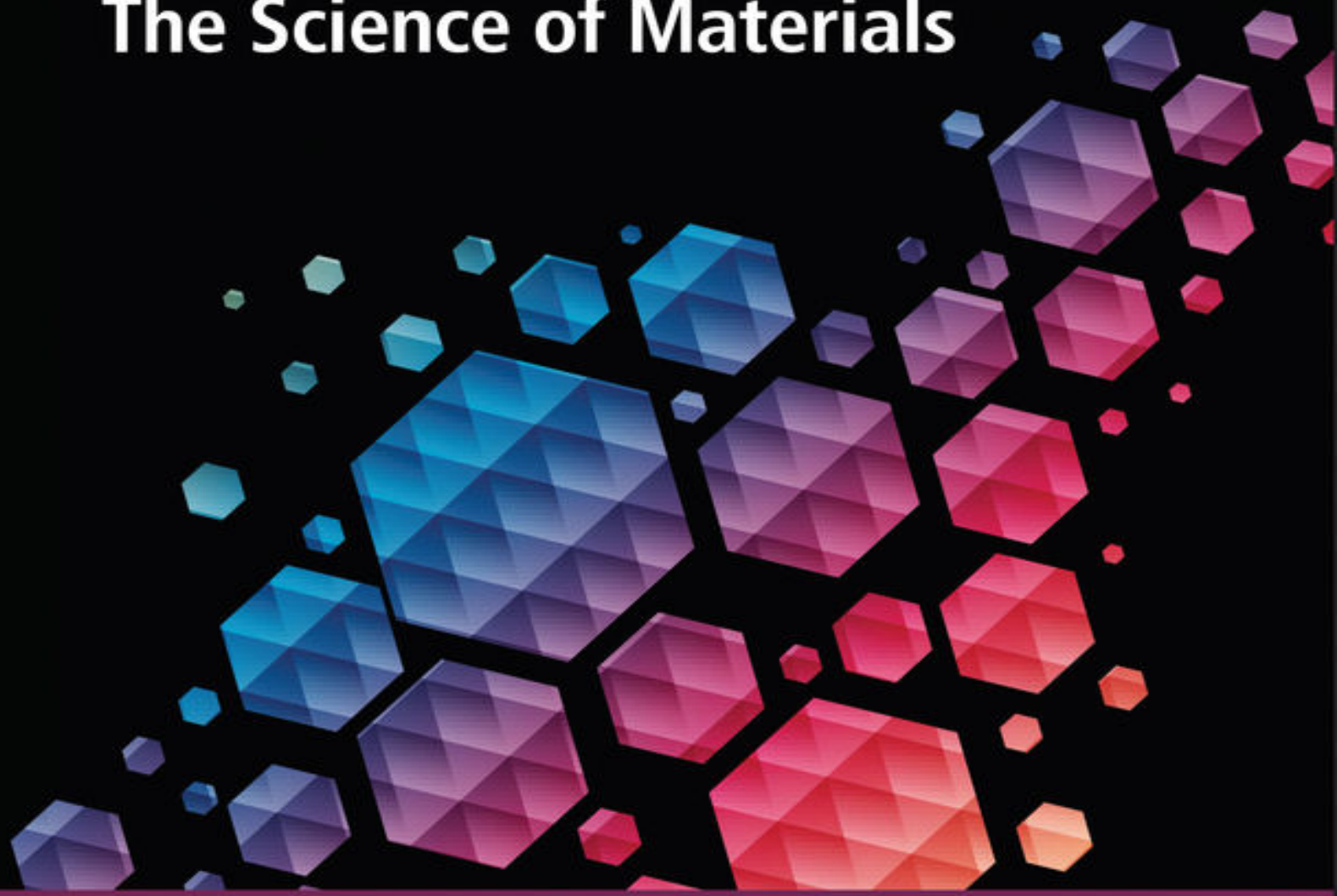
Richard J. D. Tilley

THIRD EDITION

Understanding Solids

The Science of Materials



with website

WILEY

Table of Contents

- Cover
- Title Page
- Copyright
- Preface
- About the Companion Website
- Part I: States of Aggregation
  - 1 Atoms and Bonding
    - 1.1 The Electron Structure of Atoms
    - 1.2 Ionic Bonding
    - 1.3 Covalent Bonding
    - 1.4 Metallic Bonding
    - 1.5 Weak Chemical Bonds
    - 1.6 Computation of Material Properties
    - Further Reading
    - Problems and Exercises
    - Note
  - 2 Microstructures and Phase Relationships
    - 2.1 Macrostructure, Microstructure, and Nanostructure
    - 2.2 The Development of Microstructures
    - 2.3 Phase Diagrams
    - 2.4 Ternary Systems
    - Reference
    - Further Reading
    - Problems and Exercises
  - 3 Crystal Structures and Defects
    - 3.1 Crystal Geometry
    - 3.2 Crystal Structures

3.3 Crystal Planes and Directions
3.4 Crystal Density
3.5 Structural Relationships
3.6 Point Defects
3.7 Linear, Planar, and Volume defects
Reference
Further Reading
Problems and Exercises
Notes
4 Solids: Overview
4.1 Metals
4.2 Crystalline Silicates and Inorganic Ceramic Materials
4.3 Silicate Glasses
4.4 Polymers and Organic Materials
4.5 Composite Materials
Reference
Further Reading
Problems and Exercises
Notes
Part II: Reactions and Transformations
5 Diffusion and Ionic Conductivity
5.1 Self-Diffusion and Tracer Diffusion
5.2 Non-steady-state and Steady-State Diffusion
5.3 Temperature Variation of Diffusion Coefficient
5.4 The Effect of Impurities
5.5 Random Walk Diffusion
5.6 Diffusion in Solids
5.7 Self-Diffusion in One Dimension
5.8 Self-Diffusion in Crystals

5.9 The Arrhenius Equation and Point Defects
5.10 Correlation Factors for Self-Diffusion
5.11 Ionic Conductivity
5.12 The Relationship Between Ionic Conductivity and Diffusion Coefficient
5.13 Superionic Conductors
Reference
Further Reading
Superionic Conductors: See Also References Therein:
Problems and Exercises
Note
6 Phase Transformations and Reactions
6.1 Sintering
6.2 Phase Transitions
6.3 Displacive and Reconstructive Transitions
6.4 Order–Disorder Transitions
6.5 Martensitic Transformations
6.6 Phase Diagrams and Microstructures
6.7 High Temperature Oxidation of Metals
6.8 Solid-State Reactions
Further Reading
Problems and Exercises
7 Oxidation and Reduction
7.1 Galvanic Cells
7.2 Chemical Analysis Using Galvanic Cells
7.3 Batteries
7.4 Corrosion
7.5 Electrolysis
7.6 Pourbaix Diagrams
Reference

Further Reading
Problems and Exercises
Notes
Part III: Physical Properties
8 Mechanical Properties of Solids
8.1 Strength and Hardness
8.2 Elastic Moduli
8.3 Deformation and Fracture
8.4 Time-Dependent Properties
8.5 Nanoscale Properties
8.6 Composite Materials
Further Reading
Problems and Exercises
9 Insulating Solids
9.1 Dielectrics
9.2 Piezoelectrics, Pyroelectrics, and Ferroelectrics
9.3 Ferroelectrics
Reference
Further Reading
Problems and Exercises
Note
10 Magnetic Solids
10.1 Magnetic Materials
10.2 Paramagnetic Materials
10.3 Ferromagnetic Materials
10.4 Antiferromagnetic Materials and Superexchange
10.5 Ferrimagnetic Materials
10.6 Nanostructures
10.7 Magnetic Defects
Further Reading

Problems and Exercises
Note
11 Electronic Conductivity in Solids
11.1 Metals
11.2 Semiconductors
11.3 Metal–Insulator Transitions
11.4 Conducting Polymers
11.5 Superconductivity
11.6 Nanostructures and Quantum Confinement of Electrons
Further Reading
Problems and Exercises
Notes
12 Optical Aspects of Solids
12.1 Light
12.2 Sources of Light
12.3 Refraction
12.4 Reflection
12.5 Scattering and Attenuation
12.6 Diffraction
12.7 Fibre Optics
12.8 Energy Conversion
12.9 Nanostructures
Further Reading
Problems and Exercises
Notes
13 Thermal Properties of Solids
13.1 Heat Capacity
13.2 Thermal Conductivity
13.3 Expansion and Contraction
13.4 Thermoelectric Effects

13.5 The Magnetocaloric Effect
13.6 Thermochromic Effects
References
Further Reading
Problems and Exercises
Note
Part IV: Nuclear Properties of Solids
14 Radioactivity and Nuclear Reactions
14.1 Radioactivity
14.2 Artificial Radioactive Atoms
14.3 Nuclear Decay
14.4 Nuclear Energy
14.5 Nuclear Waste
Further Reading
Problems and Exercises
Appendix A: Periodic Table
Appendix B: Energy Levels and Terms of Many-Electron Atoms
B.1 Derivation of Atomic Terms
B.2 The Ground State Term of an Atom
B.3 Energy Levels of Many Electron Atoms
Index
End User License Agreement

List of Tables

Chapter 1

Table 1.1 Orbitals of hydrogen-like atoms.

Table 1.2 Reduced Madelung constants,α.

Table 1.3 The geometry of some hybrid orbitals.

Table 1.4 The electron configurations of some homonuclear diatomic molecules.

Table 1.5 Bonds between atoms, ions, and molecules.

Chapter 3

Table 3.1 The crystal systems.

Table 3.2 Interplanar spacing and unit cell volume.

Table 3.3 Structures in terms of anion packing.

Table 3.4 The formation enthalpy of Schottky defects in alkali halide compoun…

Table 3.5 The formation enthalpy of Frenkel defects.

Table 3.6 The Kröger–Vink notation for defects in crystals.[a]

Chapter 4

Table 4.1 Allotropic structures of some metals.

Table 4.2 A summary of silicate structures.

Table 4.3 Some silicate glasses.

Table 4.4 Glass viscosity.

Table 4.5 Addition polymers.

Table 4.6 Typical Portland cement constituents.

Chapter 5

Table 5.1 Solutions of the diffusion equation.

Table 5.2 Values of the error function.[a]

Table 5.3 Some representative values for self-diffusion coefficients.[a]

Table 5.4 Approximate enthalpy values for the formation and movement of defec...

Table 5.5 Correlation factors for self-diffusion.

Chapter 7

Table 7.1 Standard reduction potentials at 25 °C.

Chapter 8

Table 8.1 Representative values of Young's modulus and Poisson's ratio.

Chapter 9

Table 9.1 The relative permittivity and refractive index of some crystals.

Table 9.2 Ferroelectrics and antiferroelectrics.

Chapter 10

Table 10.1 Calculated and observed magnetic dipole moments for the lanthanoid...

Table 10.2 The magnetic properties of the 3d-transition metal ions.

Table 10.3 Ferromagnetic and antiferromagnetic compounds.

Table 10.4 Magnetic moments and electron configurations of the ferromagnetic ...

Chapter 11

Table 11.1 Approximate values for the band gap of some semiconductors.

Table 11.2 Some cuprate superconducting oxides.

Table 11.3 Homologous series of cuprate superconductors.

Chapter 12

Table 12.1 The visible spectrum.

Table 12.2 Luminescence.

Table 12.3 Refractive indices.

Table 12.4 Refractive coefficients for some oxides.[a]

Table 12.5 The colour of a thin film in white light.

Chapter 13

Table 13.1 The Debye temperature and room temperature heat capacity of some e...

Table 13.2 Coefficients of thermal conductivity and thermal expansion.

Chapter 14

Table 14.1 Nuclides and particles in nuclear reactions.

Table 14.2 Periodic Table group seven elements.

Table 14.3 Half-lives of some radioisotopes.

List of Illustrations

Chapter 1

Figure 1.1 Electron energy levels of the H atom.

Figure 1.2 Schematic energy level diagram for a light, hydrogen-like atom.

Figure 1.3 The probability of finding an s electron at a distance *r* from the...

Figure 1.4 The boundary surfaces of the p orbitals: (a) p_x; (b) p_y; (c) p_z....

Figure 1.5 The boundary surfaces of the d orbitals: (a) d_{xy}; (b) d_{xz}; (c) d_y...

Figure 1.6 The evolution of the electron energy levels of an atom with a $3d^2$

Figure 1.7 Ionic radii, nm. Cation radii apply to ions octahedrally coordina...

Figure 1.8 The total potential energy between two monovalent ions as a funct...

Figure 1.9 Covalent bond formation: (a) two H atoms with opposed electron sp...

Figure 1.10 Covalent bonding: (a) σ bonding via overlap of s and p orbitals;...

Figure 1.11 (a) The 2s and 2p orbitals on C; (b) four sp^3 hybrid orbitals; (...

Figure 1.12 The formation of molecular orbitals in an H_2 molecule.

Figure 1.13 Schematic molecular orbital energy level diagram for homonuclear...

Figure 1.14 Molecular orbitals formed by a more metallic atom A and a less m...

Figure 1.15 The schematic development of energy bands from delocalised molec...

Figure 1.16 Schematic development of energy bands from atomic orbitals for m...

Figure 1.17 The energy levels and allowed wavelengths of an electron wave co...

Figure 1.18 The relationship between energy, E and wavenumber, k, for an ele...

Figure 1.19 The Fermi function, $P(E)$.

Figure 1.20 (a) The one-dimensional free-electron E vs. k curve broken into ...

Figure 1.21 The bonding in a molecule, M_2, and a solid metal.

Figure 1.22 Schematic development of energy bands from isolated atoms for so...

Figure 1.23 Schematic development of energy bands from isolated atoms for ge...

Figure 1.24 The boiling points of the alkanes, C_nH_{2n+2}, pl...

Figure 1.25 The Lennard-Jones potential between two atoms separated by a dis...

Chapter 2

Figure 2.1 The cubic unit cells of (a) tungsten; (b) copper.

Figure 2.2 Silicate glasses, schematic: (a) structure of an SiO_4 tetrahedron...

Figure 2.3 Polythene: chemical formula representation of (a) linear polythen...

Figure 2.4 Sheet-like nanomaterials: (a) graphene; (b) a single boron nitrid...

Figure 2.5 (a) C nanotube; (b) double tube arrangement, schematic.

Figure 2.6 Nanoparticles: (a) C_{60} buckyball; (b) Au core nanoparticle with N...

Figure 2.7 Electronic semiconductor nanostructures: (a) a quantum well; (b) ...

Figure 2.8 Polycrystalline solids: (a) granite; mica (black), quartz and fel...

Figure 2.9 The generalised form of a one-component phase diagram.

Figure 2.10 The phase diagram for water, schematic.

Figure 2.11 (a) The cooling curve generated as a sample of water is slowly c...

Figure 2.12 The copper (Cu)–nickel (Ni) phase diagram at 1 atm pressure.

Figure 2.13 Composition determination when two phases are present.

Figure 2.14 The lead (Pb)–tin (Sn) phase diagram at 1 atm pressure.

Figure 2.15 Composition determination in the Pb–Sn system.

Figure 2.16 The $CaSiO_3$ (wollastonite)–$CaAl_2O_4$ phase diagram at 1 atm pressur...

Figure 2.17 The V_2O_5–MgO phase diagram at 1 atm pressure.

Figure 2.18 The iron-rich region of the iron–carbon existence diagram. (The ...

Figure 2.19 (a) The general form of an isobaric ternary phase diagram; (b) R...

Figure 2.20 The simplified WO_3–WO_2–ZrO_2 phase diagram; (b) determination of ...

Figure 2.21 The $FeO–Fe_2O_3–TiO_2$ system at: (a) approximately 500 °C; (b) appr...

Figure 2.22 Approximate phase diagram for the $MgO–Al_2O_3–SiO_2$ system.

Chapter 3

Figure 3.1 The reference axes used to characterise the crystal systems.

Figure 3.2 The 14 Bravais lattices. (Note that the lattice points are exagge...

Figure 3.3 (a) A regular TiO_6 octahedron. (b) A regular SiO_4 tetrahedron.

Figure 3.4 The positions of an atom in a unit cell, specified as fractions o...

Figure 3.5 The face-centred cubic (fcc, A1) structure.

Figure 3.6 The body-centred cubic (bcc, A2) structure.

Figure 3.7 The hexagonal close-packed (hcp, A3) structure.

