comes from coal, and in 2012, about one-third of the United States' electricity came from coal, down from approximately 49% in 2008. As of 2012 in the United States, use of coal to generate electricity was declining, as plentiful supplies of natural gas obtained by hydraulic fracturing of tight shale formations became available at low prices.

In Denmark, a net electric efficiency of > 47% has been obtained at the coal-fired Nordjyllandsværket CHP Plant and an overall plant efficiency of up to 91% with cogeneration of electricity and district heating. The multifuel-fired Avedøreværket CHP Plant just outside Copenhagen can achieve a net electric efficiency as high as 49%. The overall plant efficiency with cogeneration of electricity and district heating can reach as much as 94%.

An alternative form of coal combustion is as coal-water slurry fuel (CWS), which was developed in the Soviet Union. CWS significantly reduces emissions, improving the heating value of coal. Other ways to use coal are combined heat and power cogeneration and an MHD topping cycle.

The total known deposits recoverable by current technologies, including highly polluting, low-energy content types of coal (i.e., lignite, bituminous), is sufficient for many years. However, consumption is increasing and maximal production could be reached within decades (see world coal reserves, below). On the other hand much may have to be left in the ground to avoid climate change.

Gasification

Coal gasification can be used to produce syngas, a mixture of carbon monoxide (CO) and hydrogen (H_2) gas. Often syngas is used to fire gas turbines to produce electricity, but the versatility of syngas also allows it to be converted into transportation fuels, such as gasoline and diesel, through the Fischer-Tropsch process; alternatively, syngas can be converted into methanol, which can be blended into fuel directly or converted to gasoline via the methanol to gasoline process.[59] Gasification combined with Fischer-Tropsch technology is currently used by the Sasol chemical company of South Africa to make motor vehicle fuels from coal and natural gas. Alternatively, the hydrogen obtained from gasification can be used for various purposes, such as powering ahydrogen economy, making ammonia, or upgrading fossil fuels.

During gasification, the coal is mixed with oxygen and steam while also being heated and pressurized. During the reaction, oxygen and water molecules oxidize the coal into carbon monoxide (CO), while also releasing hydrogen gas (H_2). This process has been conducted in both underground coal mines and in the production of town gas.

$$C\ (as\ Coal) + O_2 + H_2O \rightarrow H_2 + CO$$

If the refiner wants to produce gasoline, the syngas is collected at this state and routed into a Fischer-Tropsch reaction. If hydrogen is the desired end-product, however, the syngas is fed into thewater gas shift reaction, where more hydrogen is liberated.

$$CO + H_2O \rightarrow CO_2 + H_2$$

In the past, coal was converted to make coal gas (town gas), which was piped to customers to burn for illumination, heating, and cooking.

Liquefaction

Coal can also be converted into synthetic fuels equivalent to gasoline or diesel by several different direct processes (which do not intrinsically require gasification or indirect conversion). In the direct liquefaction processes, the coal is either hydrogenated or carbonized. Hydrogenation processes are the Bergius process, the SRC-I and SRC-II (Solvent Refined Coal) processes, the NUS Corporation hydrogenation process and several other single-stage and two-stage processes. In the process of low-temperature carbonization, coal is coked at temperatures between 360 and 750 °C (680 and 1,380 °F). These temperatures optimize the production of coal tars richer in lighter hydrocarbons than normal coal tar. The coal tar is then further processed into fuels. An overview of coal liquefaction and its future potential is available.

Coal liquefaction methods involve carbon dioxide (CO_2) emissions in the conversion process. If coal liquefaction is done without employing either carbon capture and storage (CCS) technologies or biomass blending, the result is lifecycle greenhouse gas footprints that are generally greater than those released in the extraction and refinement of liquid fuel production from crude oil. If CCS technologies are employed, reductions of 5–12% can be achieved in Coal to Liquid (CTL) plants and up to a 75% reduction is achievable when co-gasifying coal with commercially demonstrated levels of biomass (30% biomass by weight) in coal/biomass-to-liquids plants. For future synthetic fuel projects, carbon dioxide sequestration is proposed to avoid releasing CO_2 into the atmosphere. Sequestration adds to the cost of production.

Refined Coal

Refined coal is the product of a coal-upgrading technology that removes moisture and certain pollutants from lower-rank coals such as sub-bituminous and lignite (brown) coals. It is one form of several precombustion treatments and processes for coal that alter coal's characteristics before it is burned. The goals of precombustion coal technologies are to increase efficiency and reduce emissions when the coal is burned. Depending on the situation, precombustion technology can be used in place of or as a supplement to postcombustion technologies to control emissions from coal-fueled boilers.

Industrial Processes

Finely ground bituminous coal, known in this application as sea coal, is a constituent of foundry sand. While the molten metal is in the mould, the coal burns slowly, releasing reducing gases at pressure, and so preventing the metal from penetrating the pores of the sand. It is also contained in "mould wash," a paste or liquid with the same function applied to the mould before casting.Sea coal can be mixed with the clay lining (the "bod") used for the bottom of a cupola furnace. When heated, the coal decomposes and the bod becomes slightly friable, easing the process of breaking open holes for tapping the molten metal.

Production of Chemicals

Coal is an important feedstock in production of a wide range of chemical fertilizers and other chemical products. The main route to these products is coal gasification to produce syngas. Primary chemicals that are produced directly from the syngas include methanol, hydrogen and carbon monoxide, which are the chemical building blocks from which a whole spectrum of derivative chemicals are manufactured, including olefins, acetic acid, formaldehyde, ammonia,urea and others. The versatility of syngas as a precursor to primary chemicals and high-value derivative products provides the option of using relatively inexpensive coal to produce a wide range of valuable commodities.

Historically, production of chemicals from coal has been used since the 1950s and has become established in the market. According to the 2010 Worldwide Gasification Database, a survey of current and planned gasifiers, from 2004 to 2007 chemical production increased its gasification product share from 37% to 45%. From 2008 to 2010, 22% of new gasifier additions were to be for chemical production.

Because the slate of chemical products that can be made via coal gasification can in general also use feedstocks derived from natural gas and petroleum, the chemical industry tends to use whatever feedstocks are most cost-effective. Therefore, interest in using coal tends to increase for higher oil and natural gas prices and during periods of high global economic growth that may strain oil and gas production. Also, production of chemicals from coal is of much higher interest in countries like South Africa, China, India and the United States where there are abundant coal resources. The abundance of coal combined with lack of natural gas resources in China is strong inducement for the coal to chemicals industry pursued there. In the United States, the best example of the industry is Eastman Chemical Company which has been successfully operating a coal-to-chemicals plant at its Kingsport, Tennessee, site since 1983. Similarly, Sasol has built and operated coal-to-chemicals facilities in South Africa.

Coal to chemical processes do require substantial quantities of water. As of 2013 much of the coal to chemical production was in the People's Republic of China where environmental regulation and water management was weak.

World Coal Reserves

The 948 billion short tons of recoverable coal reserves estimated by the Energy Information Administration are equal to about 4,196 BBOE (billion barrels of oil equivalent). The amount of coal burned during 2007 was estimated at 7.075 billion short tons, or 133.179 quadrillion BTU's. This is an average of 18.8 million BTU per short ton. In terms of heat content, this is about 57,000,000 barrels (9,100,000 m^3) of oil equivalent per day. By comparison in 2007, natural gas provided 51,000,000 barrels (8,100,000

m^3) of oil equivalent per day, while oil provided 85,800,000 barrels (13,640,000 m^3) per day.


Figure 5. A coal mine in Wyoming, United States. The United States has the world's largest coal reserves.

British Petroleum, in its 2007 report, estimated at 2006 end that there were 147 years reserves-to-production ratio based on *proven* coal reserves worldwide. This figure only includes reserves classified as "proven"; exploration drilling programs by mining companies, particularly in under-explored areas, are continually providing new reserves. In many cases, companies are aware of coal deposits that have not been sufficiently drilled to qualify as "proven." However, some nations haven't updated their information and assume reserves remain at the same levels even with withdrawals.

Of the three fossil fuels, coal has the most widely distributed reserves; coal is mined in over 100 countries, and on all continents except Antarctica. The largest reserves are found in the United States, Russia, China, Australia and India.

Contributors and Attributions

CC licensed content, Shared previously

- Coal. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Coal. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.12.4: Coal is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.12.5: Mining


Figure 1. Surface coal mining

Mining is the extraction of valuable minerals or other geological materials from the earth from an orebody, lode, vein, seam, or reef, which forms the mineralized package of economic interest to the miner.

Ores recovered by mining include metals, coal, oil shale, gemstones, limestone, dimension stone, rock salt, potash, gravel, and clay. Mining is required to obtain any material that cannot be grown through agricultural processes, or created artificially in a laboratory or factory. Mining in a wider sense includes extraction of any non-renewable resource such as petroleum, natural gas, or even water.

Mining of stone and metal has been done since pre-historic times. Modern mining processes involve prospecting for ore bodies, analysis of the profit potential of a proposed mine, extraction of the desired materials, and final reclamation of the land after the mine is closed.

The nature of mining processes creates a potential negative impact on the environment both during the mining operations and for years after the mine is closed. This impact has led most of the world's nations to adopt regulations designed to moderate the negative effects of mining operations. Safety has long been a concern as well, and modern practices have improved safety in mines significantly.

Mine Development and Lifecycle

The process of mining from discovery of an ore body through extraction of minerals and finally to returning the land to its natural state consists of several distinct steps. The first is discovery of the ore body, which is carried out through prospecting or exploration to find and then define the extent, location and value of the ore body. This leads to a mathematical resource estimation to estimate the size and grade of the deposit.

This estimation is used to conduct a pre-feasibility study to determine the theoretical economics of the ore deposit. This identifies, early on, whether further investment in estimation and engineering studies is warranted and identifies key risks and areas for further work. The next step is to conduct a feasibility study to evaluate the financial viability, the technical and financial risks, and the robustness of the project.

This is when the mining company makes the decision whether to develop the mine or to walk away from the project. This includes mine planning to evaluate the economically recoverable portion of the deposit, the metallurgy and ore recoverability, marketability and payability of the ore concentrates, engineering concerns, milling and infrastructure costs, finance and equity requirements, and an analysis of the proposed mine from the initial excavation all the way through to reclamation. The proportion of a deposit that is economically recoverable is dependent on the enrichment factor of the ore in the area.