Figure 3.8 The structure of diamond: (a) unit cell; (b) cell viewed approxim...

Figure 3.9 The structure of graphite: (a) unit cell; (b) projection down the...

Figure 3.10 The halite (NaCl) structure: (a) perspective view; (b) projectio...

Figure 3.11 The cubic perovskite structure $SrTiO_3$.

Figure 3.12 The $MgAl_2O_4$ normal spinel structure: (a) projected down the **c**-ax...

Figure 3.13 Vegard's law, relating unit cell lattice parameters to compositi...

Figure 3.14 Miller indices of crystal planes: (a) (100); (b) (200); (c) (300...

Figure 3.15 Miller indices of crystals planes: (a) (110); (b) (101); (c) (01...

Figure 3.16 Miller indices of crystal planes: (a) (110) and ($\bar{1}$10); (b) (1$\bar{1}$0)...

Figure 3.17 Miller indices of crystal planes: (a) (111); (b) (1$\bar{1}$1).

Figure 3.18 Determination of Miller indices: (a), (b), (320); (c), (d), (3$\bar{1}$0...

Figure 3.19 Miller indices in hexagonal crystals.

Figure 3.20 (a) Directions in a lattice; (b) parallel directions have the sa...

Figure 3.21 Directions in the (001) plane of a hexagonal crystal.

Figure 3.22 Hexagonal close-packing (hcp) of spheres: (a) single layer; (b) ...

Figure 3.23 Hexagonal close-packing unit cell: (a) basal layer; (b) perspect...

Figure 3.24 Cubic close-packing (ccp) of spheres. The layers follow the sequ...

Figure 3.25 The cubic close-packing (ccp) unit cell: (a) layers A and B; (b)...

Figure 3.26 Sites between close-packed layers of spheres: (a), (b) tetrahedr...

Figure 3.27 Representations of tetrahedra: (a), (b) perspective view; (c), (...

Figure 3.28 Representations of octahedra: (a), (b) perspective view; (c), (d...

Figure 3.29 The structure of α-quartz (SiO_2) projected down the **c**-axis. The ...

Figure 3.30 (a) The cubic ReO_3 structure represented as corner-shared ReO_6 o...

Figure 3.31 Idealised point defects in a pure monatomic crystal: (a) vacancy...

Figure 3.32 The Gibbs energy of a pure crystal as a function of the number o...

Figure 3.33 Impurity or dopant point defects in a crystal, schematic: (a) su...

Figure 3.34 Point defects in an ionic crystal, formula MX, schematic: (a) Sc...

Figure 3.35 The structure of calcia-stabilised zirconia, schematic. The unit...

Figure 3.36 The layered structure of TiS_2, schematic: (a) without and (b) co...

Figure 3.37 (a) an edge dislocation in a metal crystal: (b), (c) a circuit a...

Figure 3.38 Screw dislocations: (a), (b) screw dislocation formed by displac...

Figure 3.39 Formation of dislocation loops: (a) the aggregation of vacancies...

Figure 3.40 Planar defects in crystals: (a) grain boundary; (b) twin plane; ...

Figure 3.41 A precipitate formed by the clustering of impurity atoms in a cr...

Chapter 4

Figure 4.1 The crystal structures and unit cell parameters (nm) of the metal...

Figure 4.2 Metallic radii, for 12-coordination.

Figure 4.3 The crystal structure of: (a) disordered CuAu; (b) ordered Cu_3Au;...

Figure 4.4 Octahedral (a) and tetrahedral (b) sites in the face-centred cubi...

Figure 4.5 Octahedral (a) and tetrahedral (b) sites in the hexagonal close-p...

Figure 4.6 Octahedral (a) and tetrahedral (b) sites in the body-centred cubi...

Figure 4.7 The glass forming region in the $(Fe_{1-x}Cr_xMo_y)_{75}C_{15}B_{10}$ phase...

Figure 4.8 Electron diffraction pattern from the quasicrystalline alloy Al_{72}

Figure 4.9 Icosahedra: (a) icosahedral geometry; (b) part of the As_3Co skutt...

Figure 4.10 Icosahedra, arranged with the same orientation, but not on a lat...

Figure 4.11 Linked $[SiO_4]$ units found in ionic silicates: (a) isolated $(SiO_4$

Figure 4.12 (a) Single (SiO_3) and (b) double (Si_4O_{11}) chains.

Figure 4.13 (a) Single sheet of corner-linked (SiO_4) tetrahedra; (b) the sam...

Figure 4.14 The structure of clays and related minerals: (a) single layer of...

Figure 4.15 The structures of some clay- and mica-related minerals formed fr...

Figure 4.16 (a) Truncated octahedron; (b) sodalite unit, with oxygen atoms o...

Figure 4.17 The structure of alumina, Al_2O_3: (a) projected close to [001]; (...

Figure 4.18 Cubic calcia-stabilised zirconia: (a) unit cell; (b) the structu...

Figure 4.19 The random network model of corner-linked (SiO_4) tetrahedra in a...

Figure 4.20 Specific volume vs. temperature curves: (a) crystalline solid wi...

Figure 4.21 Viscosity of a soda-lime glass: (a) viscosity vs. temperature; (...

Figure 4.22 Angel plot of fragility. The curves are for silica and soda-lime...

Figure 4.23 Crystal nucleation (a) and growth (b, c) in a glass ceramic, sch...

Figure 4.24 The polymerisation of ethylene, schematic: (a) a single monomer ...

Figure 4.25 (a, b) The structure of a tetrahedral CH_2 unit in a polymer chai...

Figure 4.26 The formation of the polyester PET, schematic.

Figure 4.27 Formation of nylon, schematic: (a) nylon 66; (b) nylon 6. The ca...

Figure 4.28 Epoxy resins: (a) epoxide group; (b) a small diepoxy molecule; (...

Figure 4.29 Reaction of molecules to give cross-linked polymers.

Figure 4.30 Molar mass of a polymer. (a) Idealised distribution: M_n, number ...

Figure 4.31 Polymer chain geometries, schematic: (a) linear; (b) branched; (...

Figure 4.32 (a) Polymer chains in polyethylene, schematic; (b) folded chains...

Figure 4.33 Schematic structure of a spherulite in a polymer such as polythe...

Figure 4.34 Polymer chains: (a) perspective view of an isotactic polymer cha...

Figure 4.35 Schematic polymer geometry: (a) alternating copolymer; (b) rando...

Figure 4.36 Elastomer, schematic: (a) unstretched; and (b) stretched.

Figure 4.37 Rubber: (a) isoprene; (b) bond redistribution and polymerisation...

Figure 4.38 The formation of cross-links between molecules as a result of vu...

Figure 4.39 Synthetic elastomers: (a) neoprene; and (b) nitrile rubber.

Figure 4.40 Polymerisation mechanisms: (a) step growth; (b) chain growth, A*...

Figure 4.41 The generation of free radicals (*) by bond breaking: (a) an azi...

Figure 4.42 Chain termination: (a) two chains meeting; (b) disproportionatio...

Figure 4.43 Metallocene catalysis: (a) the C atom framework of $[C_5H_5]^-$

Figure 4.44 MOFs: (a) schematic structure, consisting of metal nodes and org...

Figure 4.45 Covalent organic frameworks: (a) DBA; (b) HHTP; (c) MOF-5; (d) D...

Figure 4.46 The basic tetrahedral unit found in zeolitic imidazolate framewo...

Figure 4.47 Fibre-reinforced composites: (a) aligned fibres; (b) random orie...

Figure 4.48 The approximate relative strengths of the components of Portland...

Chapter 5

Figure 5.1 Self-diffusion of atoms in a solid: (a) initial state; (b) after ...

Figure 5.2 Diffusion couple formed by two MgO crystals separated by a film o...

Figure 5.3 Diffusion profiles: (a), (b), (c), gradually increasing heating t...

Figure 5.4 The straight-line graph of ln c vs. x^2 from a diffusion experimen...

Figure 5.5 Geometries for non-steady-state diffusion: (a), (b), the concentr...

Figure 5.6 Steady-state diffusion: (a) gas diffusing through a porous solid;...

Figure 5.7 Arrhenius plot of ln D vs. $1/T$.

Figure 5.8 A form of the Arrhenius equation found for crystals with lower (l...

Figure 5.9 Computed random walk in three-dimensions.

Figure 5.10 One-dimensional random walks of 100 steps of length a. The total...

Figure 5.11 The probability of a diffusing atom arriving at a point ma after...

Figure 5.12 Self-diffusion mechanisms: v, vacancy; i, interstitial; iy, inte...

Figure 5.13 Impurity diffusion mechanisms; v, vacancy; e, exchange; r, ring;...

Figure 5.14 Diffusion mechanisms: (a) a crystal containing Schottky defects;...

Figure 5.15 (a) Atom diffusion from one stable position to another; (b) The ...

Figure 5.16 Two adjacent planes in a crystal, containing n_1 and n_2 diffusing...

Figure 5.17 The variation of Arrhenius plots with impurity content. D, diffu...

Figure 5.18 Correlated motion during vacancy diffusion: (a) a vacancy can ju...

Figure 5.19 Energy barriers to be surmounted by an ion during ionic conducti...

Figure 5.20 Arrhenius plot of ln (σT) vs. $1/T$ for ionic conductivity in an a...

Figure 5.21 The structure of α-AgI: (a) body-centred cubic arrangement of I^-...

Figure 5.22 (a) Idealised β-alumina structure projected down [110] drawn as ...

Figure 5.23 Possible correlated jumps to move an Na^+ pair in (a) to that...

Figure 5.24 The fluorite structure: (a) projection of the structure down the...

Figure 5.25 The tunnel framework in NASICON phases.

Chapter 6

Figure 6.1 Initial stages in sintering: (a) Initial compact of slightly unev...

Figure 6.2 Sintering via vapour transport (a, b) does not cause shrinkage. S...

Figure 6.3 (a) Material transfers from hills to valleys during vapour-phase ...

Figure 6.4 Definitions of (a) neck radius, x, and (b) sphere radius, r, used...

Figure 6.5 First-order transitions: (a) variation of molar Gibbs energy with...

Figure 6.6 Second-order transitions: (a) variation of molar Gibbs energy wit...

Figure 6.7 Displacive transitions, schematic: (a) undistorted octahedra; (b)...

Figure 6.8 Displacive transition in La_2CuO_4: (a) The idealised structure of ...

Figure 6.9 Displacive transition in $CaTiO_3$: (a) room temperature structure; ...

Figure 6.10 A reconstructive transition, in which components are rearranged ...

Figure 6.11 A topotactic reconstructive transformation in which the ...ABAB... p...

Figure 6.12 The idealised structure of (a) TiS_2; (b) phases with composition...

Figure 6.13 Orientational transformations: (a, b) alternative orientations o...

Figure 6.14 (a) Two unit cells of the cubic face-centred cubic structure, th...

Figure 6.15 Schematic crystallographic changes occurring during the formatio...

Figure 6.16 (a) The unconstrained transformation of austenite (γ phase) to m...

Figure 6.17 The shape-change caused by deformation (a, b) can be minimised b...

Figure 6.18 Hysteresis (schematic) between martensite start and finish tempe...

Figure 6.19 The structural relationship between the forms of zirconia.

Figure 6.20 Part of the phase diagram of the Ti–Ni system.

Figure 6.21 The CsCl structure of the high-temperature form of TiNi: (a) cub...

Figure 6.22 (a) A stack of unit cells of TiNi projected down [110]. Shear di...

Figure 6.23 The sequence of events taking place during deformation and shape...

Figure 6.24 (a–e) The progressive recovery of the original shape of a deform...

Figure 6.25 Cooling a sample of composition *c* in the Cu–Ni system: (a) liqui...

Figure 6.26 Microstructures of a solidified Cu–Ni allot: (a) slow cooling; (...

Figure 6.27 The microstructure of solidified Pb–Sn alloys: (a) the Pb–Sn pha...

Figure 6.28 The microstructure of solidified Pb–Sn alloys: (a) the Pb–Sn pha...

Figure 6.29 The microstructure of solidified Pb–Sn alloys: (a) the Pb–Sn pha...

Figure 6.30 (a) The existence diagram of the Fe–C system. (b) Microstructure...

Figure 6.31 (a) The existence diagram of the Fe–C system. (b) Microstructure...

Figure 6.32 (a) The existence diagram of the Fe–C system. (b) Microstructure...

Figure 6.33 Oxide formation on a metal surface, schematic: (a) physical abso...

Figure 6.34 Growth of an oxide film on a metal surface: (a) diffusion of met...

Figure 6.35 Mechanisms of spinel formation: (a) initial state; (b) diffusion...

Figure 6.36 Markers used to determine the mechanism of $MgAl_2O_4$ formation: (a...

Figure 6.37 Photoresists, schematic: (a) negative photoresist; (b) positive ...

Figure 6.38 C_{120} dimers produced by milling C_{60} monomers.

Chapter 7

Figure 7.1 The components of a galvanic cell.

Figure 7.2 Daniell cell, schematic.

Figure 7.3 (a) A glass electrode for the measurement of pH. (b) Practical ar...

Figure 7.4 Arrangement of an ion selective electrode, schematic. (b) Practic...

Figure 7.5 The ‘rocking chair’ operation of a secondary battery, schematic: ...

Figure 7.6 Section through a zinc-carbon battery, schematic.

Figure 7.7 Section through an alkaline battery, schematic.

Figure 7.8 Section through a lithium coin cell primary battery, schematic.

Figure 7.9 Li-air battery, schematic

Figure 7.10 Fuel cells, schematic: AFC, alkaline fuel cell; PEMFC, proton ex...

Figure 7.11 Solid oxide fuel cells: (a) planar-design; (b) tubular design; (...

Figure 7.12 A single cell in a lead-acid battery, schematic.

Figure 7.13 The original configuration of a Li-ion cell in discharge operati...

Figure 7.14 Li-ion cell electrode reactions: (a, b) anode: the C layer stack...

Figure 7.15 The advance and retreat operation of a dual-ion secondary batter...

Figure 7.16 Dissimilar metal corrosion of a steel rivet in contact with copp...

Figure 7.17 Protective galvanizing: (a) a perforation in the zinc coating ca...

Figure 7.18 Differential aeration leading to pitting in steel: (a) initial s...

Figure 7.19 Crevice corrosion due to differential aeration: (a) initial situ...

Figure 7.20 The components of an electrolytic cell.

Figure 7.21 Nickel electroplating: the nickel anode is dissolved and transpo...

Figure 7.22 The FCC Cambridge cell for the production of titanium, schematic...

Figure 7.23 Pourbaix diagram showing the stability field of water.

Figure 7.24 Pourbaix diagram showing the stable species in the iron–water–ox...

Figure 7.25 Generalised Pourbaix diagram showing the stable species in a sys...

Figure 7.26 Simplified version of a Pourbaix diagram showing the range of co...

Chapter 8

Figure 8.1 Mechanical loading of solids: (a) tension; (b) compression; (c) s...

Figure 8.2 A rod in tension: (a) initial state; (b) final state.

Figure 8.3 Three-point bend testy for ceramic samples: (a) sideways view; (b...

Figure 8.4 Engineering stress-strain curves, schematic: (a) brittle and slig...

Figure 8.5 (a) Young's modulus; (b) secant modulus; (c) tangent modulus.

Figure 8.6 Toughness of a solid: curve C represents the toughest solid.

Figure 8.7 Engineering stress–strain curve, schematic, for a Nitinol shape-m...

Figure 8.8 Hardness indenters.

Figure 8.9 Schematic relationship between Vickers hardness and shear modulus...

Figure 8.10 Elastic moduli: (a) initial dimensions; (b) specimen after appli...

Figure 8.11 (a) The lattice potential energy *U* as a function of the atomic s...

Figure 8.12 Elastic moduli: (a) initial state; (b) after application of shea...

Figure 8.13 Waves in solids, schematic: (a) longitudinal waves; (b) transver...

Figure 8.14 The pulse-echo technique for the determination of elastic moduli...

Figure 8.15 Fracture surfaces: (a) cleavage fracture of single crystal silic...

Figure 8.16 Stress in a solid at: (a) an elliptical pore; (b) an elliptical ...

Figure 8.17 Plastic deformation: (a) the permanent distortion of the materia...

Figure 8.18 Slip: (a, b) slip following the application of a shear stress; (...

Figure 8.19 Application of shear stress to a crystal containing an edge disl...

Figure 8.20 (a) The (111) plane in a crystal of a metal with the face-centre...

Figure 8.21 A (110) slip-plane in the halite (NaCl) structure.

Figure 8.22 Schematic engineering stress-strain curve for a carbon steel.

Figure 8.23 Deformation of a semi-crystalline polymer: (a) unstressed state;...

Figure 8.24 Metal ductility: (a) schematic engineering stress vs. engineerin...

Figure 8.25 (a) A fractured semi-ductile steel tensile test specimen showing...

Figure 8.26 Polymer elongation: (a) schematic engineering stress vs. enginee...

Figure 8.27 Dislocation multiplication at a Frank–Read source: a dislocation...

Figure 8.28 Stress cycles: (a) irregular; (b) positive sinusoidal cycle; (c)...

Figure 8.29 Stress amplitude plotted against log (number of cycles to failur...

Figure 8.30 Fatigue failure: (a) stage 1, initial crack formation; (b) stage...

Figure 8.31 A normal creep curve for a solid subjected to constant load at a...

Figure 8.32 The effect of temperature on creep.

Figure 8.33 Creep mechanisms: (a) Coble creep: atoms diffuse along grain bou...

Figure 8.34 The structures of solid lubricants: (a) graphite; (b) graphite f...

Figure 8.35 An auxetic network in which the links between elements of the st...

Figure 8.36 A solid composed of rigid blocks of crystal linked by bonds of c...

Figure 8.37 Auxetic β-brass: (a) the CsCl structure of β-brass; strong bonds...

Figure 8.38 Nanoindentation: (a) indentation of a thin film; (b) schematic u...

Figure 8.39 Aligned fibre composites: (a) longitudinal load; (b) transverse ...

Figure 8.40 Layered materials: (a) longitudinal (parallel) tension; (b) tran...

Figure 8.41 Young's modulus of a composite, schematic: (a) configuration of ...

Chapter 9

Figure 9.1 The range of electronic conductivity in solids.

Figure 9.2 Capacitors: (a) charges will accumulate on metal plates due to an...

Figure 9.3 In an applied electric field, $\mathbf{E}_0$, an insulator gains a surface ch...

Figure 9.4 Polarisation: (a) electronic; (b) ionic; (c) orientational. Dipol...

Figure 9.5 The contribution of dipoles, ions and electrons to the polarisabi...

Figure 9.6 Schematic relationship between insulating solids: **E** is an applied...

Figure 9.7 The relationship between point-group symmetry and piezoelectric a...

Figure 9.8 The structure of hexagonal ZnS (wurtzite): (a) a ZnS_4 tetrahedron...

Figure 9.9 Piezoelectricity: (a) a force applied to a tetrahedron along a bo...

Figure 9.10 Piezoelectricity in quartz: (a) Part of the structure of room-te...

Figure 9.11 Quartz crystals: (a) section through a crystal normal to the **c**-a...

Figure 9.12 (a) dipoles present in a tetrahedral unit of PVF, poly(vinyl flu...

Figure 9.13 The electric dipoles present in (a) chains of nylon 6; (b) chain...

Figure 9.14 The crystal structures, schematic, of two forms of PVF_2 viewed d...

Figure 9.15 The α-helix, schematic: (a) the succession of carbon and nitroge...

Figure 9.16 Ferroelectric crystal formation, schematic: (a) a cation-centred...

Figure 9.17 Variation of potential energy vs. position for cation displaceme...

Figure 9.18 Two domains formed by differing alignment of electric dipoles in...

Figure 9.19 Antiparallel arrangement of electric dipoles in an antiferromagn...

Figure 9.20 Hysteresis behaviour of the polarisation, **P**, vs. applied electri...

Figure 9.21 Domain growth in ferroelectric $BiFeO_3$: (a) (001) single domain s...

Figure 9.22 Ferroelectric behaviour near to the Curie temperature: (a) first...

Figure 9.23 The Curie–Weiss behaviour of a ferroelectric above the Curie tem...

Figure 9.24 Hydrogen bonds (dashed) in ferroelectrics: (a, b) below the Curi...

Figure 9.25 Ferroelectric KH_2PO_4: (a) skeleton of the room temperature struc...

Figure 9.26 Ferroelectricity in sodium nitrite, $NaNO_2$: (a) the planar nitrit...

Figure 9.27 The barium titanate, $BaTiO_3$, structure: (a) cubic paraelectric h...

Figure 9.28 The phase diagram of the $PbZrO_3$–$PbTiO_3$ (PZT) system. The space g...

Figure 9.29 Relaxor ferroelectric properties, schematic: (a) variation of re...

Figure 9.30 A ferroelectric superlattice composed of $m = 2$ unit cells of SrT...

Figure 9.31 The flexoelectric effect: (a) evolution of the potential energy ...

Chapter 10

Figure 10.1 Magnetic materials in a Gouy balance: (a) electromagnet off; (b)...

Figure 10.2 Weak magnetic materials in an external field: (a) no solid prese...

Figure 10.3 Behaviour of a diamagnetic solid (a) in no field; (b) in an appl...

Figure 10.4 (a, b) Behaviour of a paramagnetic solid in an applied magnetic ...

Figure 10.5 (a) Magnetic dipole ordering in a ferromagnetic solid; (b) tempe...

Figure 10.6 (a) Magnetic dipole ordering in an antiferromagnetic solid; (b) ...

Figure 10.7 (a) Magnetic dipole ordering in a ferrimagnetic solid; (b) tempe...

Figure 10.8 Ordered arrays of magnetic dipoles arranged on a cubic lattice: ...

Figure 10.9 Cations in (a) an anion octahedron; (b) an anion tetrahedron.

Figure 10.10 The crystal field splitting, Δ (tet) or 10 Dq (tet) and Δ (oct)...

Figure 10.11 The electron configurations possible for d^n cations in an octah...

Figure 10.12 Splitting of atomic energy levels in a magnetic field.

Figure 10.13 The density of states for electrons in a metal: (a) in the abse...

Figure 10.14 Variation of the relative spontaneous magnetisation $\mathbf{M}/\mathbf{M}_s$ as a f...

Figure 10.15 The Bethe–Slater curve for the magnitude of the exchange ...

Figure 10.16 Schematic diagram of energy band overlap of 3d, 4s, and 4p orbi...

Figure 10.17 Ferromagnetic domains: (a) Weiss domains, schematic; (b) domain...

Figure 10.18 A domain (Bloch) wall between two magnetic domains, schematic....

Figure 10.19 The **B**–**H** (hysteresis) loop of a ferromagnetic solid.

Figure 10.20 Schematic **B**–**H** loops for soft and hard magnetic solids.

Figure 10.21 (a) The room temperature face-centred cubic structure of Mn_3Pt;...

Figure 10.22 Superexchange between Ni^{2+} and O^{2-} (schematic) leadin...

Figure 10.23 Schematic representation of the magnetic spin arrangement in Ba...

Figure 10.24 Double exchange (electron transfer) between Fe^{2+} and Fe^{3+}...

Figure 10.25 (a) Films less than a domain wide have magnetic dipoles aligned...

Figure 10.26 A single sheet of the CrX_3 structure (X = Cr, Br, I).

Figure 10.27 Part of an $m = n = 3$ magnetic $LaMnO_3/LaNiO_3$ superlattice in a (...

Figure 10.28 Photoinduced spin crossover in transition metal cations: (a) pr...

Figure 10.29 Bound magnetic polarons, schematic: (a) one bound magnetic pola...

Figure 10.30 Possible electron spin configurations for $Co^{3\pm}$ ($3d^6$) cation...

Figure 10.31 Double exchange between $Mn^{3\pm}$ and $Mn^{4\pm}$ high spin (HS) ca...

Figure 10.32 Electron configurations possible for Mn cations in an octahedra...

Chapter 11

Figure 11.1 Schematic energy band representations of materials: (a) insulato...

Figure 11.2 Free electrons in a one-dimensional metal: (a) energy vs. wave v...

Figure 11.3 The circular Fermi surface of a two-dimensional metal, dots repr...

Figure 11.4 The variation of the resistivity of a metal with temperature, sc...

Figure 11.5 The variation of the resistivity of alloys with concentration of...

Figure 11.6 The variation of resistivity vs. temperature for an intrinsic se...

Figure 11.7 Impurities in a crystal of silicon: (a) donor (P) atoms; (b) don...

Figure 11.8 The position of the Fermi energy, E_F, in semiconductors: (a) int...

Figure 11.9 The position of the Fermi energy, E_F, in a semiconductor at ordi...

Figure 11.10 The temperature variation of the position of the Fermi energy,

Figure 11.11 Arrangement of the four-point probe method of measurement of re...

Figure 11.12 Measurement of the Hall effect, schematic: (a) reference axes; ...

Figure 11.13 The p-n junction schematic: (a) energy bands of separate p-type...

Figure 11.14 Band structure across a p–n junction: (a) under forward bias; (...

Figure 11.15 Ideal current–voltage curve for a p–n junction.

Figure 11.16 The currents flowing across a p–n junction under forward bias....

Figure 11.17 Weakly interacting atoms: (a) ground state with one localised e...

Figure 11.18 The (slightly idealised) structure of Ca_2RuO_4 projected down [0...

Figure 11.19 Part of a conjugated hydrocarbon molecule (schematic) in which ...

Figure 11.20 The structures of polyacetylenes, schematic: (a) acetylene (eth...

Figure 11.21 Conjugated polymers: (a) a chain of isolated half-occupied p or...

Figure 11.22 Distortion of a polymer chain: (a) delocalized orbitals lead to...

Figure 11.23 Variation of the conductivity of polyacetylene with iodine dopa...

Figure 11.24 Variation of the conductivity of polyacetylene doped with iodin...

Figure 11.25 A soliton in *trans*-polyacetylene, schematic. The shaded ellipse...

Figure 11.26 Superconducting transition temperatures.

Figure 11.27 The Meissner effect: (a) normal metal or superconducting metal ...

Figure 11.28 The variation of superconductivity with external magnetic field...

Figure 11.29 A Type II superconductor in the mixed state.

Figure 11.30 The temperature–current–magnetic field phase space delimiting s...

Figure 11.31 Cooper pairs: (a) a single electron passing through a supercond...

Figure 11.32 A dc SQUID: (a) the circuit consists of a loop of superconducto...

Figure 11.33 The crystal structure of La_2CuO_4. The **c**-axis is vertical.

Figure 11.34 The phase region supporting superconductivity for $La_{2-x}Sr$

Figure 11.35 Comparison of the structures of (a) $LaCuO_4$ and (b) Nd_2CuO_4.

Figure 11.36 The crystal structures of: (a) $YBa_2Cu_3O_7$; (b) $YBa_2Cu_3O_6$.

Figure 11.37 The variation of the superconducting transition temperature T_c ...

Figure 11.38 Superconducting sheets found in cuprate superconductors: (a) a ...

Figure 11.39 Doping into the charge reservoir layers of a cuprate supercondu...

Figure 11.40 Idealised structures of the cuprate superconductors: (a) Bi_2Sr_2

Figure 11.41 Quantum wells: (a) a single quantum well in gallium arsenide, G...

Figure 11.42 The first three energy levels for an electron trapped in a one-...

Chapter 12

Figure 12.1 The electromagnetic spectrum; the visible portion extends from a...

Figure 12.2 The electric field of a light wave represented as a sinusoidal w...

Figure 12.3 The interference of light waves: (a, b) two waves in step add to...

Figure 12.4 Absorption and emission of radiation: (a) light absorption; (b) ...

Figure 12.5 Black body emission spectra.

Figure 12.6 Fluorescent lamps: electrons from the cathode collide with mercu...

Figure 12.7 The principle of LED operation. Electrons and holes combine in t...

Figure 12.8 Photon absorption and emission in direct (a, b) and indirect (c,...

Figure 12.9 Principle of an organic electroluminescent device (OLED); (a) th...

Figure 12.10 An MEH-PPV OLED: (a) energy level diagram for the anode, emitti...

Figure 12.11 The energy levels involved in the ruby laser: (a) spin-allowed ...

Figure 12.12 The main transitions in four-level laser operation.

Figure 12.13 The energy levels of most importance in the neodymium laser.

Figure 12.14 Semiconductor laser, schematic.

Figure 12.15 Refraction of a light beam on entering a transparent solid.

Figure 12.16 The effect of refractive index on the wavelength of light.

Figure 12.17 Total internal reflection of a beam of light: (a), (b) θ l...

Figure 12.19 Phase change on reflection at a surface of higher refractive in...

Figure 12.18 The reflection of light from a surface: (a) angles of incidence...

Figure 12.20 Multiple reflections at the upper and lower surfaces of a thin ...

Figure 12.21 The sinusoidal variation of the reflectivity of a thin film.

Figure 12.22 Reflection at a thin film in air: (a) perpendicular incidence (...

Figure 12.23 The approximate colour of a thin film when viewed from directly...

Figure 12.24 A quarter-wave stack of alternating layers of high and low refr...

Figure 12.25 Interference filters: (a) shortpass; (b) longpass; (c) bandpass...

Figure 12.26 Scattering: (a) Rayleigh scattering; (b) Mie scattering from re...

Figure 12.27 The diffraction pattern from a thin slit.

Figure 12.28 Diffraction gratings: (a) transmission; (b) reflection.

Figure 12.29 (a) Diffraction of X-rays by a crystal; (b) diffraction of ligh...

Figure 12.30 A one-dimensional photonic crystal.

Figure 12.31 Rainbow hologram: (a–c) a security hologram viewed at slightly ...

Figure 12.32 The structure of a silica glass optical fibre.

Figure 12.33 Alternative paths that light beams can take in the core of an o...

Figure 12.34 Types of optical fibre: (a) stepped index; (b) graded index; (c...

Figure 12.35 Signal amplification: an incoming weak signal is amplified on p...

Figure 12.36 Solar cells: (a) a photon incident upon a p–n junction generate...

Figure 12.37 Dye-sensitised solar cell, schematic: (a) reactions taking plac...

Figure 12.38 The structure of $(CH_3NH_3)PbCl_3$ ($MAPbCl_3$). In the layers precedi...

Figure 12.39 Perovskite solar cells, schematic: (a) dye-sensitised solar cel...

Figure 12.40 Interband transitions (schematic) between electron (upper) and ...

Figure 12.41 Green-emitting single quantum well (SQW) active layer LED.

Figure 12.42 Photoluminescent colours emitted by CdS quantum dots.

Figure 12.43 Quantum dot colours: (a) the change in energy level structure o...

Figure 12.44 A QLED screen (EL QD LED screen) schematic.

Figure 12.45 A ‘moth-eye’ surface structure.

Chapter 13

Figure 13.1 The variation of heat capacity of a solid with temperature: (a) ...

Figure 13.2 Variation of the heat capacity of a solid at a phase transition:...

Figure 13.3 Thermal conductivity: (a) the temperature gradient across a homo...

Figure 13.4 The variation of thermal conductivity with temperature of some c...

Figure 13.5 Thermal expansion: the mean coefficient of thermal expansion, [α

Figure 13.6 The variation of expansivity with temperature for tungsten, magn...

Figure 13.7 Interatomic potentials, schematic: (a) variation of potential wi...

Figure 13.8 The thermal expansion/contraction of water close to 0 °C.

Figure 13.9 The thermal expansion of the **c**-axis and contraction of the **a**- an...

Figure 13.10 Thermal contraction due to the removal of octahedral distortion...

Figure 13.11 Thermal contraction due to 'rocking' (rotation) of octahedra.

Figure 13.12 Thermal contraction brought about by silicon-oxygen 'hinges' co...

Figure 13.13 Thermoelectric effects: (a) the Seebeck effect; (b) the Peltier...

Figure 13.14 The Seebeck effect in a single phase.

Figure 13.15 Thermocouple temperature measurement.

Figure 13.16 A thermopile: thermocouples connected in series.

Figure 13.17 Use of the Peltier effect for (a) thermoelectric heating; (b) t...

Figure 13.18 Adiabatic cooling: the initial state is regenerated by the cool...

Figure 13.19 The giant magnetocaloric effect: the initial paramagnetic multi...

Figure 13.20 Liquid crystal thermometer display.