To gain access to the mineral deposit within an area it is often necessary to mine through or remove waste material which is not of immediate interest to the miner. The total movement of ore and waste constitutes the mining process. Often more waste than ore is mined during the life of a mine, depending on the nature and location of the ore body. Waste removal and placement is a

major cost to the mining operator, so a detailed characterization of the waste material forms an essential part of the geological exploration program for a mining operation.

Once the analysis determines a given ore body is worth recovering, development begins to create access to the ore body. The mine buildings and processing plants are built, and any necessary equipment is obtained. The operation of the mine to recover the ore begins and continues as long as the company operating the mine finds it economical to do so. Once all the ore that the mine can produce profitably is recovered, reclamation begins to make the land used by the mine suitable for future use.

Mining Techniques

Mining techniques can be divided into two common excavation types: surface mining and sub-surface (underground) mining. Today, surface mining is much more common, and produces, for example, 85% of minerals (excluding petroleum and natural gas) in the United States, including 98% of metallic ores.

Targets are divided into two general categories of materials: *placer deposits*, consisting of valuable minerals contained within river gravels, beach sands, and other unconsolidated materials; and *lode deposits*, where valuable minerals are found in veins, in layers, or in mineral grains generally distributed throughout a mass of actual rock. Both types of ore deposit, placer or lode, are mined by both surface and underground methods.

Some mining, including much of the rare earth elements and uranium mining, is done by less-common methods, such as in-situ leaching: this technique involves digging neither at the surface nor underground. The extraction of target minerals by this technique requires that they be soluble, e.g., potash, potassium chloride, sodium chloride, sodium sulfate, which dissolve in water. Some minerals, such as copper minerals and uranium oxide, require acid or carbonate solutions to dissolve.

Surface Mining

Surface mining is done by removing (stripping) surface vegetation, dirt, and, if necessary, layers of bedrock in order to reach buried ore deposits. Techniques of surface mining include: open-pit mining, which is the recovery of materials from an open pit in the ground, quarrying or gathering building materials from an open-pit mine; strip mining, which consists of stripping surface layers off to reveal ore/seams underneath; and mountaintop removal, commonly associated with coal mining, which involves taking the top of a mountain off to reach ore deposits at depth. Most (but not all) placer deposits, because of their shallowly buried nature, are mined by surface methods. Finally, landfill mining involves sites where landfills are excavated and processed.



Figure 2. Garzweiler surface mine, Germany

Open-Pit Mining

Open-pit mining, or **open-cast mining** is a surface mining technique of extracting rock or minerals from the earth by their removal from an open pit or borrow.

This form of mining differs from extractive methods that require tunneling into the earth, such as long wall mining. Open-pit mines are used when deposits of commercially useful minerals or rocks are found near the surface; that is, where the overburden (surface material covering the valuable deposit) is relatively thin or the material of interest is structurally unsuitable for tunneling (as would be the case for sand, cinder, and gravel). For minerals that occur deep below the surface—where the overburden is thick or the mineral occurs as veins in hard rock—underground mining methods extract the valued material.

Open-pit mines that produce building materials and dimension stone are commonly referred to as “quarries.”

Open-pit mines are typically enlarged until either the mineral resource is exhausted, or an increasing ratio of overburden to ore makes further mining uneconomic. When this occurs, the exhausted mines are sometimes converted to landfills for disposal of solid wastes. However, some form of water control is usually required to keep the mine pit from becoming a lake, if the mine is situated in a climate of considerable precipitation or if any layers of the pit forming the mine border productive aquifers.

 Figure 3. Note the angled and stepped sides of the Sunrise Dam Gold Mine, Australia.

Open-cast mines are dug on benches, which describe vertical levels of the hole. These benches are usually on four to sixty meter intervals, depending on the size of the machinery that is being used. Many quarries do not use benches, as they are usually shallow.

Most walls of the pit are generally dug on an angle less than vertical, to prevent and minimize damage and danger from rock falls. This depends on how weathered the rocks are, and the type of rock, and also how many structural weaknesses occur within the rocks, such as a faults, shears, joints orfoliations.

The walls are stepped. The inclined section of the wall is known as the batter, and the flat part of the step is known as the bench or berm. The steps in the walls help prevent rock falls continuing down the entire face of the wall. In some instances additional ground support is required and rock bolts, cable bolts and shotcrete are used. De-watering bores may be used to relieve water pressure by drilling horizontally into the wall, which is often enough to cause failures in the wall by itself.

A haul road is usually situated at the side of the pit, forming a ramp up which trucks can drive, carrying ore and waste rock.

Waste rock is piled up at the surface, near the edge of the open pit. This is known as the waste dump. The waste dump is also tiered and stepped, to minimize degradation.

Ore which has been processed is known as tailings, and is generally a slurry. This is pumped to a tailings dam or settling pond, where the water evaporates. Tailings dams can often be toxic due to the presence of unextracted sulfide minerals, some forms of toxic minerals in the gangue, and often cyanide which is used to treat gold ore via the cyanide leach process. This toxicity can harm the surrounding environment.

Typical Open Cut Grades

Gold is generally extracted in open-pit mines at 1 to 2 ppm (parts per million) but in certain cases, 0.75 ppm gold is economical. This was achieved by bulk heap leaching at the Peak Hill mine in western New South Wales, near Dubbo, Australia.

Nickel, generally as laterite, is extracted via open-pit down to 0.2%. Copper is extracted at grades as low as 0.15% to 0.2%, generally in massive open-pit mines in Chile, where the size of the resources and favorable metallurgy allows economies of scale.

Materials typically extracted from open-pit mines include:

- Bitumen
- Clay
- Coal
- Copper
- Coquina
- Diamonds
- Gravel and stone (stone refers to bedrock, while gravel is unconsolidated material)
- Granite
- Gritstone
- Gypsum

- Limestone
- Marble
- Metal ores, such as Copper, Iron, Gold, Silver and Molybdenum
- Uranium
- Phosphate

Underground Mining


Figure 4. Mantrip used for transporting miners within an underground mine

Sub-surface mining consists of digging tunnels or shafts into the earth to reach buried ore deposits. Ore, for processing, and waste rock, for disposal, are brought to the surface through the tunnels and shafts. Sub-surface mining can be classified by the type of access shafts used, the extraction method or the technique used to reach the mineral deposit. Drift mining utilizes horizontal access tunnels, slope mining uses diagonally sloping access shafts, and shaft mining utilizes vertical access shafts. Mining in hard and soft rock formations require different techniques.

Other methods include shrinkage stope mining, which is mining upward, creating a sloping underground room, long wall mining, which is grinding a long ore surface underground, and room and pillar mining, which is removing ore from rooms while leaving pillars in place to support the roof of the room. Room and pillar mining often leads to retreat mining, in which supporting pillars are removed as miners retreat, allowing the room to cave in, thereby loosening more ore. Additional sub-surface mining methods include hard rock mining, which is mining of hard rock (igneous, metamorphic or sedimentary) materials, bore hole mining, drift and fill mining, long hole slope mining, sub level caving, and block caving.

Machines

Heavy machinery is used in mining to explore and develop sites, to remove and stockpile overburden, to break and remove rocks of various hardness and toughness, to process the ore, and to carry out reclamation projects after the mine is closed. Bulldozers, drills, explosives and trucks are all necessary for excavating the land. In the case of placer mining, unconsolidated gravel, or alluvium, is fed into machinery consisting of a hopper and a shaking screen or trommel which frees the desired minerals from the waste gravel. The minerals are then concentrated using sluices or jigs.

Large drills are used to sink shafts, excavate stopes, and obtain samples for analysis. Trams are used to transport miners, minerals and waste. Lifts carry miners into and out of mines, and move rock and ore out, and machinery in and out, of underground mines. Huge trucks, shovels and cranes are employed in surface mining to move large quantities of overburden and ore. Processing plants utilize large crushers, mills, reactors, roasters and other equipment to consolidate the mineral-rich material and extract the desired compounds and metals from the ore.



Figure 5. The Bagger 288 is a bucket-wheel excavator used in strip mining. It is also the largest land vehicle of all time.

Processing

Once the mineral is extracted, it is often then processed. The science of extractive metallurgy is a specialized area in the science of metallurgy that studies the extraction of valuable metals from their ores, especially through chemical or mechanical means.

Mineral processing (or mineral dressing) is a specialized area in the science of metallurgy that studies the mechanical means of crushing, grinding, and washing that enable the separation (extractive metallurgy) of valuable metals or minerals from their gangue (waste material). Processing of placer ore material consists of gravity-dependent methods of separation, such as sluice boxes. Only minor shaking or washing may be necessary to disaggregate (unclump) the sands or gravels before processing. Processing of ore from a lode mine, whether it is a surface or subsurface mine, requires that the rock ore be crushed and pulverized before extraction of the valuable minerals begins. After lode ore is crushed, recovery of the valuable minerals is done by one, or a combination of several, mechanical and chemical techniques.

Since most metals are present in ores as oxides or sulfides, the metal needs to be reduced to its metallic form. This can be accomplished through chemical means such as smelting or through electrolytic reduction, as in the case of aluminium. Geometallurgy combines the geologic sciences with extractive metallurgy and mining.

Mining Industry

Mining exists in many countries. London is known as the capital of global “mining houses” such as Rio Tinto Group, BHP Billiton, and Anglo American PLC. The US mining industry is also large, but it is dominated by the coal and other nonmetal minerals (e.g., rock and sand), and various regulations have worked to reduce the significance of mining in the United States. In 2007 the totalmarket capitalization of mining companies was reported at US$962 billion, which compares to a total global market cap of publicly traded companies of about US$50 trillion in 2007. In 2002, Chile and Peru were reportedly the major mining countries of South America. The mineral industry of Africa includes the mining of various minerals; it produces relatively little of the industrial metals copper, lead, and zinc, but according to one estimate has as a percent of world reserves 40% of gold, 60% of cobalt, and 90% of the world’s platinum group metals. Mining in India is a significant part of that country’s economy. In the developed world, mining in Australia, with BHP Billiton founded and headquartered in the country, and mining in Canada are particularly significant. For rare earth minerals mining, China reportedly controlled 95% of production in 2013.