Figure 13.21 The cholesteric chiral nematic liquid crystal structure, schema...

Figure 13.22 The rutile structure of metallic VO_2 above 68 °C.

Figure 13.23 VO_2 thermochromic coating, schematic: (a) below the transition ...

Chapter 14

Figure 14.1 The uranium-238 decay series.

Figure 14.2 The thorium-232 decay series.

Figure 14.3 The actinium-227 (uranium-235) decay series.

Figure 14.4 The plutonium-241 decay series.

Figure 14.5 The band of stability of the elements.

Figure 14.6 The exponential rate of radioactive decay.

Figure 14.7 Plot of isotope ratios $^{87}Sr/^{86}Sr$ vs. $^{87}Rb/^{86}Sr$, schematic.

Figure 14.8 The isotopic binding energy per nucleon.

Figure 14.9 A fission chain reaction, schematic. In the chain illustrated, e...

Figure 14.10 A pressurised water nuclear reactor for power generation, schem...

Understanding Solids

The Science of Materials

Richard J. D. Tilley
University of Cardiff
United Kingdom

Third Edition

WILEY

This edition first published 2021

© 2021 John Wiley & Sons Ltd

Edition History

Richard J. D. Tilley, Understanding Solids, 1st edition (978-0-470-85275-0 Cloth, 978-0-470-85276-7 Paper)

Richard J. D. Tilley, Understanding Solids, 2nd edition (978-1-118-42328-8 Cloth, 978-1-118-42346-2 Paper)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at http://www.wiley.com/go/permissions.

The right of Richard J. D. Tilley to be identified as the author of this work has been asserted in accordance with law.

Registered Office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional

statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Name: Tilley, R. J. D., author.

Title: Understanding solids : the science of materials / Richard J. D. Tilley, Professor Emeritus, University of Cardiff, United Kingdom.

Description: Third edition. | Chichester, West Sussex, United Kingdom : John Wiley & Sons Ltd., 2021. | Includes bibliographical references and index.

Identifiers: LCCN 2021005722 (print) | LCCN 2021005723 (ebook) | ISBN 9781119716501 (hardback) | ISBN 9781119716549 (adobe pdf) | ISBN 9781119716556 (epub)

Subjects: LCSH: Materials science. | Solids. | BISAC: TECHNOLOGY & ENGINEERING / Material Science.

Classification: LCC TA403 .T63 2021 (print) | LCC TA403 (ebook) | DDC 620.1/1–dc23

LC record available at https://lccn.loc.gov/2021005722

LC ebook record available at https://lccn.loc.gov/2021005723

Cover Design: Wiley

Cover Image: © phochi/iStock/Getty Images

Preface

As with the previous editions, this new edition is aimed at introducing undergraduate students in sciences, engineering, and technology to the properties of solids and how these arise. The Chapters 1–4 cover basics such as atomic structure, chemical bonding, crystallography, and phase relationships, so as to provide a foundation for Chapters 5–14, which are concerned with describing the chemical and physical properties of solids in more detail.

All of the chapters in the second edition have been completely revised for this edition. The first five chapters in the earlier version have been condensed into three so as to present the material in a more coherent fashion, better suited to current requirements, without losing the overall aims of the previous edition: that of summarising basic properties essential to the understanding of the chemical and physical behaviours described in later chapters.

Since the previous edition, the scientific literature has been dominated by research on batteries and fuel cells, perovskite solar cells, lighting and displays, and solids with small dimensions: nanoparticles, whiskers, and sheets. All of these areas are introduced within the overall scheme of this book. The sections concerned with batteries and fuel cells have been updated, particularly with respect to lithium-ion and lithium-air cells. Sections concerned with organic framework structures, superionic conductors, mechanochemistry, bi-layer graphene, hologram formation and recording, perovskite solar cells, the optics of nanoparticle arrays, and thermochromic materials have been added.

New problems have been added throughout where helpful to the assimilation of newer material. In addition, to assist in making progress in these areas, the Further Reading sections at the end of each chapter include reference to a reasonable number of recent review articles on these topics. These, together with the other references, are designed to expand the material presented and provide a starting point for further exploration of the topics covered.

It is a pleasure to record help from the staff at John Wiley, including Emma Strickland, Jenny Cossham, and Dr. M. A. Preuss, and the

staff of the Trevithick library in Cardiff University for the continuing help with current literature. Finally, I must acknowledge my wife Anne, who has continually supported my efforts, and without whose help this revision would not have been possible.

R.J.D. Tilley
October 2020

About the Companion Website

This book is accompanied by a companion website for professors and instructors

www.wiley.com/go/tilley/understandingsolids3e



The website includes:

- answers to problems
- exercises



Part I
States of Aggregation

1
Atoms and Bonding

- What is an atomic term symbol?
- How are ionic radii determined?
- What are energy bands?

1.1 The Electron Structure of Atoms

An atom is made up of a small massive *nucleus*, in which almost all of the mass resides, surrounded by an *electron cloud*. Each element is differentiated from all others by the amount of positive charge on the nucleus, called the *proton number* or *atomic number*, *Z*. In a neutral atom, the nuclear charge is exactly balanced by *Z* electrons in the outer electron cloud, each of which carries one unit of negative charge. Variants of atoms that have slightly more or less electrons than are required for charge neutrality are called *ions*; those that have lost electrons have an overall positive charge and are called *cations*, while those that have gained electrons have an overall negative charge and are called *anions*. Many of the physical and chemical properties of solids described in later chapters are controlled by the outer electron structure of the component atoms.

1.1.1 Hydrogen

Hydrogen is the simplest atom and consists of a nucleus consisting of a single proton carrying one unit of positive charge together with a single bound electron carrying one unit of negative charge. The quantum mechanics of the hydrogen atom allows the probability of the location of the electron and its energy, defined in terms of *wavefunctions*, to be completely specified. The region of space defining the most probable location of the electron is termed an *orbital*. The wavefunctions are described by three *quantum numbers*: *n*, the *principal quantum number*; *l*, the *orbital angular momentum quantum number*; and m_l, the *magnetic quantum number*. The principal quantum number, *n*, defines the energy of the electron. It can take integral values 1, 2, 3, ... to infinity. The energy of the electron is lowest for $n = 1$ and this represents the most stable or *ground state* of the hydrogen atom. The next lowest energy is given by $n = 2$, then by $n = 3$, and so on.

The energy of each state is given by the formula:

$$E = -A/n^2$$

where *A* is a constant equal to 2.179×10^{-18} J (13.6 eV)[1] and *E* is the energy of the level with principal quantum number *n*. The negative

sign in the equation indicates that the energy of the electron is chosen as zero when n is infinite, that is to say, when the electron is no longer bound to the nucleus.

There is only one wavefunction and orbital associated with the lowest energy, $n = 1$, state. The states of higher energy each have n^2 different wavefunctions, corresponding to n^2 different orbitals, all of which have the same energy. There are four different wavefunctions and orbitals corresponding to $n = 2$, nine different wavefunctions and orbitals for $n = 3$, and so on. These are differentiated from each other by different values of the quantum numbers l and m_l, as explained in the following text. Wavefunctions with the same energy are said to be *degenerate*.

It is often convenient to represent the energy associated with each value of the principal quantum number, n, as a series of steps or *energy levels* (Figure 1.1). When an electron gains energy, it jumps from an energy level with a lower value of n to a level with a higher value of n. When an electron loses energy, it drops from an energy level with a higher value of n to an energy level with a lower value. The discrete packets of energy given out or taken up in this way are *photons* of electromagnetic radiation. The energy of a photon needed to excite an electron from energy E_1, corresponding to an energy level n_1 to energy E_2, corresponding to an energy level n_2 is given by:

$$\begin{aligned} E &= E_1 - E_2 = -2.179 \\ &\quad \times 10^{-18}[1/{n_1}^2 - 1/{n_2}^2]\,\mathrm{J} \\ &= -13.6\,[\,1/{n_1}^2 - 1/{n_2}^2]\,\mathrm{eV} \end{aligned}$$

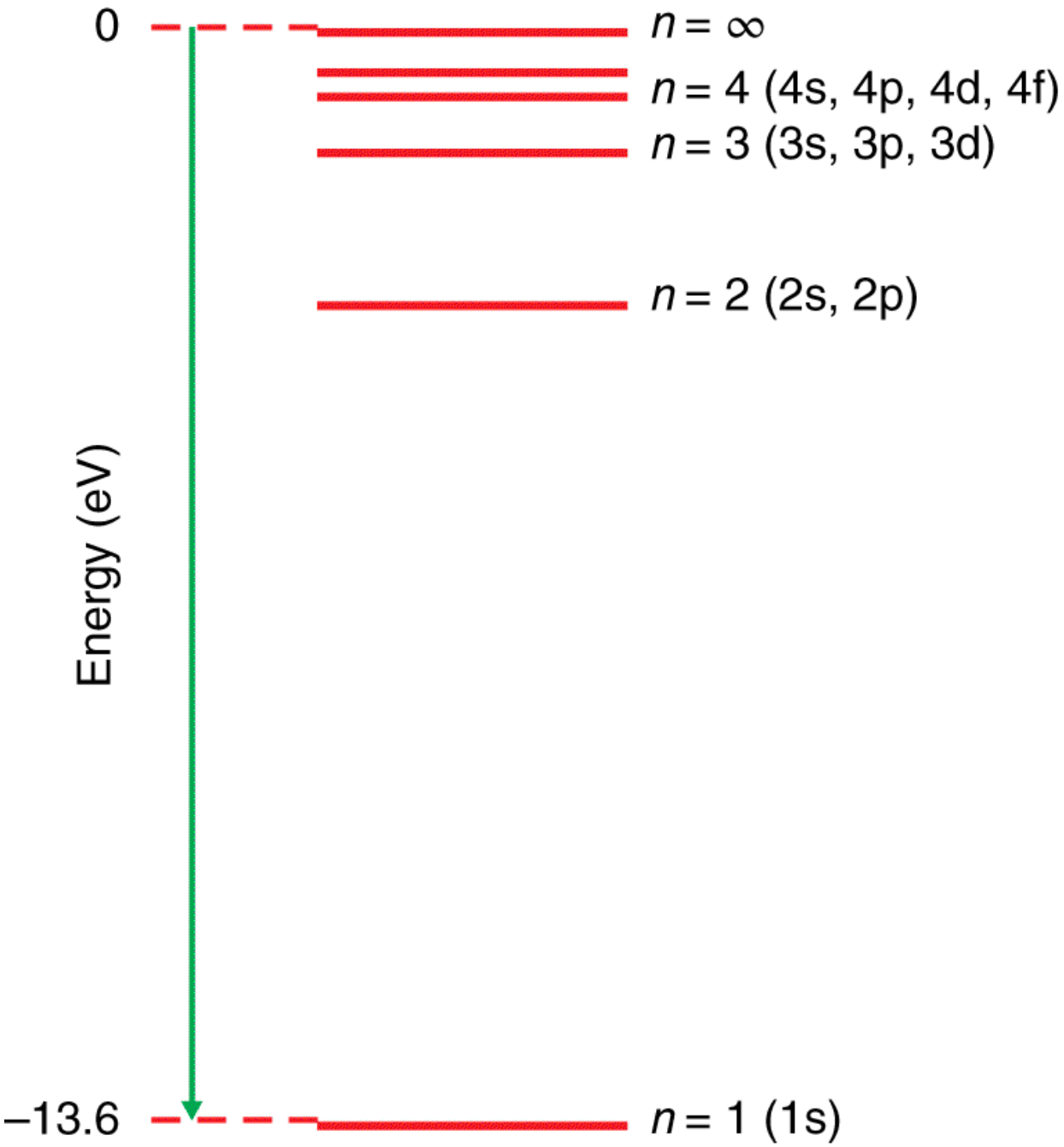


Figure 1.1 Electron energy levels of the H atom.

The energy of the photon emitted when the electron falls back from E_2 to E_1 is the same as that required for the excitation. The frequency ν (or the equivalent wavelength, λ) of the photons that are either emitted or absorbed during these energy changes is given by the equation:

$$E = E_1 - E_2 = h\nu = h\,c/\lambda$$

where h is the Planck constant. (Note that this equation applies to the transition between any two energy levels on any atom, not just between energy levels on hydrogen.) The energy needed to free the electron completely from the proton, which is called the *ionisation energy* of the hydrogen atom, is given by putting $n_1 = 1$ and $n_2 = \infty$ in the equation. The ionisation energy is 13.6 eV (2.179×10^{-18} J).

In the case of a single electron attracted to a nucleus of charge $+Ze$, the energy levels are given by:

$$E = -AZ^2/n^2$$

This shows that the energy levels are much lower in energy than in hydrogen, and that the ionisation energy of such atoms is considerably higher.

1.1.2 Many Electron Atoms

The simplest description of an atom with a nuclear charge of $+Z$ surrounded by Z electrons, called the *orbital approximation*, supposes that a single electron moves in a potential due to the nucleus and the average field of all the other electrons present in the atom, that is, it mirrors the hydrogen atom. The electron experiences an *effective nuclear charge*, Z_{eff}, which is considered to be located as a point charge at the nucleus of the atom. The energy levels of all of the orbitals drop sharply as Z_{eff} increases. When one reaches lithium, $Z = 3$, the 1s orbital energy has already decreased so much that it forms a chemically unreactive shell. This is translated into the concept of an atom as consisting of unreactive *core electrons*, surrounded by a small number of outermost *valence electrons*, which are of chemical and physical significance. Moreover, the change of energy as Z increases justifies the approximation that the valence electrons of all atoms are at similar energies.

The orbitals are specified by two interdependent quantum numbers, l and m_l, where:

l takes values of 0, 1, 2, ..., $(n - 1)$

m_l takes values of 0, ±1, ±2, ..., ±l

Each set of quantum numbers defines the *state* of the system. For a value of n = 1, there is only one state, corresponding to n = 1, l = 0, and m_l = 0. For n = 2, l can take values of 0 and 1, and m_l can then take values of 0, associated with l = 0, and −1, 0, and +1, associated with l = 1. For n = 3, l can take values of 0, 1, and 2, and m_l then can take values of 0, associated with l = 0, −1, 0, and +1, associated with l = 1 and −2, −1, 0, +1, +2, associated with l = 2. The same procedure applies to higher values of n. The various orbitals are given letter symbols. Orbitals with l = 0 are called *s orbitals*, those with l = 1 are called *p orbitals*, those with l = 2 are called *d orbitals*, those with l = 3 are called *f orbitals*, and those with l = 4 are called *g orbitals* (Table 1.1).

Table 1.1 Orbitals of hydrogen-like atoms.

n	l	m_l	Orbital	Number of electrons	Shell	Capacity
1	0	0	1s (1)	2	K	2
2	0	0	2s (1)	2	L	8
1	0, ±1	2p (3)	6			
3	0	0	3s (1)	2	M	18
1	0, ±1	3p (3)	6			
2	0, ±1, ±2	3d (5)	10			
4	0	0	4s (1)	2	N	32
1	0, ±1	4p (3)	6			
2	0, ±1, ±2	4d (5)	10			
3	0, ±1, ±2, ±3	4f (7)	14			
5	0	0	5s (1)	2	O	50
1	0, ±1	5p (3)	6			
2	0, ±1, ±2	5d (5)	10			
3	0, ±1, ±2, ±3	5f (7)	14			
4	0, ±1, ±2, ±3, ±4	5g (9)	18			
6	0	0	6s (1)	2	P	72
1	0, ±1	6p (3)	6			
2	0, ±1, ±2	6d (5)	10			

n	l	m_l	Orbital	Number of electrons	Shell	Capacity
3	0, ±1, ±2, ±3	6f (7)	14			
4	0, ±1, ±2, ±3, ±4	6g (9)	18			
5	0, ±1, ±2, ±3, ±4, ±5	6h (11)	22			

The set of orbitals derived from a single value of the principal quantum number form a *shell*. The lowest energy shell is called the K shell, and corresponds to $n = 1$. The other shells are labelled alphabetically (Table 1.1). There is only one s orbital in any shell, labelled 1s, 2s, and so on. There are three p orbitals in all shells from $n = 2$ upwards, collectively called 3p, 4p, and so on. There are five d orbitals in the shells from $n = 3$ upwards, collectively called 3d, 4d, 5d, and so on. There are seven f orbitals in the shells from $n = 4$ upwards, collectively called 4f, 5f, and so on. There are 9g orbitals added in the shells from $n = 5$ upwards and 22 h orbitals added in the shells from $n = 6$ upwards.

The extension of the various orbitals increases as the effective nuclear charge increases. This corresponds to the idea that heavy atoms are larger than light atoms. In addition, a different effective nuclear charge is experienced by electrons in differing orbitals. This has the effect of separating the energy of the *n*s, *n*p, *n*d, *n*f, and *n*g orbitals that are identical in hydrogen. It is found that for any value of n, the s orbitals have lowest energy, the three p orbitals have equal and slightly higher energy, the five d orbitals have equal and slightly higher energy again, and the seven f orbitals have equal and slightly higher energy again (Figure 1.2). However, the energy differences between the higher energy orbitals are very small and the energy of orbitals in different shells overlap so that the *n*d orbitals falls below that of the $(n + 1)$p orbitals (3d below 4p, 4d below 5p, and 5d below 6p) and the *n*f orbitals falls below that of the $(n + 2)$p orbitals (4f below 6p and 5f below 7p). The outer electron configuration of the elements is given in Appendix A, the Periodic Table. Note that in the case of the heavier elements the outer electron configurations are not as clearly defined as those of the lighter elements, and in these cases,

especially with some of the lanthanoids and Period 7 elements, the quoted electron configuration must be treated with caution.

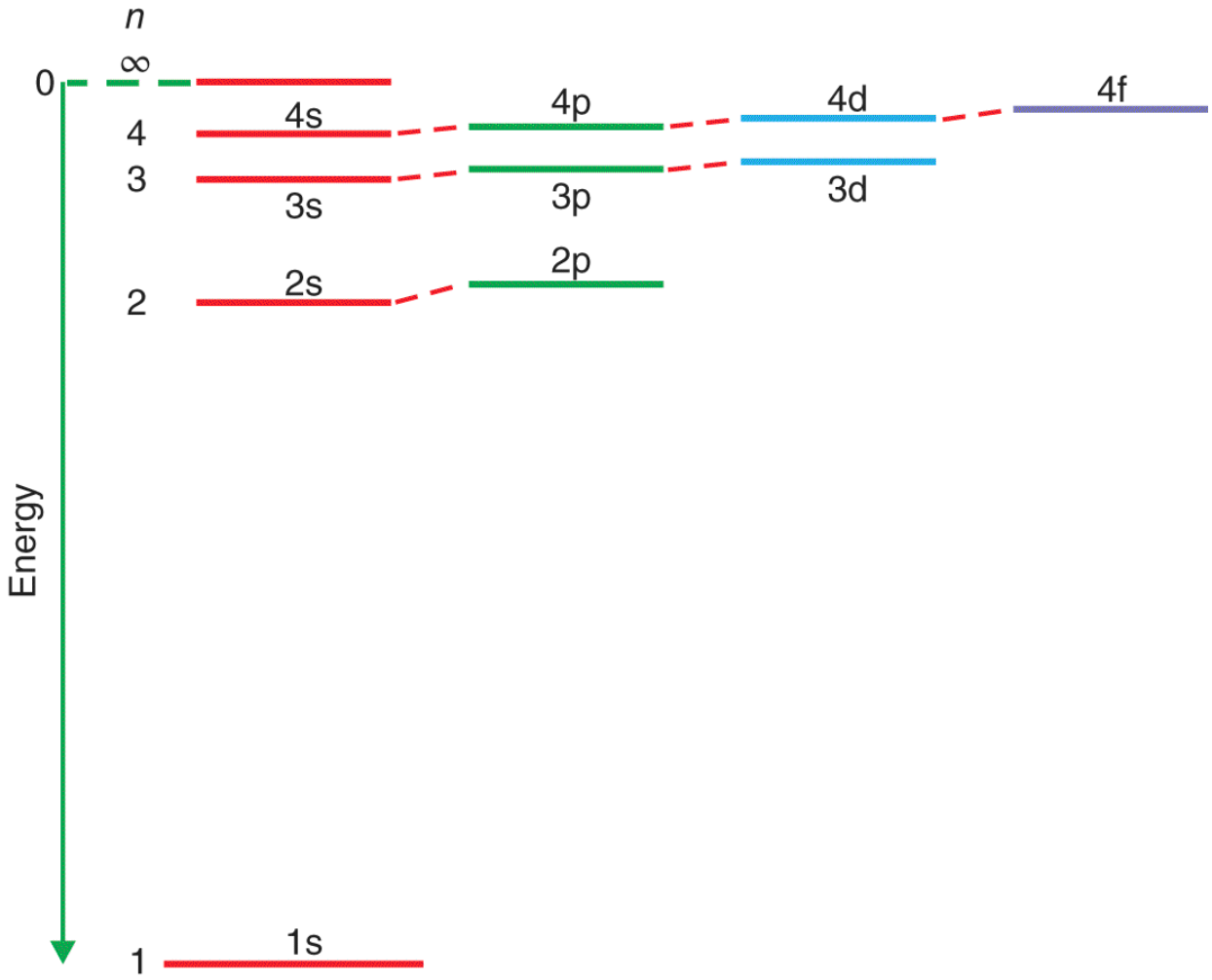


Figure 1.2 Schematic energy level diagram for a light, hydrogen-like atom.

1.1.3 Orbital Shapes

The shapes of the electron orbitals, or more properly, the probability of encountering an electron in a certain small volume of space surrounding the nucleus, is of importance as it controls many of the properties exhibited in solids, such as the colour and magnetic properties of the 3d transition metal compounds. The probability of encountering an electron in an s orbital does not depend upon direction but does vary with distance from the nucleus (Figure 1.3). For hydrogen, this probability peaks at a radial distance of 0.052

92 nm for a 1s orbital; equal to the distance calculated by Bohr as the minimum allowed radius of an orbiting 'planetary' electron around a proton, and is called the *Bohr radius*. As the electron is promoted to the 2s, 3s, and 4s orbitals, the maximum probability peaks further and further from the nucleus. Generally, s orbitals are drawn as spherical *boundary surfaces* that enclose an arbitrary volume in which there is a high probability, say, 95%, that the electron will be found (Figure 1.3a,b).

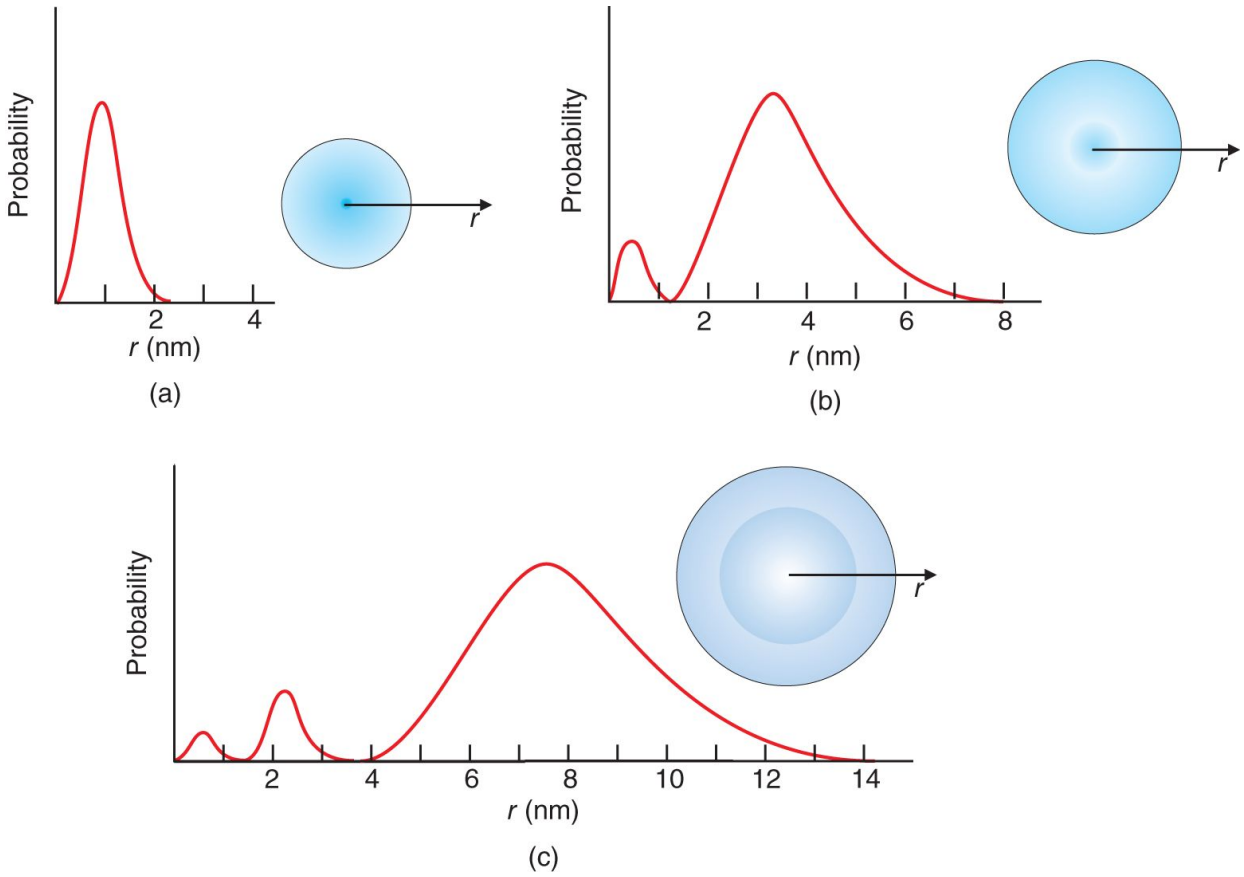


Figure 1.3 The probability of finding an s electron at a distance *r* from the nucleus: (a) 1s; (b) 2s; (c) 3s.

All other wavefunctions are specified by three quantum numbers and can be divided into two parts, a radial part, with similar probability shapes to those shown in Figure 1.3, multiplied by an angular part. The maximum probability of finding the electron depends on both the radial and angular parts of the wavefunction and the resulting boundary surfaces have complex shapes. For many purposes, however, it is sufficient to describe only the angular part of the wavefunction.

The boundary surfaces of the angular parts of the three p orbitals are approximately dumbbell shaped, each consisting of two lobes. These lie along three mutually perpendicular directions, which it is natural to equate to *x*-, *y*-, and *z*-axes (Figure 1.4). The corresponding orbitals are labelled np_x, np_y, and np_z, for example, $2p_x$, $2p_y$, and $2p_z$. Note that the electron occupies both lobes of the p orbital. The probability of encountering a p electron on the perpendicular plane that separates the two halves of the dumbbell is zero, and this plane is called a *nodal plane*. The sign of the wavefunction is of importance when orbitals overlap to form bonds. The two lobes of each p orbital are labelled as + and −, and the sign changes as a nodal plane is crossed. The radial probability of encountering an electron in a p orbital is zero at the nucleus, and increases with distance from the nucleus. The maximum probability is further from the nucleus for an electron in a 3p orbital than a 2p orbital, and so on, so that 3p orbitals have a greater extension in space than 2p orbitals.

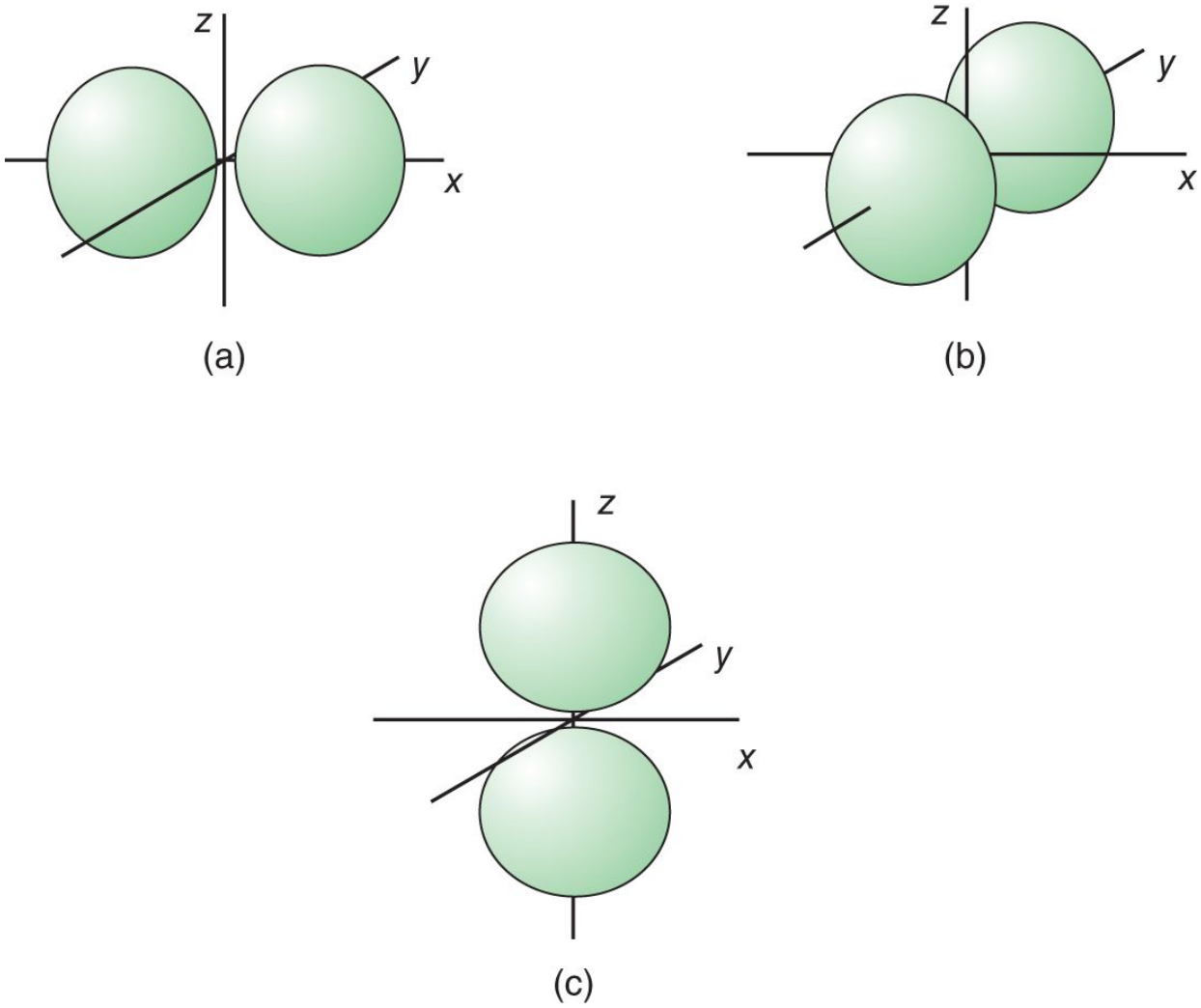
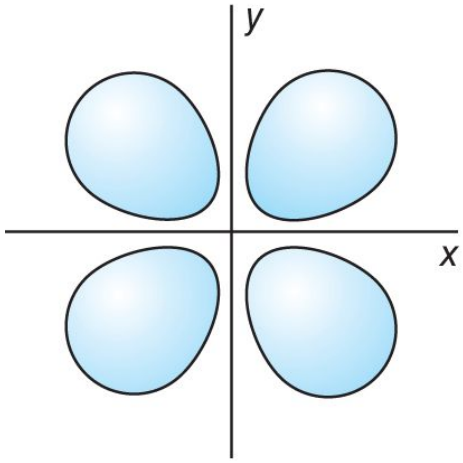
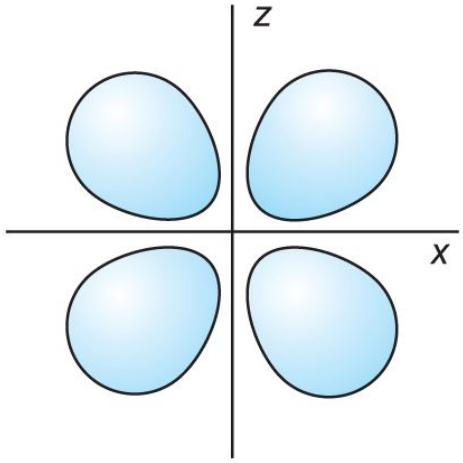


Figure 1.4 The boundary surfaces of the p orbitals: (a) p_x; (b) p_y; (c) p_z.