Mining operations can be grouped into five major categories in terms of their respective resources. These are oil and gas extraction, coal mining, metal ore mining, nonmetallic mineral mining and quarrying, and mining support activities. Of all of these categories, oil and gas extraction remains one of the largest in terms of its global economic importance. Prospecting potential mining sites, a vital area of concern for the mining industry, is now done using sophisticated new technologies such as seismic prospecting and remote-sensing satellites. Mining is heavily affected by the prices of the commodity minerals, which are often volatile. The 2000s commodities boom (“commodities supercycle”) increased the prices of commodities, driving aggressive mining. In addition, the price of gold increased dramatically in the 2000s, which increased gold mining; for example, one study found that conversion of forest in the Amazon increased six-fold from the period 2003–2006 (292 ha/yr) to the period 2006–2009 (1,915 ha/yr), largely due to artisanal mining.

Safety

Safety has long been a concern in the mining business especially in sub-surface mining. The Courrières mine disaster, Europe's worst mining accident, involved the death of 1,099 miners in Northern France on March 10, 1906. This disaster was surpassed only by the Benxihu Colliery accident in China on April 26, 1942, which killed 1,549 miners. While mining today is substantially safer than it was in previous decades, mining accidents still occur. Government figures indicate that 5,000 Chinese miners die in accidents each year, while other reports have suggested a figure as high as 20,000.Mining accidents continue worldwide, including accidents causing dozens of fatalities at a time such as the 2007 Ulyanovskaya Mine disaster in Russia, the 2009 Heilongjiang mine explosion in China, and the 2010 Upper Big Branch Mine disaster in the United States.

Mining ventilation is a significant safety concern for many miners. Poor ventilation inside sub-surface mines causes exposure to harmful gases, heat, and dust, which can cause illness, injury, and death. The concentration of methane and other airborne contaminants underground can generally be controlled by dilution (ventilation), capture before entering the host air stream (methane drainage), or isolation (seals and stoppings). Rock dusts, including coal dust and silicon dust, can cause long-term lung problems including silicosis, asbestosis, and pneumoconiosis (also known as miners lung or black lungdisease). A ventilation system is set up to force a stream of air through the working areas of the mine. The air circulation necessary for effective ventilation of a mine is generated by one or more large mine fans, usually located above ground. Air flows in one direction only, making circuits through the mine such that each main work area constantly receives a supply of fresh air. Watering down in coal mines also helps to keep dust levels down: by spraying the machine with water and filtering the dust-laden water with a scrubber fan, miners can successfully trap the dust.

Gases in mines can poison the workers or displace the oxygen in the mine, causing asphyxiation. For this reason, the U.S. Mine Safety and Health Administration requires that groups of miners in the United States carry gas detection equipment that can detect common gases, such as CO, O_2, H_2S, CH_4, as well as calculate % Lower Explosive Limit. Regulation requires that all production stop if there is a concentration of 1.4% of flammable gas present. Additionally, further regulation is being requested for more gas detection as newer technology such as nanotechnology is introduced.

Ignited methane gas is a common source of explosions in coal mines, which in turn can initiate more extensive coal dust explosions. For this reason, rock dusts such as limestone dust are spread throughout coal mines to diminish the chances of coal dust explosions as well as to limit the extent of potential explosions, in a process known as rock dusting. Coal dust explosions can also begin independently of methane gas explosions. Frictional heat and sparks generated by mining equipment can ignite both methane gas and coal dust. For this reason, water is often used to cool rock-cutting sites.

Miners utilize equipment strong enough to break through extremely hard layers of the Earth's crust. This equipment, combined with the closed work space in which underground miners work, can cause hearing loss. For example, a roof bolter (commonly used by mine roof bolter operators) can reach sound power levels of up to 115 dB. Combined with the reverberant effects of underground mines, a miner without proper hearing protection is at a high risk forhearing loss. By age 50, nearly 90% of U.S. coal miners have some hearing loss, compared to only 10% among workers not exposed to loud noises.Roof bolters are among the loudest machines, but auger miners, bulldozers, continuous mining machines, front end loaders, and shuttle cars and trucks are also among those machines most responsible for excessive noise in mine work.

Since mining entails removing dirt and rock from its natural location, thereby creating large empty pits, rooms, and tunnels, cave-ins as well as ground and rock falls are a major concern within mines. Modern techniques for timbering and bracing walls and ceilings within sub-surface mines have reduced the number of fatalities due to cave-ins, but ground falls continue to represent up to 50% of mining fatalities. Even in cases where mine collapses are not instantly fatal, they can trap mine workers deep underground. Cases such as these often lead to high-profile rescue efforts, such as when 33 Chilean miners were trapped deep underground for 69 days in 2010.

High temperatures and humidity may result in heat-related illnesses, including heat stroke, which can be fatal. The presence of heavy equipment in confined spaces also poses a risk to miners. To improve the safety of mine workers, modern mines use automation and remote operation including, for example, such equipment as automated loaders and remotely operated rockbreakers. However, despite modern improvements to safety practices, mining remains a dangerous occupation throughout the world.

Contributors and Attributions

CC licensed content, Original

- Revision and adaptation. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY-SA: Attribution-ShareAlike*

CC licensed content, Shared previously

- Mining. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Mining. **License**: *CC BY-SA: Attribution-ShareAlike*
- Open-pit mining. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Open-pit_mining. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.12.5: Mining is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.12.6: Effects of Mining

Environmental Effects

Environmental issues can include erosion, formation of sinkholes, loss of biodiversity, and contamination of soil, groundwater and surface water by chemicals from mining processes. In some cases, additional forest logging is done in the vicinity of mines to create space for the storage of the created debris and soil. Contamination resulting from leakage of chemicals can also affect the health of the local population if not properly controlled. Extreme examples of pollution from mining activities include coal fires, which can last for years or even decades, producing massive amounts of environmental damage.

Mining companies in most countries are required to follow stringent environmental and rehabilitation codes in order to minimize environmental impact and avoid impacting human health. These codes and regulations all require the common steps of environmental impact assessment, development of environmental management plans, mine closure planning (which must be done before the start of mining operations), and environmental monitoring during operation and after closure. However, in some areas, particularly in the developing world, government regulations may not be well enforced.

Waste

Ore mills generate large amounts of waste, called tailings. For example, 99 tons of waste are generated per ton of copper, with even higher ratios in gold mining. These tailings can be toxic. Tailings, which are usually produced as a slurry, are most commonly dumped into ponds made from naturally existing valleys. These ponds are secured by impoundments (dams or embankment dams). In 2000 it was estimated that 3,500 tailings impoundments existed, and that every year, 2 to 5 major failures and 35 minor failures occurred; for example, in the Marcopper mining disaster at least 2 million tons of tailings were released into a local river. Subaqueous tailings disposal is another option. The mining industry has argued that submarine tailings disposal (STD), which disposes of tailings in the sea, is ideal because it avoids the risks of tailings ponds; although the practice is illegal in the United States and Canada, it is used in the developing world.

The waste is classified as either sterile or mineralised, with acid generating potential, and the movement and storage of this material forms a major part of the mine planning process. When the mineralised package is determined by an economic cut-off, the near-grade mineralised waste is usually dumped separately with view to later treatment should market conditions change and it becomes economically viable. Civil engineering design parameters are used in the design of the waste dumps, and special conditions apply to high-rainfall areas and to seismically active areas. Waste dump designs must meet all regulatory requirements of the country in whose jurisdiction the mine is located. It is also common practice to rehabilitate dumps to an internationally acceptable standard, which in some cases means that higher standards than the local regulatory standard are applied.

Open-Pit Mining



Figure 1. Opencut coal mine loadout station and reclaimed land at the North Antelope Rochelle coal mine in Wyoming, United States.

After mining finishes, the mine area must undergo rehabilitation. Waste dumps are contoured to flatten them out, to further stabilise them. If the ore contains sulfides it is usually covered with a layer of clay to prevent access of rain and oxygen from the air, which can oxidise the sulfides to produce sulfuric acid, a phenomenon known as acid mine drainage. This is then generally covered with soil, and vegetation is planted to help consolidate the material. Eventually this layer will erode, but it is generally hoped that the rate of leaching or acid will be slowed by the cover such that the environment can handle the load of acid and associated heavy metals. There are no long term studies on the success of these covers due to the relatively short time in which large scale open pit mining has existed. It may take hundreds to thousands of years for some waste dumps to become “acid neutral” and stop leaching to the environment. The dumps are usually fenced off to prevent livestock denuding them of vegetation. The open pit is then surrounded with afence, to prevent access, and it generally eventually fills up with ground water. In arid areas it may not fill due to deep groundwater levels.



Figure 2. An open-pit sulfur mine at Tarnobrzeg, Poland undergoing land rehabilitation

Metal Reserves and Recycling

During the twentieth century, the variety of metals used in society grew rapidly. Today, the development of major nations such as China and India and advances in technologies are fueling an ever greater demand. The result is that metal mining activities are expanding and more and more of the world’s metal stocks are above ground in use rather than below ground as unused reserves. An example is the in-use stock of copper. Between 1932 and 1999, copper in use in the USA rose from 73 kilograms (161 lb) to 238 kilograms (525 lb) per person.

95% of the energy used to make aluminum from bauxite ore is saved by using recycled material. However, levels of metals recycling are generally low. In 2010, the International Resource Panel, hosted by the United Nations Environment Programme (UNEP), published reports on metal stocks that exist within society and their recycling rates.

The report’s authors observed that the metal stocks in society can serve as huge mines above ground. However, they warned that the recycling rates of some rare metals used in applications such as mobile phones, battery packs for hybrid cars, and fuel cells are so low that unless future end-of-life recycling rates are dramatically stepped up these critical metals will become unavailable for use in modern technology.

Contributors and Attributions

CC licensed content, Original

- Revision and Adaptation. **Authored by**: Kimberly Schulte and Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY-SA: Attribution-ShareAlike*

CC licensed content, Shared previously

- Open-pit mining. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Open-pit_mining. **License**: *CC BY-SA: Attribution-ShareAlike*
- Mining. **Provided by**: Wikipedia. **Located at**: http://en.Wikipedia.org/wiki/Mining. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 4.12.6: Effects of Mining is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.12.7: Earth's Energy

Much of Earth's energy comes from the Sun. Nearly all life on Earth depends on solar energy since plants use sunlight to make food through the process of photosynthesis. Photosynthesis was the process that fed plants and animals, which in turn, over the course of millions of years, became fossil fuels. The Sun heats some areas of Earth more than other areas, which causes wind. The Sun's energy also drives the water cycle, which moves water over the surface of the Earth. Some of these types of energy can be harnessed for use by people.