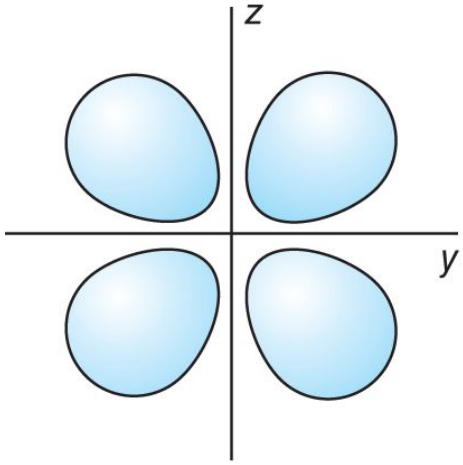
Of the five d orbitals, three of which have lobes lying between pairs of axes, d_{xy}, between the *x*- and *y*-axes, d_{xz}, between the *x*- and *z*-axes, and d_{yz}, between the *y*- and *z*-axes. The other two orbitals have lobes along the axes, $d_{x^2-y^2}$ pointing along *x*- and *y*-, and d_{z^2} pointing along the *z*-axis (Figure 1.5). Except for the d_{z^2} orbital, two perpendicular planar nodal planes separate the lobes and intersect at the nucleus. In the d_{z^2} orbital, the nodes are conical surfaces.



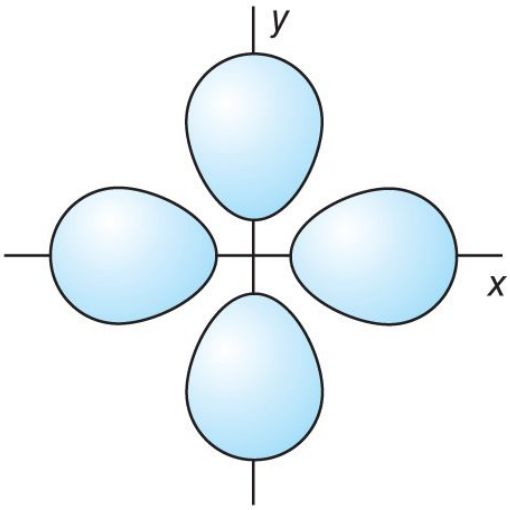
(a)



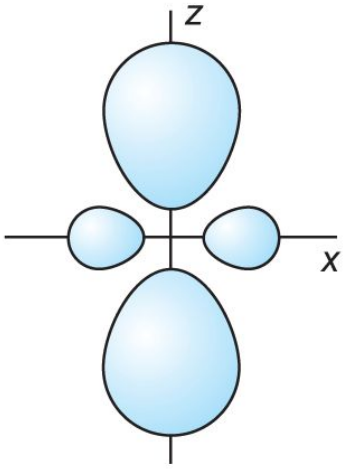
(b)



(c)



(d)



(e)

Figure 1.5 The boundary surfaces of the d orbitals: (a) d_{xy}; (b) d_{xz}; (c) d_{yz}; (d) $d_{x^2-y^2}$; (e) d_{z^2}.

1.1.4 Electron Spin and Electron Configuration

In order to account for other atomic properties, especially optical and magnetic properties, the electrons are allocated a fourth quantum number called the *spin quantum number*, *s*. The spin quantum number has a value of ½. The spin of an electron on an atom can adopt one of two different directions, represented by a quantum number, m_s, which takes values of + ½ or −½. These two spin directions are frequently represented by ↑, *spin up*, or *α*, and ↓, *spin down*, or *β*.

The *electron configuration* of an atom is the description of the number of electrons in each orbital, based upon the orbital approximation. This is usually given for the lowest energy possible, called the *ground state*. To obtain the electron configuration of an atom, the electrons are fed into the orbitals, starting with the lowest energy orbital, 1s, and then continuing to the higher energy orbitals so as fill them up systematically from the bottom up. This is called the *aufbau* (or *building up*) *principle*. Each orbital can hold a maximum of two electrons, which must have opposite values of m_s, +½ and −½. Two electrons in a single orbital are said to be *spin paired*. Thus, the s orbitals can hold 2 electrons, the three p orbitals can hold a total of 6 electrons, the five d orbitals can hold a total of 10 electrons, and the seven f orbitals a total of 14 electrons. When electrons are allocated to the p, d, and f orbitals, the lowest energy situation is that in which the electrons go into an unoccupied orbital if possible. This situation is expressed in *Hund's First Rule*: when electrons have a choice of several orbitals of equal energy, the lowest energy, or ground state, configuration corresponds to the occupation of separate orbitals with parallel spins rather than fewer orbitals with paired spins. For the heavier atoms the electron configuration does not always follow a strictly repetitive sequence, as the difference in energy between the various configurations is often miniscule. In particular, half-filled d and f orbitals are favoured when possible.

This simple approach is able to account for the overall variation in properties of the elements across the Periodic Table (Appendix A). The alkali metals, typified by Na, all have an outer electron configuration of ns^1, the alkaline earth elements, typified by Mg, a configuration of ns^2, the transition metals, typically Fe, have partly filled *n*d orbitals, the lanthanoids, partly filled 4f orbitals, and the inert gases such as Ne have a completely filled shell configurations.

1.1.5 Atomic Energy Levels

The spectra of atoms are a record of transitions that the outer electrons make between the available electron energy levels. The frequency ν (or the equivalent wavelength, λ) of the spectral line is related to the energy separation of the two energy levels, ΔE, by:

$$\Delta E = h\nu = h\,c/\lambda$$

where h is the Planck constant and c is the velocity of a plane light wave in a vacuum. The electron energy level scheme described earlier (Figure 1.2) applies to a one electron 'hydrogen-like' atom that has a single electron moving under the combined electrostatic field of the nucleus and all of the other electrons. While this is useful chemically, it is unable to account for the spectrum of an atom or provide any information on the possible energy levels appropriate to the electron configuration. There are a number of ways of obtaining these *many-electron* energy levels of an atom, which are described by *many electron* quantum numbers. The approach most frequently encountered, called the *Russell–Saunders coupling*, makes the approximation that the electrostatic repulsion between electrons is the most important energy term. To obtain revised atomic configurations, all of the individual *s* values of the electrons are summed to yield a total spin angular momentum quantum number S. Similarly, all of the individual l values for the electrons present are summed to give a total orbital angular momentum quantum number L. (Note that one-electron quantum numbers are written in lower case while many electron quantum numbers are written in upper case.) The value of S is not used directly, but is replaced by the *spin multiplicity*, $2S + 1$. Similarly, the total angular momentum quantum

number L, is replaced by a letter symbol similar to that used for the single electron quantum number l. The correspondence is

L	0	1	2	3	4	5	6
Letter	S	P	D	F	G	H	I

After $L = 3$, F, the sequence of letters is alphabetic, omitting J. Be aware that the symbol S has two interpretations, as the value of L (S Roman) and as the value of total spin (S *italic*). The compatible combinations of S and L are written in the form ^{2S+1}L, called a *term symbol*, which represents the energy level, called a *term* in spectroscopic parlance. States with a multiplicity of 1 are called *singlet* states, states with a multiplicity of 2 are called *doublet* states, with multiplicity of three, *triplets*, with multiplicity 4, *quartets*, and so on. Hence, ^{1}S is called singlet S, and ^{3}P is called triplet P.

To give an example, for a $3d^2$ transition metal cation such as Ti^{2+} or V^{3+} (which has a single energy level in Figure 1.2), the allowed Russell–Saunders terms are ^{3}F, ^{3}P, ^{1}G, ^{1}D, and ^{1}S. Although the derivation of terms is straightforward (see Appendix B), it is difficult to calculate the energies of these states. The lowest energy, ground state term, is, however, readily found using *Hund's first and second rules*: the term with the lowest energy has the highest multiplicity and for terms with the same multiplicity the term with the highest value of L is lowest in energy. The lowest energy term for a $3d^2$ ion is thus ^{3}F (Figure 1.6).

$3d^2$
^1S
^1G
^3P
^1D
^3F
$^1S_0$
$^1G_4$
$^3P_2$
$^3P_1$
$^3P_0$
$^1D_2$
$^3F_4$
$^3F_3$
$^3F_2$
0
+4
–4
+2
–2
+1
–1
0
+2
–2
+4
–4
+3
–3
+2
–2
No interaction
Electron–electron repulsion
(Russell-saunders terms)
Spin–orbit interaction
(j–j levels)
Magnetic field
2J + 1 levels

Figure 1.6 The evolution of the electron energy levels of an atom with a $3d^2$ configuration.

The term symbol does not account for the total complexity of an atomic spectrum and additional factors must be taken into account. The most important of these is the interaction between the electron orbital angular momentum and the electron spin angular momentum (*spin–orbit coupling* or *j–j* coupling). To include this, a new quantum number, *J*, is needed, given by combining *L* and *S* values (Appendix B). The new quantum number is incorporated as a subscript to the term, now written ${}^{2S+1}L_J$ and this is no longer called a term, but a *level*. Each value of *J* represents a different energy level. It is found that a singlet term always gives one level, a doublet, two, a triplet three, and so on. Thus, ground state term ^3P is composed of three levels $^3P_0$, ^3P, and $^3P_2$ (Figure 1.6). The magnitude of the energy difference between these levels depends upon the strength of the interaction between *L* and *S*. In order to find which of these states is the lowest energy, use can be made of *Hund's third rule*: the level with the lowest energy is that with the lowest *J* value if the valence shell is up to half full and that with the highest *J* value if the valence shell is more than half full.

Additionally, the spectrum of an atom in a magnetic field has more lines present than the same atom in the absence of the magnetic field, a feature called the *Zeeman effect*. In a magnetic field, each of the ${}^{2S+1}L_J$ levels splits into $(2J + 1)$ separated energy levels. The spacing between the levels is given by $g_J\mu_B\mathbf{B}$, where g_J is the Landé *g*-value,

$$g_J = 1 + \{[J(J+1) - L(L+1) + S(S+1)]\}/[2J(J+1)]$$

μ_B is a fundamental physical constant, the Bohr magneton, and **B** is the magnetic induction (Figure 1.6). A similar change, called the *Stark effect*, arises in the presence of a strong electric field.

An alternative approach to the Russell–Saunders coupling is to assume that the interaction between the orbital angular momentum and the spin angular momentum, i.e. the spin–orbit coupling, is the

most important. In this case, the *s* and *l* quantum numbers for an individual electron are added to give a total angular momentum number *j* for a single electron. These values of *j* are then added to give the total angular momentum quantum number *J*, for the whole atom. The technique of adding *j* values to obtain energy levels is called the *j–j coupling*.

Note that in a heavy atom it might be preferable to go from the electron configuration to levels derived by *j–j* coupling and then add on a smaller effect due to electron–electron repulsion (Russell–Saunders coupling) before finally including the magnetic field splitting. In real atoms, the energy levels determined experimentally are often best described by an intermediate model between the two extremes of Russell–Saunders and *j–j* coupling.

1.2 Ionic Bonding

1.2.1 Ionic Size and Bonding

Ions are charged species that form when the number of electrons surrounding a nucleus varies slightly from that required for electric neutrality. Metallic elements, which have few electrons outside a closed noble gas core, tend to lose electrons and form positively charged cations; the charge written as a superscript. Non-metals tend to gain electrons and form negatively charged anions; the charge written as a superscript. Groups of atoms can also form anions, for example, carbonate, CO_3^{2-}, and nitrate, NO_3^-. Ions are called *monovalent* if they carry a charge of ±1, *divalent* if they carry a charge of ±2, *trivalent* if they carry a charge of ±3, and so on. This does not depend upon the number of atoms in an ion. Thus, both Sr^{2+} and CO_3^{2-} are regarded as divalent ions.

Positive and negative ions attract each other electrostatically to form an *ionic bond*. Electrostatic interactions are *long range* and *non-directional* and these features typify ionic bonding. Ionic bonding has the advantage that it readily accounts for the formula of a solid. This follows directly from the idea that cations and anions have integer charges and ionic compounds are neutral overall. Thus, a cation with a charge $+a$ can combine with an anion of charge $-a$ to give neutral compounds MX. When the cations and anions have different charges, they combine in numbers to render the compound electrically neutral; for example, Al^{3+} and O^{2-} combine to give neutral Al_2O_3, while Sr^{2+} and Ti^{4+} combine with O^{2-} to give neutral perovskite, $SrTiO_3$.

The concept of allocating a fixed size to each ion is an attractive one and has been extensively utilised. Ionic radii are generally derived from X-ray crystallographic structure determinations, which give a precise knowledge of the distances between the ions. To derive ionic radii, it is assumed that the individual ions are spherical and in contact. The radius of one commonly occurring ion, such as O^{2-}, is taken as a standard. Other consistent radii can then be derived by

subtracting the standard radius from measured inter-ionic distances (Figure 1.7).

+1	+2	+3	+4 (+3)	+5 (+4) [3+]	+6 (+4) [3+]	+6 (+4) [3+] {2+}	+4 (+3) [2+]	+4 (+3) [2+]	+4 (+2)	+2 (+1)	+2	+3	+4 (2+)	+5 (3+)	−2 (6+)	−1
Li 0.088	Be 0.041(t)											B 0.026 (t)	C	N	O 0.126	F 0.119
Na 0.116	Mg 0.086											Al 0.067	Si 0.040 (t)	P	S 0.170	Cl 0.167
K 0.1521	Ca 0.114	Sc 0.0885	Ti 0.0745 (0.081)	V 0.068 (0.073) [0.078]	Cr [0.0755]	Mn (0.068) [0.079*] {0.097*}	Fe (0.079*) [0.092*]	Co (0.075*) [0.089*]	Ni (0.083)	Cu 0.087 (0.108)	Zn 0.089	Ga 0.076	Ge 0.068	As 0.064	Se (0.043 t)	Br 0.182
Rb 0.163	Sr 0.1217	Y 0.104	Zr 0.086	Nb 0.078	Mo 0.074	Tc (0.078)	Ru 0.076	Rh 0.0755	Pd (0.100)	Ag (0.129)	Cd 0.109	In 0.094	Sn 0.083 (0.105)	Sb 0.075	Te (0.068)	I 0.206
Cs 0.184	Ba 0.150	La 0.1185	Hf 0.085	Ta 0.078	W 0.074 (0.079)	Re 0.066	Os 0.077	Ir 0.077	Pt 0.077 (0.092)	Au	Hg 0.116	Tl 0.1025	Pb 0.0915 (0.132)	Bi 0.086 (0.116)	Po	At

* High spin configuration

Figure 1.7 Ionic radii, nm. Cation radii apply to ions octahedrally coordinated to O^{2-} unless marked (t), tetrahedral coordination. (For high spin configuration, see Section 10.2.2.)

The fact that the ionic radius depends upon the standard ion by which the radii were determined has led to a number of different tables of ionic radii. Although these are all internally self-consistent, radii from different sources should not be mixed. Additionally, cation radii are sensitive to the immediate surroundings. A cation surrounded by six oxygen ions in octahedral coordination has a different radius when surrounded by four oxygen ions in tetrahedral coordination or six sulfur ions in octahedral coordination. Ideally, tables of cationic radii should apply to a specific anion and coordination geometry.

The principal size characteristics of ions are the following:

(i) Cations are usually smaller than anions. The reason for this is that removal of electrons to form cations leads to a contraction of the electron orbitals due to the relative increase in nuclear charge while addition of electrons to form anions leads to an expansion of the charge clouds due to a relative decrease in the nuclear charge.

(ii) The radius of an ion increases with atomic number.

(iii) Cation radius decreases rapidly with increase of positive charge for a series of isoelectronic ions such as Na^+, Mg^{2+}, and Al^{3+}, all of which have the electronic configuration [Ne].

(iv) Anion radius is largely unaffected by an increase in negative charge. For example, F^- is similar in size to O^{2-} and Cl^- is similar in size to S^{2-}.

(v) Successive valence increases decrease cation radius. For example, Fe^{3+} is smaller than Fe^{2+}.

(vi) The radius of transition metal cations, and to a lesser extent lanthanoid and actinoid cations, depends upon whether the electron spins on the cation are paired, giving a low spin state, or largely unpaired, giving a high spin state.

While the majority of the ions of elements can be considered to be spherical, cations with a lone pair of electrons in the outer orbitals, In^+, Tl^+, Sn^{2+}, Pb^{2+}, Sb^{3+}, and Bi^{3+} are definitely not so. Similarly, ions, such as CO_3^{2-}, NO_3^-, are not spherical, although at high temperatures, rotation often makes them appear so.

1.2.2 Lattice Energies

An advantage of the ionic bonding model is that ionic interaction energies can be fairly easily assessed. The electrostatic potential energy (often called the *Coulomb interaction* or *Coulomb potential*) between a pair of oppositely charged monovalent ions E_e is given by:

$$E_e = (+e)(-e)/4\pi\varepsilon_0 r = -e^2/4\pi\varepsilon_0 r$$

where the point charges on the interacting species are $\pm e$, and the distance separating the charges is r. A negative overall energy implies stability. The electrostatic energy of an ionic crystal, called the *Madelung energy*, has a form identical to this multiplied by a constant that arises from the geometry of the structure, a term representing the charges on the ions and a term giving the number of ions present.

$$E_e = [-e^2/4\pi\varepsilon_0 r] \times [\text{geometry}] \times [\text{ionic charges}] \times [\text{amount}]$$

For a crystal composed of ions of charge $\pm Ze$:

$$E_e = N_A[-e^2/4\pi\varepsilon_0 r]\,[\alpha]\,[Z^2]$$

where N_A is Avogadro's constant and r is the equilibrium distance between neighbouring oppositely charged ions in the crystal. The term reflecting the geometry of the structure, α, is called the *Madelung constant*. The Madelung constant for each structure is different and is determined by summing all of the ionic interactions over the whole of the structure.

When the formula of the compound is M_mX_n, the charges on the cations and anions differ and as a result two alternative Madelung constants, usually designated A or M, may be quoted. They differ from each other in the molecular unit chosen for evaluation, M_mX_n or $MX_{n/m}$, and whether the ionic charges are included in the constant. In such cases it is simplest to separate out the charge contribution and stoichiometry to give a r*educed Madelung constant*, which is a purely geometric term. The reduced Madelung constant of a compound M_mX_n is defined by:

$$E_e = -N_A \frac{e^2}{4\pi\varepsilon_0 r} \alpha(Z_M Z_X) \frac{m+n}{2}$$

where, to maintain charge neutrality, $mZ_M = nZ_X$. Surprisingly, the reduced Madelung constant is very similar for a wide range of structures, equal to 1.68 ± 0.08 (Table 1.2). This means that the approximate electrostatic energy of any crystal structure can be estimated as long as the chemical formula is available. To convert between literature values of the Madelung constant A or M, and the reduced Madelung constant α, for a compound of formula M_mX_n, use:

Table 1.2 Reduced Madelung constants, α.

Structure	Formula	Example	α
Halite	M^+X^-	NaCl	1.748
Caesium chloride	M^+X^-	CsCl	1.763
Sphalerite	$M^{2+}X^{2-}$	ZnS	1.638
Wurtzite	$M^{2+}X^{2-}$	ZnO	1.641
Fluorite	$M^{2+}X^-_2$	CaF_2	1.680
Rutile	$M^{4+}X^{2-}_2$	TiO_2	1.605
β-Quartz	$M^{4+}X^{2-}_2$	SiO_2	1.480
Corundum	$M^{3+}_2X^{2-}_3$	Al_2O_3	1.669

$$\alpha Z_M Z_X (m+n)/2 = A$$

$$\alpha (m+n)/2 = M$$

In general, the electrostatic attraction between cations and anions will increase as they approach until, at some interionic distance, the electron clouds of the ions begin to overlap, leading to repulsion. Ultimately the two opposing energies will balance and the ions will adopt an equilibrium separation. The *repulsive potential energy*, E_r, can be formulated in a number of ways, typically:

$$E_r = B/r^n$$

where B and n are empirical constants.

The *potential energy*, U, per mole of an ionic crystal may be represented as the sum of the electrostatic and repulsive energy terms:

$$U = E_e + E_r = [-N_A \alpha Z^2 e^2/4\pi\varepsilon_0 r] + N_A B/r^n$$

per mole. The energy is a function of the distance between the ions, r, and at equilibrium this energy must pass through a minimum (Figure 1.8). Thus, we can write:

$$dU/dr = [N_A \alpha Z^2 e^2 / 4\pi\varepsilon_0 r^2] - nN_A B / r^{n+1} = 0$$

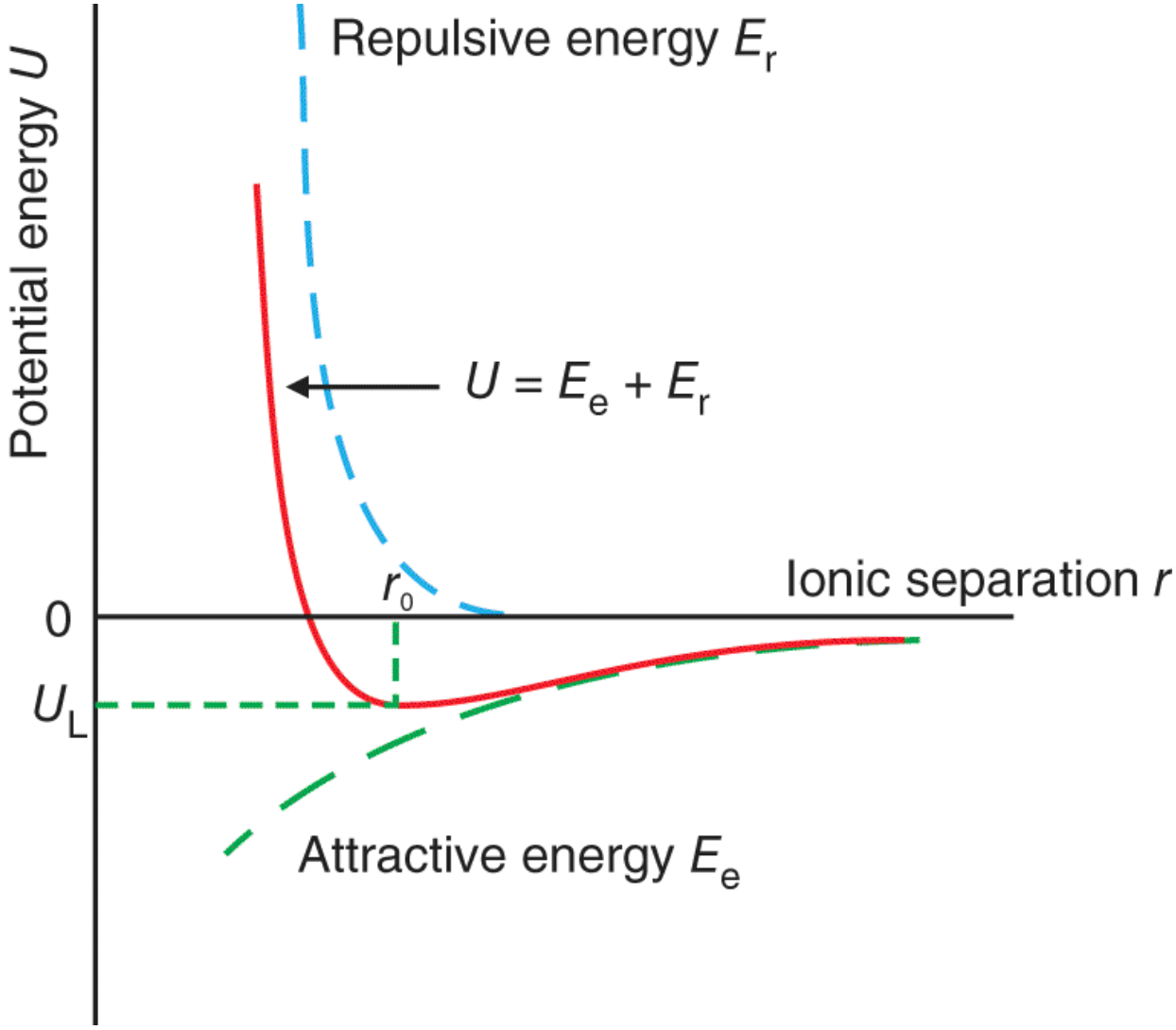


Figure 1.8 The total potential energy between two monovalent ions as a function of the ionic separation. The lattice energy U_L corresponds to the minimum, at an interatomic separation r_o.

This allows the constant B to be eliminated, to give:

$$U_L = [-N_A \alpha Z^2 e^2 / 4\pi\varepsilon_0 r_0][1 - 1/n]$$

where U_L is the *lattice energy* and r_0 is the equilibrium value of the interionic separation.

Using:

$$E_r = +N_A B \exp(-r/r^*)$$

where B and r^* are constants that are structure sensitive and eliminating the constant B, as earlier, gives the *Born–Mayer equation* for the lattice energy:

$$U_L = [N_A \alpha Z^2 e^2 / 4\pi\varepsilon_0 r_0][1 - (r^*/r_0)]$$

1.2.3 Atomistic Simulation

The analysis in Section 1.2.2 forms the basis for the computer evaluation of crystal structure information using the technique of *atomistic simulation*. In this technique interactions between pairs of atoms, written as *pair potentials*, are derived using experimental data or theoretical calculations. These can be listed as Coulombic (electrostatic) and non-Coulombic (mainly repulsive). Coulombic forces are of the form:

$$E_e = (q_i q_j)/4\pi\varepsilon_0 r_{ij}$$

where q_i and q_j are the charges on the ions and r_{ij} is the separation of the ions. The non-Coulombic potential, which is attractive at large interatomic spacing and repulsive at short interatomic spacing, is frequently given by one of three equations:

(a) *The Lennard-Jones potential*: $V_{LJ} = Ar^{-12} - Br^{-6}$

where A and B are empirical constants and r is the interatomic distance. This equation is found to fit the behaviour of rare gas and molecular interactions well.

(b) *The Morse potential*: $V_M = D\{1 - \exp[-\beta(r - r_e)]\}^2$

where D, β, and r_e are empirical constants. This form works well for covalently bonded solids (Section 1.3).

(c) *The Buckingham potential*: $V_B = B \exp(-r/\rho) - C/r^6$

which is of most value for ionic and semi-ionic solids. In this equation B, C, and ρ are empirical parameters, which vary from one ion pair to another.

Many ions are polarizable and this must be taken into account if reasonable answers are to be obtained. It is estimated by using the *shell model*, in which the polarizability is imagined to deform the electron cloud surrounding an atom as if the electron cloud were a shell attached to the nucleus by a spring. The simplest form for the polarizability α is:

$$\alpha = Y^2/k$$

where Y is the charge on the (massless) shell and k is the empirically determined *spring constant*. The potential energy stored in the spring is given by:

$$E_p = \frac{1}{2} k\, \delta r^2$$

or, if this approximation is not sufficiently accurate, by:

$$E_p = \frac{1}{2} k_2\, \delta r^2 + \left(\frac{1}{24}\right) k_4\, \delta r^4$$

where δr is the amount by which the spring between the shell and the core is stretched.

Computer routines sum all of these interactions over as many atoms as is reasonable and use techniques that effectively allow the finite answer to be extended to a crystal of any size. Pair potentials are somewhat limited and these are gradually being replaced, where the computational requirements allow it, with *many-body potentials*. These are more sophisticated than pair potentials and are able to more faithfully model processes at an atomic level.

There are two broad genres of computation. In *static lattice* simulations the atoms in the crystal, positioned at r_1, r_2, r_3, and so on. A first approximation energy is derived and further iterations are then made by changing the ionic positions until a minimum energy is reached. The final set of atomic coordinates give the crystal structure and unit cell dimensions of the atom array, which are then easily compared to those obtained via X-ray diffraction. By applying virtual forces or electric fields to the optimal structure other properties, such as crystal stability under high pressure, elastic constants and dielectric behaviour can be assessed. By removing or adding ions the energy of formation of point defects can be determined.

The other broad area of atomistic simulation concerns dynamic properties. Here the atoms are given a position and a velocity. Successive configurations of the structure are allowed to evolve with time, either the atomic trajectories following Newton's laws of motion (*molecular* or *lattice dynamics*) or else determined using probability theory (*Monte Carlo* methods). Using these approaches, defect formation and migration, diffusion, the annealing of materials, fracture, grain growth and sintering, nanoscale materials, and damage from radiation emitted by radioactive atoms can be simulated. Molecular dynamics simulations have been used, for example, to study the enormously complex problem of protein folding.

1.3 Covalent Bonding

1.3.1 Bond Geometry

Ionic bonding is unable to explain the properties of most molecules, particularly organic molecules, which have definite shapes. Methane, CH_4, has the central carbon atom surrounded by hydrogen atoms to form a tetrahedral molecule; benzene C_6H_6, consists of six carbon atoms linked in a hexagonal ring, with one H atom linked to each C atom. The links between the atoms in such molecules are described as *covalent bonds*. A covalent bond is formed by the overlap of partly-filled orbitals. The *direction* of the bond will be such as to make the overlap as great as possible and the *strongest* bonds are formed when the overlapping of the orbitals is at a maximum.

The simplest covalent bond is, perhaps, the single bond between two hydrogen atoms that make up the hydrogen (H_2) molecule. An isolated hydrogen atom has a single electron in a spherical 1s orbital. As distance between two H atoms is reduced, two different kinds of interaction are possible, depending on whether the spins of the electrons in the s orbitals of the two atoms are parallel or opposed. If the spins are opposed, covalent bonding occurs and the nuclei are pulled together. It is found that the electron density, which was originally spherically distributed around each atom is now concentrated between the nuclei (Figure 1.9a,b). If the spins of the two electrons are parallel, the electron clouds repel and bonding does not occur, even if the atoms approach closely (Figure 1.9c,d). The resulting orbitals that encompass both atoms are called *molecular orbitals*.

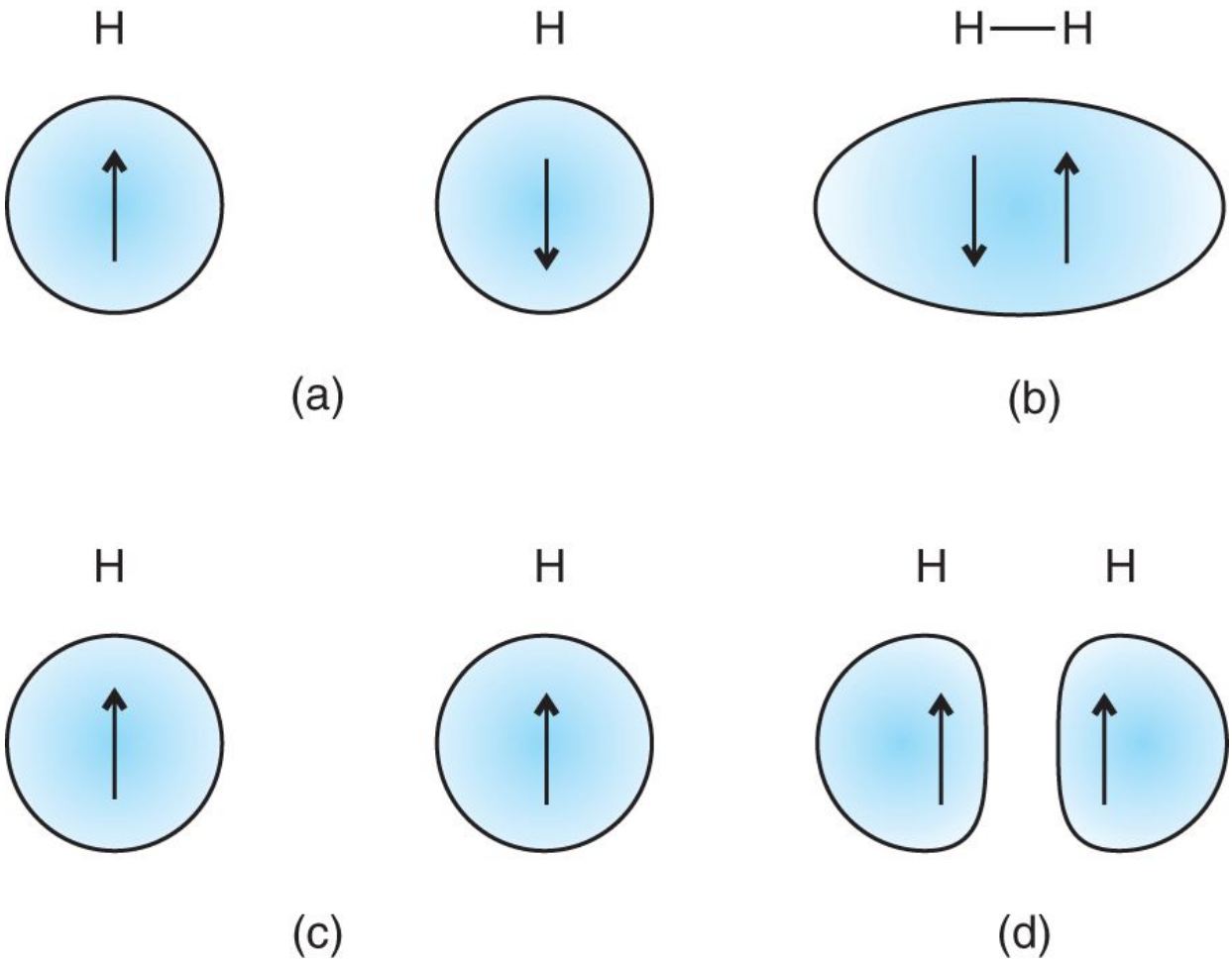


Figure 1.9 Covalent bond formation: (a) two H atoms with opposed electron spins form an H_2 molecule (b); (c) two H atoms with parallel electron spins do not bond (d).