The other main source of energy is Earth's internal heat. This heat has two origins: the breakdown of chemical elements by radioactivity, and the heat that is left over from when the planet came together. These two sources will be described in more detail in later chapters.

Energy Resources

Everything requires energy. Even when you are sitting as still as you possibly can, your body is using energy to breathe, circulate blood, digest food, and perform many other functions. Producing light or heat requires energy. Making something requires energy. Plants and animals all require energy to function. To repeat, everything requires energy!

The Need for Energy



Figure 1. Electrical transmission towers like the one shown in this picture help deliver the electricity people use for energy every day.

Energy is the ability to do work or produce change. Every living thing needs energy to perform its daily functions and even more energy to grow. Plants get energy from the "food" they make by photosynthesis, and animals get energy directly or indirectly from that food. People also use energy for many things, such as cooking food, keeping ice cream cold in the freezer, heating a house, constructing a skyscraper, or lighting their homes. Because billions of people all around the world use energy, there is a huge need for energy resources (figure 1). Energy conservation is something everyone can do now to help reduce the strain on energy resources.

The **law of conservation of energy** says that energy cannot be created or destroyed. This means that even though energy changes form, the total amount of energy always stays the same. How does energy get converted from one type to another when you kick a soccer ball? When your body breaks down the food you eat, it stores the energy from the food as **chemical energy**. Chemical energy is stored within chemical bonds. But some of this stored energy has to be released to make your leg muscles move. The chemical energy is converted to another form of energy called **kinetic energy**. Kinetic energy is the energy of anything in motion. Your muscles move your leg, your foot kicks the ball, and the ball gains kinetic energy from the kick. So you can think of the action of kicking the ball as a story of energy changing forms.

To learn the quadratic equations related to getting a rapidly moving car to overcome its kinetic energy and come to a stop, watch

this video:

A link to an interactive elements can be found at the bottom of this page.

Potential energy is energy that is stored. Potential energy has the potential to do work or the potential to be converted into other forms of energy. If a ball is sitting on the very edge at the top of the hill, it is not moving, but it has a lot of potential energy.

Animations showing the conversion of potential energy to kinetic energy can be seen at the following sites:

- Energy Transformation for Downhill Skiing
- Energy Transformation on a Roller Coaster
- Energy Transformation for a Dart

Energy, Fuel, and Heat

If you read a book beneath a lit lamp, that lamp has energy from electricity. The energy to make the electricity comes from **fuel.** Fuel has energy that it releases. A fuel is any material that can release energy in a chemical change.

What are some examples of fuel, and what are they used for?

1. Food is fuel for your body.
2. Sunlight is the energy plants need to make food by photosynthesis.
3. Gasoline is fuel for cars.
4. Hydrogen is fuel for the Sun.


Figure 2. A controlled fire.

For a fuel to be useful, its energy must be released in a way that can be controlled. Controlling the release of energy makes it possible for the energy to be used to do work. When fuel is used for its energy, it is usually burned, and most of the energy is released as **heat** (figure 2). The heat may then be used to do work. Think of a person striking a match to set some small twigs on fire. After the twigs burn for a while, they get hot enough to make some larger sticks burn. The fire keeps getting hotter, and soon it is hot enough to burn whole logs. Pretty soon the fire is roaring, and a pot of water placed on the fire starts to boil. Some of the liquid water evaporates.

What is the source of energy for boiling and evaporating the water? Although some chemical energy from the match was put into starting the fire, the heat to boil and evaporate the water comes from the energy that was stored in the wood. The wood is the fuel for the fire.

Types of Energy Resources



Figure 3. Anthracite coal is a non-renewable energy resource.

Energy resources are either renewable or non-renewable. **Non-renewable resources** are used faster than they can be replaced,so the supply available to society is limited (see example in figure 3). **Renewable resources** will not run out because they are replaced as quickly as they are used. Can you think of some renewable and non-renewable energy sources?

Non-renewable Resources

Fossil fuels—coal, oil, and natural gas—are the most common example of non-renewable energy resources. Fossil fuels are formed from fossils, the partially decomposed remains of once living plants and animals. These fossils took millions of years to form. When fossil fuels are burned for energy, they release pollutants into the atmosphere. Fossil fuels also release carbon dioxide and other greenhouse gases, which are causing global temperatures to rise. The environmental effects of fossil fuel use are discussed in the “Climate” and “Human Actions and the Atmosphere” chapters.

Renewable Resources

Renewable energy resources include solar, water, wind, biomass, and geothermal. These resources are either virtually limitless like the Sun, which will continue to shine for billions of years, or will be replaced faster than we can use them. Amounts of falling water or wind will change over the course of time, but they are quite abundant. Biomass energy, like wood for fire, can be replaced quickly.

The use of renewable resources may also cause problems. Some are expensive, while some, such as trees, have other uses. Some cause environmental problems. As the technology improves and more people use renewable energy, the prices may come down. At the same time, as we use up fossil fuels, coal, oil, and natural gas, these non-renewable resources will become more expensive. At some point, even if renewable energy costs are high, non-renewable energy will be even more expensive. Ultimately, we will have to use renewable sources.

Important Things to Consider about Energy Resources

With both renewable and non-renewable resources, there are at least two important things to consider. One is that we have to have a practical way to turn the resource into a useful form of energy. The other is that we have to consider what happens when we turn the resource into energy.

For example, if we get much less energy from burning a fuel than we put into making it, then that fuel is probably not a practical energy resource. On the other hand, if another fuel gives us large amounts of energy but creates large amounts of pollution, that fuel also may not be the best choice for an energy resource.

Today we rely on electricity more than ever, but the resources that currently supply our power are finite. The race is on to harness more renewable resources, but getting all that clean energy from production sites to homes and businesses is proving to be a major challenge.

A link to an interactive elements can be found at the bottom of this page.

Non-renewable Energy Resources

Millions of years ago, plants used energy from the Sun to form sugars, carbohydrates, and other energy-rich carbon compounds that were later transformed into coal, oil, or natural gas. The solar energy stored in these fuels is a rich source of energy. Although fossil fuels provide very high quality energy, they are non-renewable.

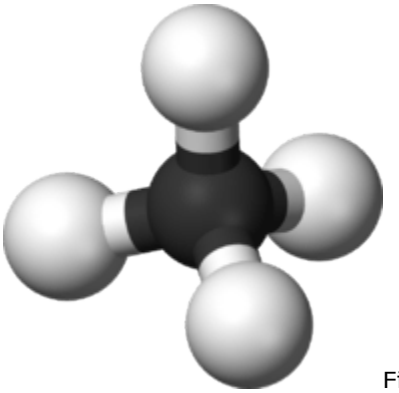
In large part, non-renewable energy sources are responsible for the world's lights seen in this animation.

Formation of Fossil Fuels


Figure 4. This wetland may look something like an ancient coal-forming swamp.

Can you name some fossils? How about dinosaur bones or dinosaur footprints? Animal skeletons, teeth, shells, coprolites (otherwise known as feces), or any other remains or trace from a living creature that becomes a rock is a **fossil**.

The same processes that formed these fossils also created some of our most important energy resources, **fossil fuels**. Coal, oil, and natural gas are fossil fuels. Fossil fuels come from living matter starting about 500 million years ago. As plants and animals died, their remains settled on the ground on land and in swamps, lakes, and seas (figure 4).


Figure 5. Hydrocarbons are made of carbon and hydrogen atoms. This molecule with one carbon and four hydrogen atoms is methane.

Over time, layer upon layer of these remains accumulated. Eventually, the layers were buried so deeply that they were crushed by an enormous mass of earth. The weight of this earth pressing down on these plant and animal remains created intense heat and pressure. After millions of years of heat and pressure, the material in these layers turned into chemicals called **hydrocarbons** (figure 5). Here is an animated view of a hydrocarbon.

Hydrocarbons can be solid, liquid, or gaseous. The solid form is what we know as coal. The liquid form is petroleum, or crude oil. Natural gas is the gaseous form.

Coal

Coal, a solid fossil fuel formed from the partially decomposed remains of ancient forests, is burned primarily to produce electricity. Coal use is undergoing enormous growth as the availability of oil and natural gas decreases and cost increases. This

increase in coal use is happening particularly in developing nations, such as China, where coal is cheap and plentiful.

Coal Formation

Coal forms from dead plants that settled at the bottom of ancient swamps. Lush coal swamps were common in the tropics during the Carboniferous period, which took place more than 300 million years ago (figure 6). The climate was warmer then.



Figure 6. The location of the continents during the Carboniferous period. Notice that quite a lot of land area is in the region of the tropics.



Figure 7. Bituminous coal is a sedimentary rock.

Mud and other dead plants buried the organic material in the swamp, and burial kept oxygen away. When plants are buried without oxygen, the organic material can be preserved or fossilized. Sand and clay settling on top of the decaying plants squeezed out the water and other substances. Millions of years later, what remains is a carbon-containing rock that we know as coal.

Coal is black or brownish-black. The most common form of coal is bituminous, a sedimentary rock that contains impurities such as sulfur (figure 7). Anthracite coal, seen in figure 3, has been metamorphosed and is nearly all carbon. For this reason, anthracite coal burns more cleanly than bituminous coal.

Coal Use

Around the world, coal is the largest source of energy for electricity. The United States is rich in coal (figure 8). California once had a number of small coal mines, but the state no longer produces coal. To turn coal into electricity, the rock is crushed into powder, which is then burned in a furnace that has a boiler. Like other fuels, coal releases its energy as heat when it burns. Heat from the burning coal boils the water in the boiler to make steam. The steam spins turbines, which turn generators to create electricity. In this way, the energy stored in the coal is converted to useful energy like electricity.

Figure 8. United States coal-producing regions in 1996. Orange is highest grade anthracite; red is low volatile bituminous; gray and gray-green is medium to high-volatile

bituminous; green is subbituminous; and yellow is the lowest grade lignite

Coal that has been located but is not being used is part of our reserves. Reserves are important because if the price of the resource goes up or the cost of extracting it goes down, they may be useful.

Consequences of Coal Use

For coal to be used as an energy source, it must first be mined. Coal mining occurs at the surface or underground by methods that are described in the "Earth's Minerals" chapter (figure 9). Mining, especially underground mining, can be dangerous. In April 2010, twenty-nine miners were killed at a West Virginia coal mine when gas that had accumulated in the mine tunnels exploded and started a fire.



Coal being mined by mountaintop removal.



A small coal-fired power plant.

Figure 9. The coal used in power plants must be mined. One method to mine coal is by mountaintop removal.

Some possible types of environmental damage from mining are discussed in the "Earth's Minerals" chapter. Coal mining exposes minerals and rocks from underground to air and water at the surface. Many of these minerals contain the element sulfur, which mixes with air and water to make sulfuric acid, a highly corrosive chemical. If the sulfuric acid gets into streams, it can kill fish, plants, and animals that live in or near the water.

Oil

Oil is a liquid fossil fuel that is extremely useful because it can be transported easily and can be used in cars and other vehicles. Oil is currently the single largest source of energy in the world.

Oil Formation

Oil from the ground is called **crude oil**, which is a mixture of many different hydrocarbons. Crude oil is a thick dark brown or black liquid hydrocarbon. Oil also forms from buried dead organisms, but these are tiny organisms that live on the sea surface and then sink to the seafloor when they die. The dead organisms are kept away from oxygen by layers of other dead creatures and sediments. As the layers pile up, heat and pressure increase. Over millions of years, the dead organisms turn into liquid oil.

Oil Production

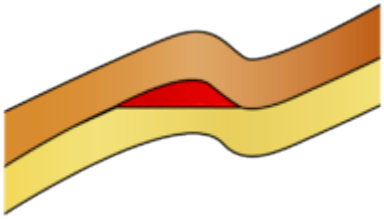


Figure 10. Oil (red) is found in the porous rock layer (yellow) and trapped by the impermeable

layer (brown). The folded structure has allowed the oil to pool so a well can be drilled into the reservoir.

In order to be collected, the oil must be located between a porous rock layer and an impermeable layer (figure 10). Trapped above the porous rock layer and beneath the impermeable layer, the oil will remain between these layers until it is extracted from the rock.

- Here is an animation of an oil deposit forming.
- The oil pocket is then drilled into from the surface. Here is an animation of an oil deposit being drilled.
- Sideways drilling allows a deposit that lies beneath land that cannot be drilled to be mined for oil.

To separate the different types of hydrocarbons in crude oil for different uses, the crude oil must be refined in refineries like the one shown in figure 11. Refining is possible because each hydrocarbon in crude oil boils at a different temperature. When the oil is boiled in the refinery, separate equipment collects the different compounds.



Figure 11. Refineries like this one separate crude oil into many useful fuels and other chemicals.

Oil Use

Most of the compounds that come out of the refining process are fuels, such as gasoline, diesel, and heating oil. Because these fuels are rich sources of energy and can be easily transported, oil provides about 90% of the energy used for transportation around the world. The rest of the compounds from crude oil are used for waxes, plastics, fertilizers, and other products.

Gasoline is in a convenient form for use in cars and other transportation vehicles. In a car engine, the burned gasoline mostly turns into carbon dioxide and water vapor. The fuel releases most of its energy as heat, which causes the gases to expand. This creates enough force to move the pistons inside the engine and to power the car.

Consequences of Oil Use

The United States does produce oil, but the amount produced is only about one-quarter as much as the nation uses. The United States has only about 1.5% of the world's proven oil reserves, and so most of the oil used by Americans must be imported from other nations.

The main oil-producing regions in the United States are the Gulf of Mexico, Texas, Alaska, and California. Most offshore drilling occurs in the Gulf of Mexico, but there are offshore platforms in California as well (Figure 12). Here is an animation of the location of petroleum basins in the contiguous United States.

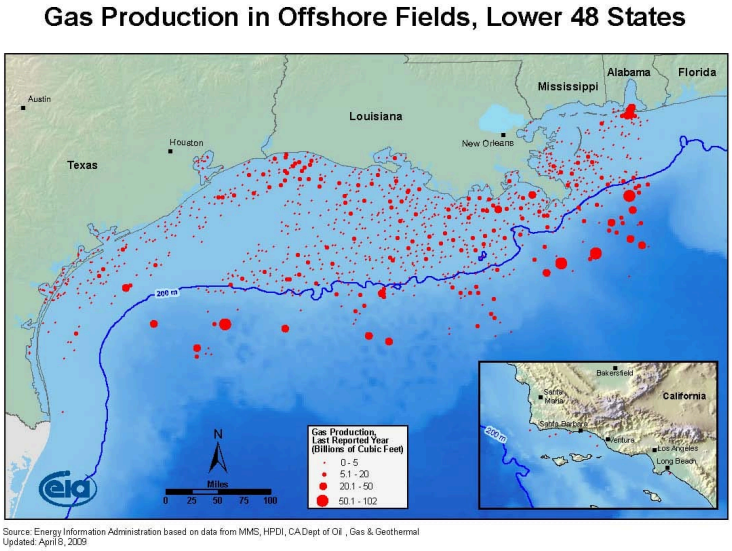


Figure 12. Offshore well locations in the Gulf of Mexico. Note that some wells are located in very deep water.

As in every type of mining, mining for oil has environmental consequences. Oil rigs are unsightly (figure 13), and spills are too common (figure 14).



Figure 13. Drill rigs at the San Ardo



Oil Field in Monterey, California. A deadly explosion on an oil rig in the Gulf of Mexico in April 2010 led to a massive oil spill. When this picture was taken in July 2010, oil was still spewing into the Gulf. The long-term consequences of the spill are being studied and are as yet unknown.

Natural Gas

Natural gas, often known simply as gas, is composed mostly of the hydrocarbon methane (refer to figure 5 for the structure).

Natural Gas Formation

Natural gas forms under the same conditions that create oil. Organic material buried in the sediments harden to become a shale formation that is the source of the gas. Although natural gas forms at higher temperatures than crude oil, the two are often found together. The formation of a minable oil and gas deposit is seen in this animation.

The largest natural gas reserves in the United States are in the Appalachian Basin, Texas, and the Gulf of Mexico region (Figure 15). California also has natural gas, found mostly in the Central Valley. In the northern Sacramento Valley and the Sacramento Delta, a sediment-filled trough formed along a location where crust was pushed together (an ancient convergent margin). Here is an animation of global natural gas reserves.

Figure 15. Gas production in the Lower 48 United States.

Natural Gas Use

Like crude oil, natural gas must be processed before it can be used as a fuel. Some of the chemicals in unprocessed natural gas are poisonous to humans. Other chemicals, such as water, make the gas less useful as a fuel. Processing natural gas removes almost everything except the methane. Once the gas is processed, it is ready to be delivered and used. Natural gas is delivered to homes for uses such as cooking and heating. Like coal and oil, natural gas is also burned to generate heat for powering turbines. The spinning turbines turn generators, and the generators create electricity.

Consequences of Natural Gas Use


Figure 16. A natural gas drill rig.

Natural gas burns much cleaner than other fossil fuels, meaning that it causes less air pollution. Natural gas also produces less carbon dioxide than other fossil fuels do for the same amount of energy, so its global warming effects are less (figure 16). See the pollution created by a car burning gasoline and a car burning natural gas in this animation.

Unfortunately, drilling for natural gas can be environmentally destructive. One technique used is hydraulic fracturing, also called fracking, which increases the rate of recovery of natural gas. Fluids are pumped through a borehole to create fractures in the reservoir rock that contains the natural gas. Material is added to the fluid to prevent the fractures from closing. The damage comes primarily from chemicals in the fracturing fluids. Chemicals that have been found in the fluids may be carcinogens (cancer-causing), radioactive materials, or endocrine disruptors, which interrupt hormones in the bodies of humans and animals. The fluids may get into groundwater or may runoff into streams and other surface waters.

Fossil Fuel Reserves

Fossil fuels provide about 85% of the world's energy at this time. Worldwide fossil fuel usage has increased many times over in the past half century (coal: 2.6x, oil: 8x, natural gas: 14x) because of population increases, because of increases in the number of cars, televisions, and other fuel-consuming uses in the developed world, and because of lifestyle improvements in the developing world. Past and predicted use of different types of energy in the United States can be seen in this animation.

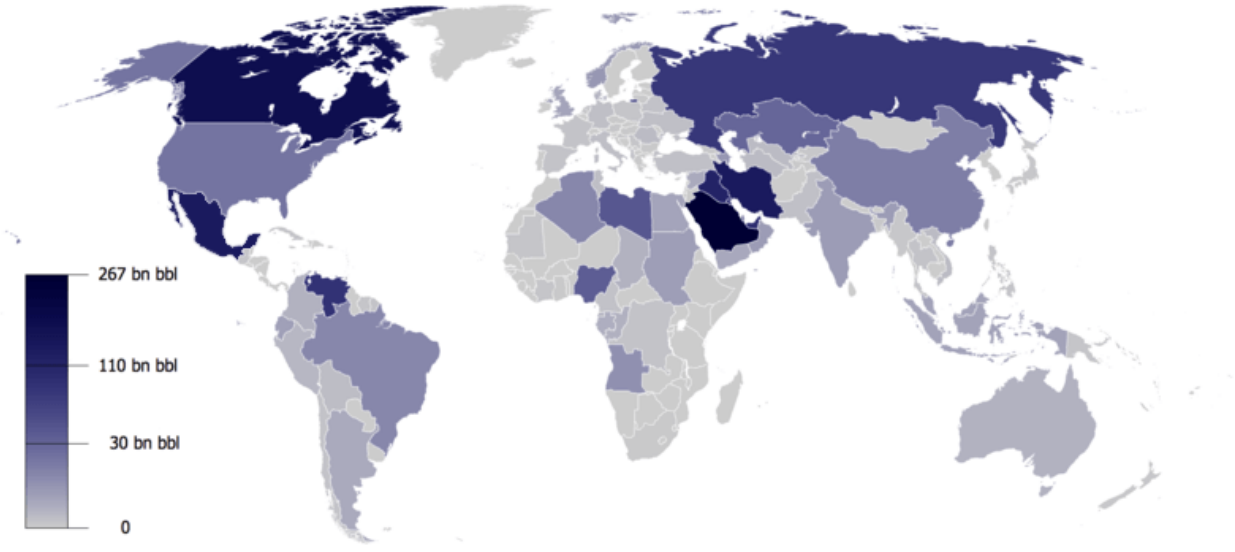


Figure 17. Worldwide oil reserves.

The amount of fossil fuels that remain untapped is unknown but can likely be measured in decades for oil and natural gas and in a few centuries for coal (figure 17). Alternative sources of fossil fuels, such as oil shales and tar sands, are increasingly being exploited (figure 18).



Figure 18. A satellite image of an oil-sands mine in Canada.

The environmental consequences of mining these fuels, and of fossil fuel use in general, along with the fact that these fuels do not have a limitless supply, are prompting the development of alternative energy sources.

Nuclear Energy

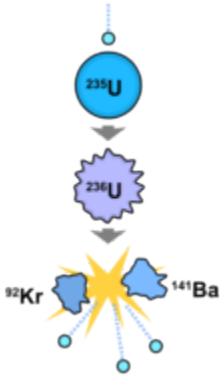


Figure 19. When struck by a tiny particle, Uranium-235 breaks apart and releases energy.

When the nucleus of an atom is split, it releases a huge amount of energy called **nuclear energy.** For nuclear energy to be used as a power source, scientists and engineers have learned to split nuclei and to control the release of energy (Figure 19).

Nuclear Energy Use

Nuclear power plants, such as the one seen in figure 20, use uranium, which is mined, processed, and then concentrated into fuel rods. When the uranium atoms in the fuel rods are hit by other extremely tiny particles, they split apart. The number of tiny particles allowed to hit the fuel rods needs to be controlled or they would cause a dangerous explosion. The energy from a nuclear power plant heats water, which creates steam and causes a turbine to spin. The spinning turbine turns a generator, which in turn produces electricity.



Figure 20. Nuclear power plants like this one provide France with almost 80% of its electricity.

Many countries around the world use nuclear energy as a source of electricity. In the United States, a little less than 20% of electricity comes from nuclear energy.

Consequences of Nuclear Power



Figure 21. Uranium mine in Kakadu National Park, Australia.

Nuclear power is clean. It does not pollute the air or release carbon dioxide. However, the use of nuclear energy does create other environmental problems. Uranium must be mined (figure 21). The process of splitting atoms creates radioactive waste, which remains dangerous for thousands or hundreds of thousands of years. As yet, there is no long-term solution for storing this waste.

The development of nuclear power plants has been on hold for three decades. Accidents at Three Mile Island and Chernobyl, Ukraine verified people's worst fears about the dangers of harnessing nuclear power (figure 22).



Figure 22. Damaged building near the site of the Chernobyl disaster.

Recently, nuclear power appeared to be making a comeback as society looked for alternatives to fossil fuels. But the 2011 disaster at the Fukushima Daiichi Nuclear Power Plant in Japan may have resulted in a new fear of nuclear power. The cause of the disaster was a 9.0 magnitude earthquake and subsequent tsunami, which compromised the plant. Although a total meltdown was averted, the plant experienced multiple partial meltdowns, core breaches, radiation releases, and cooling failures. The plant is scheduled for a complete cold shutdown before the end of 2011.

KQED: Nuclear Energy Use. Nuclear power is a controversial subject in California and most other places. Nuclear power has no pollutants including carbon emissions, but power plants are not always safe and the long-term disposal of wastes is a problem that has not yet been solved. The future of nuclear power is murky.

Renewable Energy Resources

Fossil fuels have the advantage of being cheap and transportable, but they cause environmental damage and will eventually run out. Renewable energy sources, by definition, will not run out, and most do not cause much pollution. But renewable energy sources do have a downside, too. Both the advantages and disadvantages of solar, water, wind, biomass, and geothermal energy will be described in this lesson.

Solar Power

The Sun is Earth's main source of energy, making the development of solar power a natural choice for an alternative energy source.

Solar Energy

Energy from the Sun comes from the lightest element, hydrogen, fusing together to create the second lightest element, helium. Nuclear fusion releases tremendous amounts of solar energy. The energy travels to the Earth, mostly as visible light. The light carries the energy through the empty space between the Sun and the Earth as **radiation**.

Solar Power Use

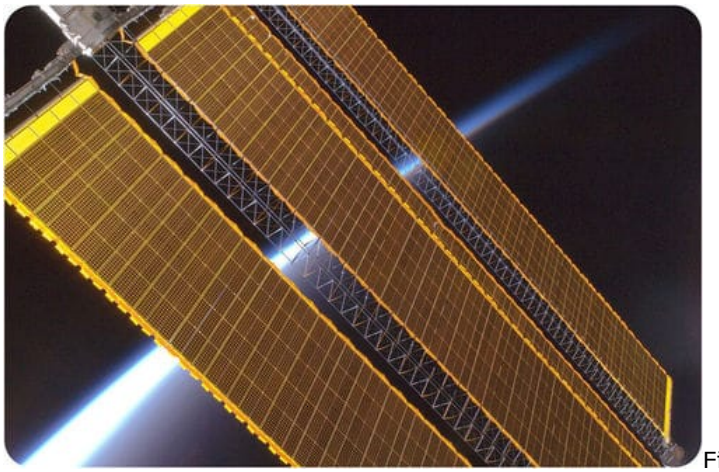


Figure 23. Solar panels supply power to the International Space Station.

Solar energy has been used for power on a small scale for hundreds of years, and plants have used it for billions of year. Unlike energy from fossil fuels, which almost always come from a central power plant or refinery, solar power can be harnessed locally (figure 23). A set of solar panels on a home's rooftop can be used to heat water for a swimming pool or can provide electricity to the house.

Society's use of solar power on a larger scale is just starting to increase. Scientists and engineers have very active, ongoing research into new ways to harness energy from the Sun more efficiently. Because of the tremendous amount of incoming sunlight, solar power is being developed in the United States in southeastern California, Nevada, and Arizona.

Solar power plants turn sunlight into electricity using a large group of mirrors to focus sunlight on one place, called a receiver (figure 24). A liquid, such as oil or water, flows through this receiver and is heated to a high temperature by the focused sunlight. The heated liquid transfers its heat to a nearby object that is at a lower temperature through a process called **conduction**. The energy conducted by the heated liquid is used to make electricity.



Figure 24. This solar power plant uses mirrors to focus sunlight on the tower in the center. The sunlight heats a liquid inside the tower to a very high temperature, producing energy to make electricity.

Here's a video of how solar energy can be concentrated so that it can be used for power.

Consequences of Solar Power Use



Figure 25. This experimental car is one example of the many uses that engineers have found for solar energy.

Solar energy has many benefits. It is extremely abundant, widespread, and will never run out. But there are problems with the widespread use of solar power.

- Sunlight must be present. Solar power is not useful in locations that are often cloudy or at night. However, storage technology is being developed.
- The technology needed for solar power is still expensive. An increase in interested customers will provide incentive for companies to research and develop new technologies and to figure out how to mass-produce existing technologies (figure 25).
- Solar panels require a lot of space. Fortunately, solar panels can be placed on any rooftop to supply at least some of the power required for a home or business.

Water Power

Water covers 70% of the planet's surface, and water power (hydroelectric power) is the most widely used form of renewable energy in the world. Hydroelectric power from streams provides almost one fifth of the world's electricity.

Hydroelectric Power

Remember that potential energy is the energy of an object waiting to fall. Water held behind a dam has a lot of potential energy. In a hydroelectric plant, a dam across a riverbed holds a stream to create a reservoir. Instead of flowing down its normal channel, the water is allowed to flow into a large turbine. As the water moves, it has kinetic energy, which makes the turbine spin. The turbine is connected to a generator, which makes electricity (figure 26).

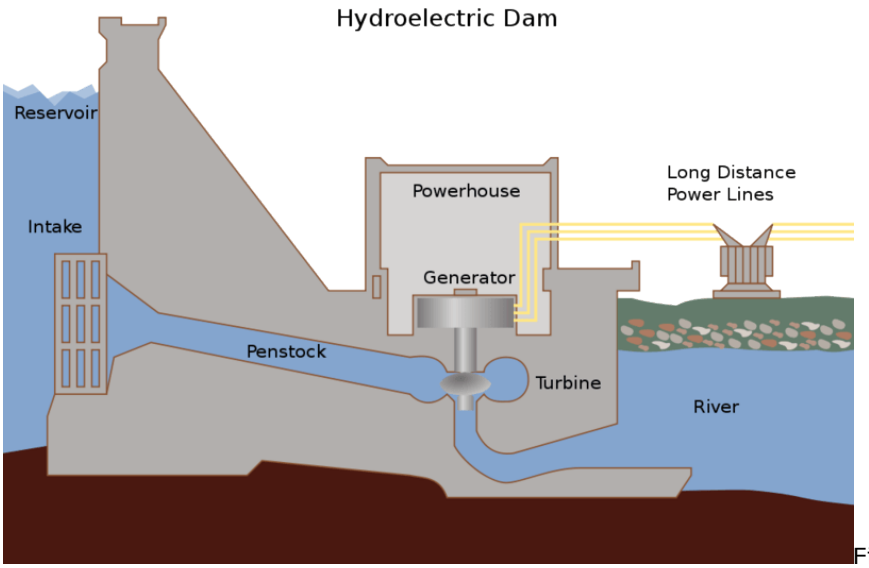


Figure 26. A cross-section of a hydroelectric plant.

Most of the streams in the United States and elsewhere in the developed world that are suitable for hydroelectric power have already been dammed (figure 27). In California, about 14.5% of the total electricity comes from hydropower. The state's nearly 400 hydropower plants are mostly located in the eastern mountain ranges where large streams descend down a steep grade.


Figure 27. Hydroelectric dams like this one use the power of moving water to create electricity.

Consequences of Water Power Use



Figure 28. Glen Canyon Dam in Arizona created Lake Powell. The dam was controversial because it flooded Glen Canyon, a beautiful desert canyon.

The major benefit of hydropower is that it generates power without releasing any pollution. Hydropower is also a renewable resource since the stream will keep on flowing. However, there are a limited number of suitable dam sites. Hydropower also has environmental problems. When a large dam disrupts a river's flow, it changes the ecosystem upstream. As the land is flooded by rising water, plants and animals are displaced or killed. Many beautiful landscapes, villages, and archeological sites have been drowned by the water in a reservoir (figure 28).

The dam and turbines also change the downstream environment for fish and other living things. Dams slow the release of silt so that downstream deltas retreat and seaside cities become dangerously exposed to storms and rising sea levels.

Ocean Water Power

The energy of waves and tides can be used to produce water power. Tidal power stations may need to close off a narrow bay or estuary. Wave power applications have to be able to withstand coastal storms and the corrosion of seawater. Because of the many problems with them, tide and wave power plants are not very common.

KQED: Harnessing Power from the Sea. Although not yet widely used, many believe tidal power has more potential than wind or solar power for meeting alternative energy needs. Quest radio looks at plans for harnessing power from the sea by San Francisco and along the northern California coast.

Wind Power

Wind power is the fastest growing renewable energy source in the world. Windmills are now seen in many locations, either individually or, more commonly, in large fields. Wind Powering America follows the development of wind power in the United States over the past several years.

Wind Energy

Energy from the sun also creates wind, which can be used as wind power. The sun heats different locations on Earth by different amounts. Air that becomes warm rises and then sucks cooler air into that spot. The movement of air from one spot to another along the ground creates wind. Since wind is moving, it has kinetic energy.

Wind Power Use



Figure 29. Wind turbines like the ones shown here turn wind into electricity without creating pollution.

Wind is the source of energy for wind power. Wind has been used for power for centuries. For example, windmills were used to grind grain and pump water. Sailing ships traveled by wind power long before ships were powered by fossil fuels. Wind can be used to generate electricity, as the moving air spins a turbine to create electricity (figure 29). This animation shows how wind power works.

Consequences of Wind Power

Wind power has many advantages. It does not burn, so it does not release pollution or carbon dioxide. Also, wind is plentiful in many places. Wind, however, does not blow all of the time, even though power is needed all of the time. Just as with solar power, engineers are working on technologies that can store wind power for later use.

Windmills are expensive and wear out quickly. A lot of windmills are needed to power a region, so nearby residents may complain about the loss of a nice view if a wind farm is built. Coastlines typically receive a lot of wind, but wind farms built near beaches may cause unhappiness for local residents and tourists.

The Cape Wind Project off of Cape Cod has been approved but is generating much controversy. Opponents are in favor of green power but not at that location. Proponents say that clean energy is needed and the project would supply 75% of the electricity needed for Cape Cod and nearby islands (figure 30).

Figure 30. Cape Wind off of Cape Cod in Massachusetts receives a great deal of wind (red color) but is also popular with tourists for its beauty.

California was an early adopter of wind power. Windmills are found in mountain passes where the cooler Pacific Ocean air is sucked through on its way to warmer inland valleys. Large fields of windmills can be seen at Altamont pass in the eastern San Francisco Bay Area, San Gorgonio Pass east of Los Angeles, and Tehachapi Pass at the southern end of the San Joaquin Valley.

Geothermal Power

Geothermal energy comes from heat deep below the surface of the Earth. Nothing must be done to the geothermal energy. It is a resource that can be used without processing.

Geothermal Energy

The heat that is used for geothermal power may come to the surface naturally as hot springs or geysers, like The Geysers in northern California. Where water does not naturally come to the surface, engineers may pump cool water into the ground. The water is heated by the hot rock and then pumped back to the surface for use. The hot water or steam from a geothermal well

spins a turbine to make electricity.



Figure 31. A geothermal energy plant in Iceland. Iceland gets about one fourth of its electricity from geothermal sources.

Geothermal energy is clean and safe. The energy source is renewable since hot rock is found everywhere in the Earth, although in many parts of the world the hot rock is not close enough to the surface for building geothermal power plants. In some areas, geothermal power is common (figure 31).

In the United States, California is a leader in producing geothermal energy. The largest geothermal power plant in the state is in the Geysers Geothermal Resource Area in Napa and Sonoma Counties, north of San Francisco. The source of heat is thought to be a large magma chamber lying beneath the area.

KQED: Geothermal Heats Up. Where Earth's internal heat gets close to the surface, geothermal power is a clean source of energy. In California, The Geysers supplies energy for many nearby homes and businesses.

A link to an interactive elements can be found at the bottom of this page.

Biomass



Figure 32. Biofuels, such as ethanol, are added to gasoline to cut down the amount of fossil fuels that are used.

Biomass is the material that comes from plants and animals that were recently living. Biomass can be burned directly, such as setting fire to wood. For as long as humans have had fire, people have used biomass for heating and cooking. People can also process biomass to make fuel, called **biofuel.** Biofuel can be created from crops, such as corn or algae, and processed for use in a car (figure 32). The advantage to biofuels is that they burn more cleanly than fossil fuels. As a result, they create less pollution and less carbon dioxide. Critics say, however, that the amount of energy, fertilizer, and land needed to produce the crops used make biofuels only a slightly better alternative than fossil fuels.

- KQED: How Green is Biomass Energy? Organic material, like almond shells, can be made into electricity. Biomass power is a great use of wastes and is more reliable than other renewable energy sources, but harvesting biomass energy uses energy and biomass plants produce pollutants including greenhouse gases.
- KQED: From Waste to Watts: Biofuel Bonanza. Cow manure can have a second life as a source of methane gas, which can be converted to electricity. Not only that food scraps can also be converted into green energy.
- KQED: Biofuels: Beyond Ethanol. To generate biomass energy, break down the cell walls of plants to release the sugars and then ferment those sugars to create fuel. Corn is a very inefficient source; scientists are looking for much better sources of biomass energy.
- Algae Power. Many people think that the best source of biomass energy for the future is algae. Compared to corn, algae is not a food crop, it can grow in many places, its much easier to convert to a usable fuel and its carbon neutral.
- Power Up with Leftovers. Food that is tossed out produces methane, a potent greenhouse gas. But that methane from leftovers can be harnessed and used as fuel. Sounds like a win-win situation.

Lesson Summary

- According to the law of conservation of energy, energy is neither created nor destroyed.
- Renewable resources can be replaced at the rate they are being used.
- Non-renewable resources are available in limited amounts or are being used faster than they can be replaced.
- Coal, oil, and natural gas are fossil fuels formed from the remains of living organisms.
- Coal is the largest source of energy for producing electricity.
- Oil and natural gas are important energy sources for vehicles and electricity generation.
- Nuclear energy is produced by splitting atoms. It also produces radioactive wastes that are very dangerous for many years.
- Solar energy, water power, wind power, geothermal energy, and biomass energy are renewable energy sources.
- Solar energy can be used either by passively storing and holding the Sun's heat, converting it to electricity, or concentrating it.
- There are many ways to use the energy of moving water, including hydroelectric dams and tidal and wave plants.
- Wind power uses the energy of moving air to turn turbines.
- Geothermal energy uses heat from deep within the earth to heat homes or produce steam that turns turbines.
- Biomass energy uses renewable materials such as wood or grains to produce energy.

Contributors and Attributions

CC licensed content, Shared previously

- Chapter 5: HS Earth's Energy. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/5.0/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.12.7: Earth's Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.12.8: Energy Consumption on Earth



What resources are in those electronics?

Everyone may realize that we use resources like trees, copper, water, and gemstones, but how many of us realize the tremendous variety of elements we need to make a single electronic device? A tablet computer with a touch screen contains many common chemical elements and a variety of rare earth elements.

Common Materials We Use from the Earth

People depend on natural resources for just about everything that keeps us fed and sheltered, as well as for the things that keep us entertained. Every person in the United States uses about 20,000 kilograms (40,000 pounds) of minerals every year for a wide range of products, such as cell phones, TVs, jewelry, and cars. Table 1 shows some common objects, the materials they are made from, and whether they are renewable or non-renewable.

Table 1. Common Objects We Use from the Earth

Common Object	Natural Resources Uses	Renewable or Non-Renewable?
Cars	15 different metals, such as iron, lead, and chromium to make the body.	Non-renewable
Jewelry	Precious metals like gold, silver, and platinum. Gems like diamonds, rubies, emeralds, turquoise.	Non-renewable
Electronic Appliances (TV's, computers, DVD players, cell phones, etc.)	Many different metals, like copper, mercury, gold.	Non-renewable

Table 1. Common Objects We Use from the Earth		
Clothing	Soil to grow fibers such as cotton. Sunlight for the plants to grow. Animals for fur and leather.	Renewable
Food	Soil to grow plants. Wildlife and agricultural animals.	Renewable
Bottled Water	Water from streams or springs. Petroleum products to make plastic bottles.	Non-renewable and Renewable
Gasoline	Petroleum drilled from wells.	Non-renewable
Household Electricity	Coal, natural gas, solar power, wind power, hydroelectric power.	Non-renewable and Renewable
Paper	Trees Sunlight Soil	Renewable
Houses	Trees for timber. Rocks and minerals for construction materials, for example, granite, gravel, sand.	Non-renewable and Renewable

Summary

- Many objects, such as a car, contain many types of resources.
- Resources may be renewable or non-renewable, and an object may contain some of each.
- Rare earth elements and other unusual materials are used in some electronic devices.

Practice

Use this resource to answer the questions that follow.

A link to an interactive elements can be found at the bottom of this page.

1. What do we use neodymium for?
2. What are rare earth elements used for in general?
3. Where do we get our REEs? Why are there signs that this can't continue?
4. Can we develop alternatives?
5. What is the problem with the deposit of REEs that is offshore of Japan?
6. What is the danger for the future?

Contributors and Attributions

CC licensed content, Shared previously

- Materials Humans Use. **Provided by**: CK-12. **Located at**: http://www.ck12.org/earth-science/Materials-Humans-Use/lesson/Materials-Humans-Use-HS-ES/. **License**: *CC BY-NC: Attribution-NonCommercial*

All rights reserved content

- Why We Need Rare Earth Elements. **Authored by**: DNews. **Located at**: https://youtu.be/xqoQfN9DgNs. **License**: *All Rights Reserved*. **License Terms**: Standard YouTube License

This page titled 4.12.8: Energy Consumption on Earth is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

4.12.9: Use and Conservation of Resources

Introduction

Natural resources may be living or non-living. Their value may be tangible, such as the price of an ounce of gold, or intangible, like the psychological value of being able to visit pristine natural areas. Some natural resources must be used and used wisely, but some must be preserved to maintain their value.

Mystery in the Forest

Like all forests, the Monongahela National Forest of West Virginia is an important natural resource. A forest is a resource in ways that are obvious and ways that are not so obvious. This forest is used for many things including:

- Recreation, such as hiking, camping, and picnicking.
- Habitat for many organisms, including nine endangered species and 50 species of rare plants.
- Streams [207 kilometers (129 miles)] for fishing, particularly trout fishing.
- Wildlife management areas for hunting deer, squirrels, turkeys, rabbits, mink, and foxes.
- Mineral and energy resources such as coal, gas, limestone, and gravel.
- Hardwood trees used for **timber**, which brings in over $7 million a year.

 Figure 1. The Monongahela National Forest in West Virginia contains many natural resources. Notice the air pollution that obscures the view.

But Monongahela National Forest has a problem; for several years, trees in the forest have not grown well. What are some reasons that trees might not grow well (figure 1)?

Scientists have been working for several years to solve the mystery. The scientists suspected that the soil is missing nutrients that the trees and other plants need to grow. Can you design an experiment that scientists could do to test this hypothesis? (There is a clue in the caption for figure 1.)

The scientists sampled the soil and tested it for important nutrients. They discovered that the soil has very low levels of plant nutrients, such as magnesium and calcium. Can you develop a hypothesis for why these nutrients might be missing from the soil? The scientists thought that air pollution from nearby factories had released chemicals into the environment that removed the nutrients from the soil and carried them away. How would the scientists test that hypothesis?

Scientists in the Monongahela National Forest are still researching the missing plant nutrients. They are trying to learn what they can do to help keep the nutrients in the soil so the trees will grow better.



Figure 2. We use Earth's resources for many purposes, including recreation and natural beauty.

Like the Monongahela National Forest, people use parts of the Earth for many reasons, such as food, water, building materials, timber, recreation, and energy (figure 2). As you've already learned, human activities can degrade natural resources, just like air pollution from factories is speeding up the loss of soil nutrients in West Virginia.

For natural resources to continue to be available, they need to be protected. We also need to **conserve** natural resources so they will last longer. When we practice conservation, we make sure resources will be available in the future, both for ourselves and for other organisms.

Renewable versus Non-Renewable Resources

In the Earth's Energy chapter, energy resources were classified as renewable or non-renewable. How do you think other natural resources, such as minerals and forests, are classified? Like energy resources, all natural resources are divided into renewable and non-renewable. Can you define these terms?



Figure 3. An old growth forest, like this tropical rainforest in Malaysia, is a complex ecosystem with many types of plants and animals. When a forest is destroyed by logging, it takes hundreds or thousands of years for the forest to regenerate.

Renewable resources can be regenerated or grown so rapidly that they reappear at the same rate or even faster as they are being used (figure 3). Are forests a renewable resource? Why are they a renewable resource? Why aren't they a renewable resource? Although new trees can grow to replace logged trees, their growth is often too slow for the trees to be of use for a long time. Loggers just move to a new area rather than wait for the forest to regenerate.

Other examples of resources that are renewable but not entirely renewable include soil, wildlife, and water. How do these

resources fit in both categories? Soil has a very slow renewal rate, so they are often non-renewable. Fish and other wildlife can reproduce and so are a renewable resource, yet it is possible to take so many of these creatures that the populations are not able to rebound, making them a non-renewable resource (figure 4). Organisms can be over-hunted, over-fished or have populations decline because of habitat loss so that their numbers go so low they are no longer a renewable resource.



Figure 4. Chimpanzees are eaten and taken as pets so their numbers in the wild are declining.

Non-renewable resources are resources that cannot be regenerated on a useful timescale. Fossil fuels and most minerals are non-renewable resources. We can (and eventually will) run out of these resources.

Resource Availability

From the table on the previous page you can see that many of the resources we depend on are non-renewable. Non-renewable resources vary in their availability; some are very abundant and others are rare. Materials, such as gravel or sand are technically non-renewable but are so abundant that running out is no issue. Some resources are truly limited in quantity: When they are gone, they are gone and something must be found that will replace them. There are even resources, such as diamonds and rubies, that are valuable in part because they are so rare.



Figure 5. Tampa Bay, Florida, has one of the few desalination plants in the United States.

Besides abundance, resource value is determined by how easy it is to locate and extract. If a resource is difficult to use, it will not be used until the price for that resource becomes so great that it is worth paying for. For example, the oceans are filled with an abundant supply of water, but desalination is costly, so it is used only where water is really limited (figure 5). As the cost of desalination plants comes down, more will likely be built.

Politics is also part of determining resource availability and cost. Nations that have a desired resource in abundance will often **export** that resource to other countries, while countries that need that resource must **import** it from one of the countries that produces it. This situation is a potential source of economic and political trouble.

Of course the greatest example of this is oil. Only 12 countries have approximately 80% of all of the world's oil (Figure 6). However, the biggest users of oil, the United States, China, and Japan, are all located outside this oil-rich region. This leads to a situation in which the availability and price of the oil is determined largely by one set of countries that have their own interests to look out for. The result has sometimes been war, which may have been attributed to all sorts of reasons, but at the bottom, the reason is oil.

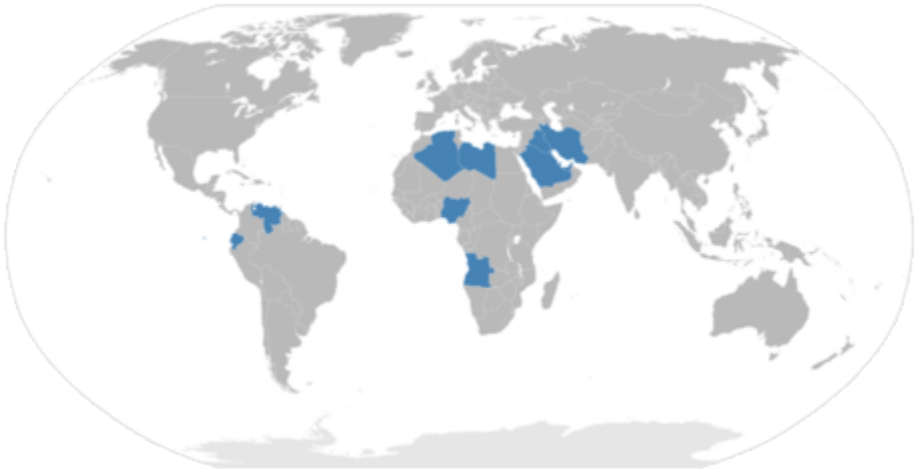


Figure 6. The nations in blue are the 12 biggest producers of oil; they are Algeria, Angola, Ecuador, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, the United Arab Emirates, and Venezuela.



Figure 7. Pollution from discarded materials degrades the environment and reduces the availability of natural resources.

The topic of overconsumption was touched on in the Ecosystems and Human Populations chapter. Many people in developed countries, such as the United States and most of Europe, use many more natural resources than people in many other countries. We have many luxury and recreational items, and it is often cheaper for us to throw something away than to fix it or just hang on to it for a while longer. This consumerism leads to greater resource use, but it also leads to more waste. Pollution from discarded materials degrades the land, air, and water (figure 7).

Natural resource use is generally lower in developing countries because people cannot afford many products. Some of these nations export natural resources to the developed world since their deposits may be richer and the cost of labor lower. Environmental regulations are often more lax, further lowering the cost of resource extraction.

Besides obtaining resources, we also dump waste on these nations. Many of our electronic wastes, which we think are being

recycled, end up in developing countries where they pose a problem for human health and the environment. Electronic wastes are sent to developing nations where people pick through them for valuable materials. These wastes contain many toxic compounds and are hazardous.

Conserving Natural Resources



Figure 8. Recycling can help conserve natural resources.

So that people in developed nations maintain a good lifestyle and people in developing nations have the ability to improve their lifestyles, natural resources must be conserved and protected (figure 8). People are researching ways to find renewable alternatives to non-renewable resources. Here is a checklist of ways to conserve resources:

- Buy less stuff (use items as long as you can, and ask yourself if you really need something new).
- Reduce excess packaging (drink tap water instead of water from plastic bottles).
- Recycle materials such as metal cans, old cell phones, and plastic bottles.
- Purchase products made from recycled materials.
- Reduce pollution so that resources are maintained.
- Prevent soil erosion.
- Plant new trees to replace those that are cut down.
- Drive cars less, take public transportation, bicycle, or walk.
- Conserve energy at home (turn out lights when they are not needed).

National Geographic videos found on this site in Environment Videos, Environmental Threats, Deforestation.

- “Sustainable Logging”

Or Environment Videos, Habitats, Rainforest.

- “Mamirarua” is a sustainable development reserve that is protecting the Amazon
- “Vancouver Rain Forest” explores an alliance between conservationists and logging companies

Or find ways to go green from National Geographic videos, Environment Videos, Going Green.

- The problem with plastic bags is discussed in this Conservation in action, “Edward Norton: Bag the Bag”
- Trying to mitigate problems caused by intensive logging in Ecuador while helping the people who live there improve their standards of living is in “Ecuador Conservation”

Lesson Summary

- We use natural resources for many things. Natural resources give us food, water, recreation, energy, building materials, and luxury items.
- Many resources vary in their availability throughout the world. Some are rare, difficult to get, or in short supply.
- Natural resources must be preserved and protected from pollution and overuse.

- Buying fewer new products and recycling will help to conserve resources.

Contributors and Attributions

CC licensed content, Shared previously

- 20.1: Use and Conservation of Resources. **Provided by**: CK-12. **Located at**: http://www.ck12.org/book/CK-12-Earth-Science-For-High-School/section/20.1/. **License**: *CC BY-NC: Attribution-NonCommercial*

This page titled 4.12.9: Use and Conservation of Resources is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Machele Kindle via source content that was edited to the style and standards of the LibreTexts platform.

INDEX

No index terms found.

This page titled Index is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled Glossary is shared under a Public Domain license and was authored, remixed, and/or curated by Machele Kindle.

DETAILED LICENSING

Detailed licensing information for this resource is not available at this time.

This page titled Physical Science: A Two-Semester Course is shared under a license and was authored, remixed, and/or curated by Machele Kindle.