Any partly filled orbitals can overlap to form bonds in this way. A bond formed by two s orbitals, two end-on p orbitals, or by a single s and a single p orbital has rotational symmetry about the bond axis (Figure 1.10a,b) and is termed σ *bond*. Two p orbitals, linked sideways, cause an increase of the electron density on either side of a nodal plane on which the two nuclei are situated. Such bonds are termed π *bonds* (Figure 1.10c,d). It is important to note that the designation of a bond as σ or π does not depend on the type of orbital forming the bond, only the geometry of overlap of the orbitals.

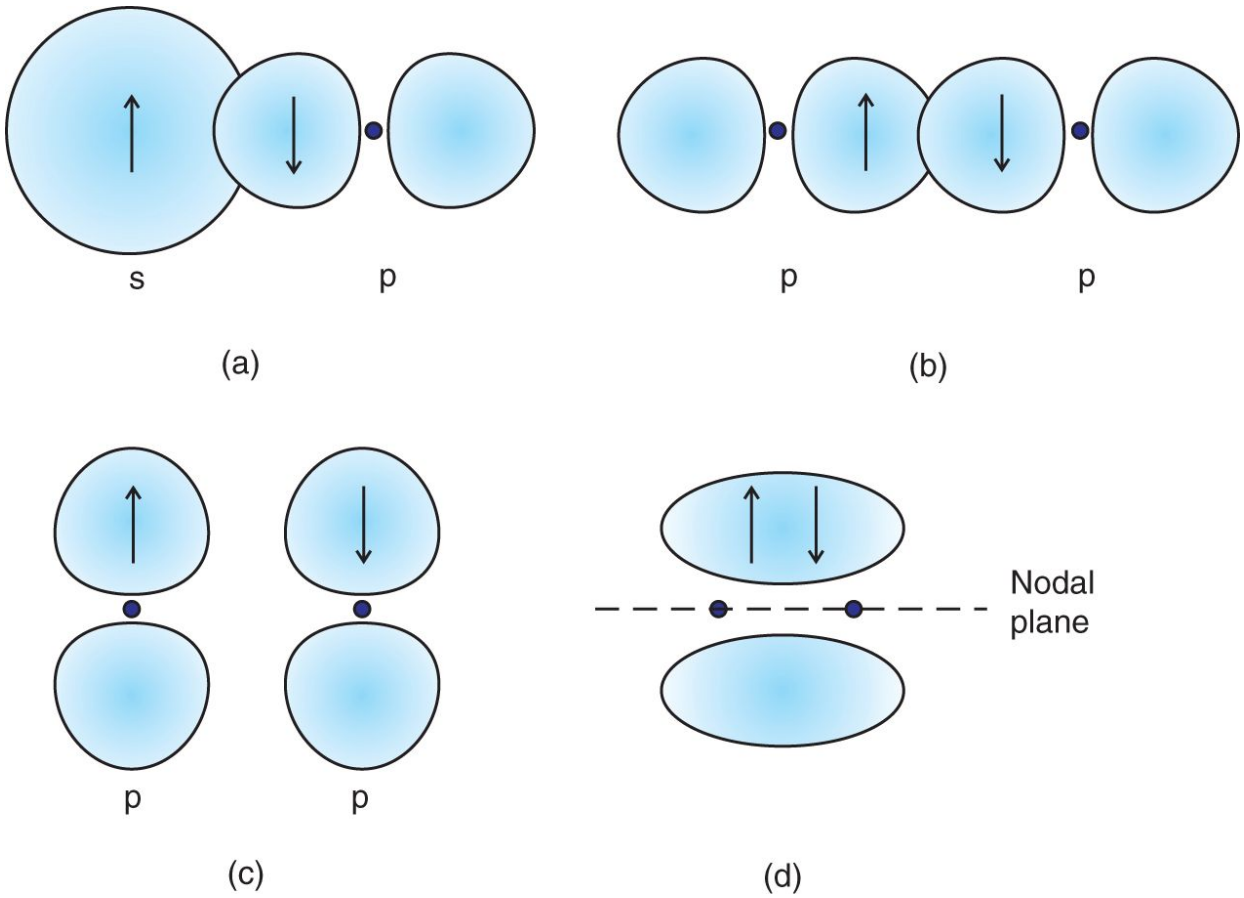


Figure 1.10 Covalent bonding: (a) σ bonding via overlap of s and p orbitals; (b) σ bonding via overlap of two p orbitals; (c) sideways on approach of two p orbitals leading to (d) π bond formation.

Although covalent bonds due to simple orbital overlap accounts for a large number of molecular structures, it does not explain the geometry of many other simple molecules such as methane, CH_4, ammonia, NH_3, and water, H_2O. In these cases, the simple atomic orbitals are rearranged in such a way that they can make stronger bonds (with greater overlap) than the unaltered orbitals. This can be illustrated by the tetrahedral geometry of many bonds in organic compounds. The outer electron configuration of carbon is $2s^2\ 2p^2$. If one electron is promoted from the filled $2s^2$ orbital into the empty p orbital, the four singly occupied orbitals rearrange into *sp^3-hybrid* orbitals. The resulting four orbitals point towards the corners of a tetrahedron, at angles of 109.5° to each other (Figure 1.11). These orbitals have an overlapping power of twice the overlapping power of

s orbitals so that the bonds are extremely strong. The C□C bond energy in diamond, the hardest of all solids, is 245 kJ mol^{-1}.

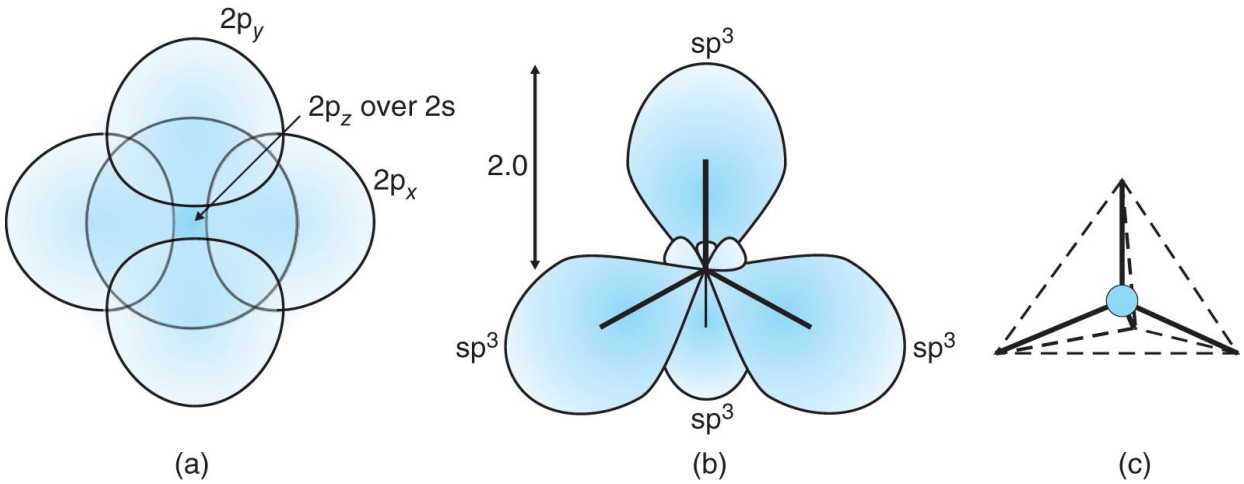


Figure 1.11 (a) The 2s and 2p orbitals on C; (b) four sp^3 hybrid orbitals; (c) tetrahedral geometry of the sp^3 orbitals.

It is a general rule of hybrid bond formation that the same number of hybrid orbitals form and can be used for bonding as the number of atomic orbitals used in the initial combination. Thus, one s and one p orbital yield two sp-hybrid orbitals resulting in linear molecules. One s orbital and two p orbitals yield three new sp^2-hybrid orbitals for bond formation. For maximum overlap these orbitals point as far away from each other as possible, forming bonds at angles of 120° yielding triangular molecules. This type of bonding is commonly encountered in borosilicate glasses, where the boron atoms are linked to three oxygen atoms at the corners of an equilateral triangle. The geometry of some hybrid orbitals is given in Table 1.3.

Table 1.3 The geometry of some hybrid orbitals.

Orbital configuration	Coordination number	Geometry	Example
sp	2	Linear	$HgCl_2$
sp^2	3	Trigonal	BCl_3
sp^3	4	Tetrahedral	CH_4
dsp^3	5	Trigonal bipyramidal	PCl_5
d^2sp^3	6	Octahedral	SF_6

One of the characteristic features of covalent compounds is the presence of multiple bonds between atoms. Single bonds, described earlier, are *always* σ bonds. Multiple bonds result when atoms link via σ and π bonds at the same time. A double bond between two atoms then consists of σ and one π bond, while a triple bond consists of one σ and two π bonds.

1.3.2 Bond Energies

As atoms come together to form a covalent molecule, the atomic orbitals interact and spread over all of the nuclei to give molecular orbitals that extend over the whole molecule. Each pair of atomic orbitals gives rise to two molecular orbitals, one of which has a higher energy than the original pair and one with a lower energy than the original pair. The molecular orbital of lower energy than the parent atomic orbitals has the greatest concentration of electron density between the nuclei, and is a *bonding orbital*. The molecular orbital of higher energy than the parent atomic orbitals, the *antibonding orbital*, has the electron density concentrated in the region outside of the line joining the nuclei. If they have cylindrical symmetry, they are called σ orbitals, the lower energy orbital is labelled σ and upper antibonding one labelled σ*. If the molecular orbitals are derived from sideways on overlap of p orbitals, they lack this symmetry and are called π orbitals, the lower bonding orbital labelled π and the upper antibonding orbital labelled π*. The available electrons are fed into the molecular orbitals, using the Pauli principle and Hund's rules

(the aufbau principle), to arrive at a molecular electron configuration. For a stable molecule to form, the total energy of the electrons in the molecular orbitals that are occupied must be less than the energy total when the electrons occupy separate atomic orbitals.

The principle can be illustrated by describing the ground state electron configuration and bond energy for homonuclear diatomic molecules. For H_2 the two 1s orbitals of the isolated atoms combine to give two molecular orbitals, σ1s (bonding) and, σ*1s (antibonding) (Figure 1.12). Both electrons will occupy the bonding, σ, orbital provided that they have opposed spins. This will be the lowest energy configuration, or *ground state*, of the pair, and a covalently bonded hydrogen molecule, H_2, will form. The bond energy will be $2E_{bond}$.

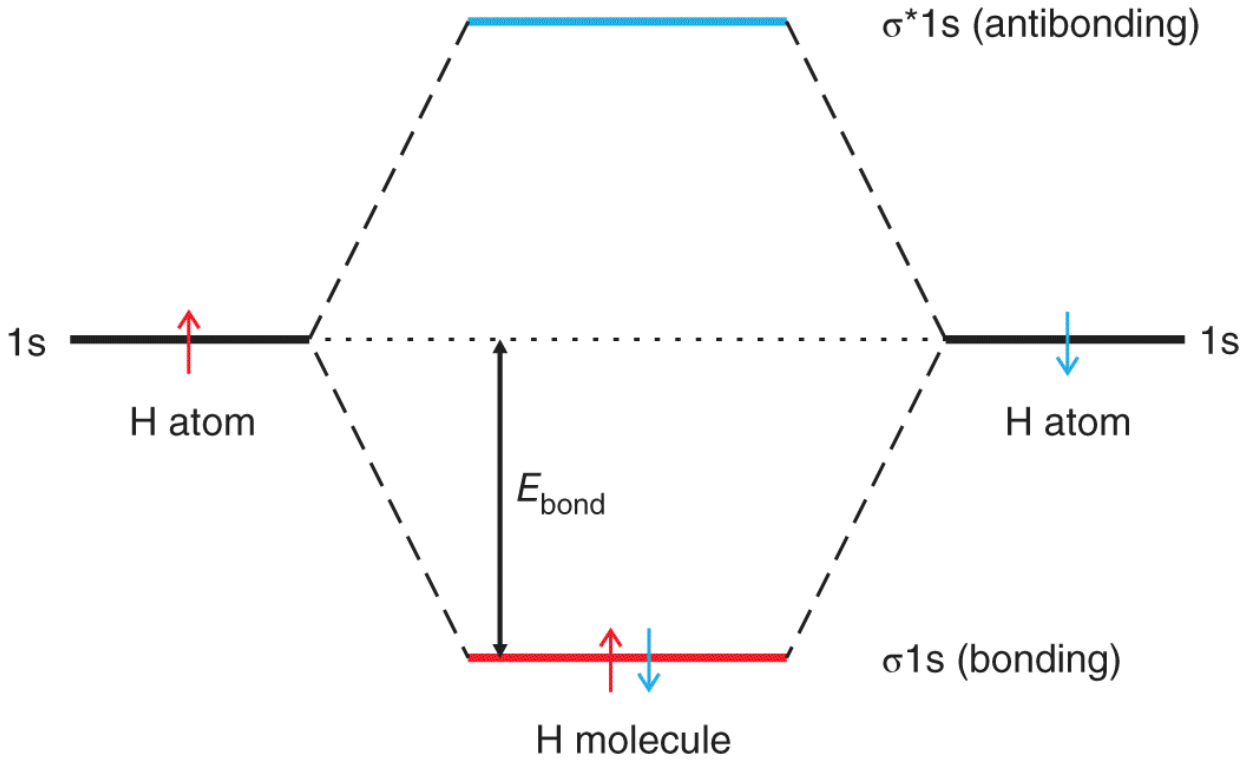


Figure 1.12 The formation of molecular orbitals in an H_2 molecule.

When two helium atoms interact, there are four electrons to be placed in the orbitals and so both the σ1s and σ*1s orbitals will be filled. The effect of the filled antibonding orbital completely negates the effect of the filled bonding orbital. No energy is gained by the system and so He_2 does not form.

To derive the electron configurations of the other homonuclear X_2 molecules, formed from the elements of the second period, Li_2 to Ne_2, exactly the same procedure is followed. The interaction of the 2s outer orbitals will form σ2s and σ*2s orbitals. End-on overlap of the p orbitals produces $\sigma 2p_x$ and $\sigma^* 2p_x$ orbitals. The sideways on overlap of a pair of p orbitals forms one $\pi 2p_y$ bonding orbital, one $\pi 2p_z$ bonding orbital, one $\pi^* 2p_y$ antibonding orbital, and one $\pi^* 2p_z$ antibonding orbital. The difference in energy between the $\sigma 2p_x$ and π2p orbitals is small and gradually changes along the series, so that the $\sigma 2p_x$ orbital drops below the π2p orbitals for the last three molecules, O_2, F_2, and (hypothetical) Ne_2 (Figure 1.13). The molecular configurations of the homonuclear diatomic molecules can now be obtained (Table 1.4).

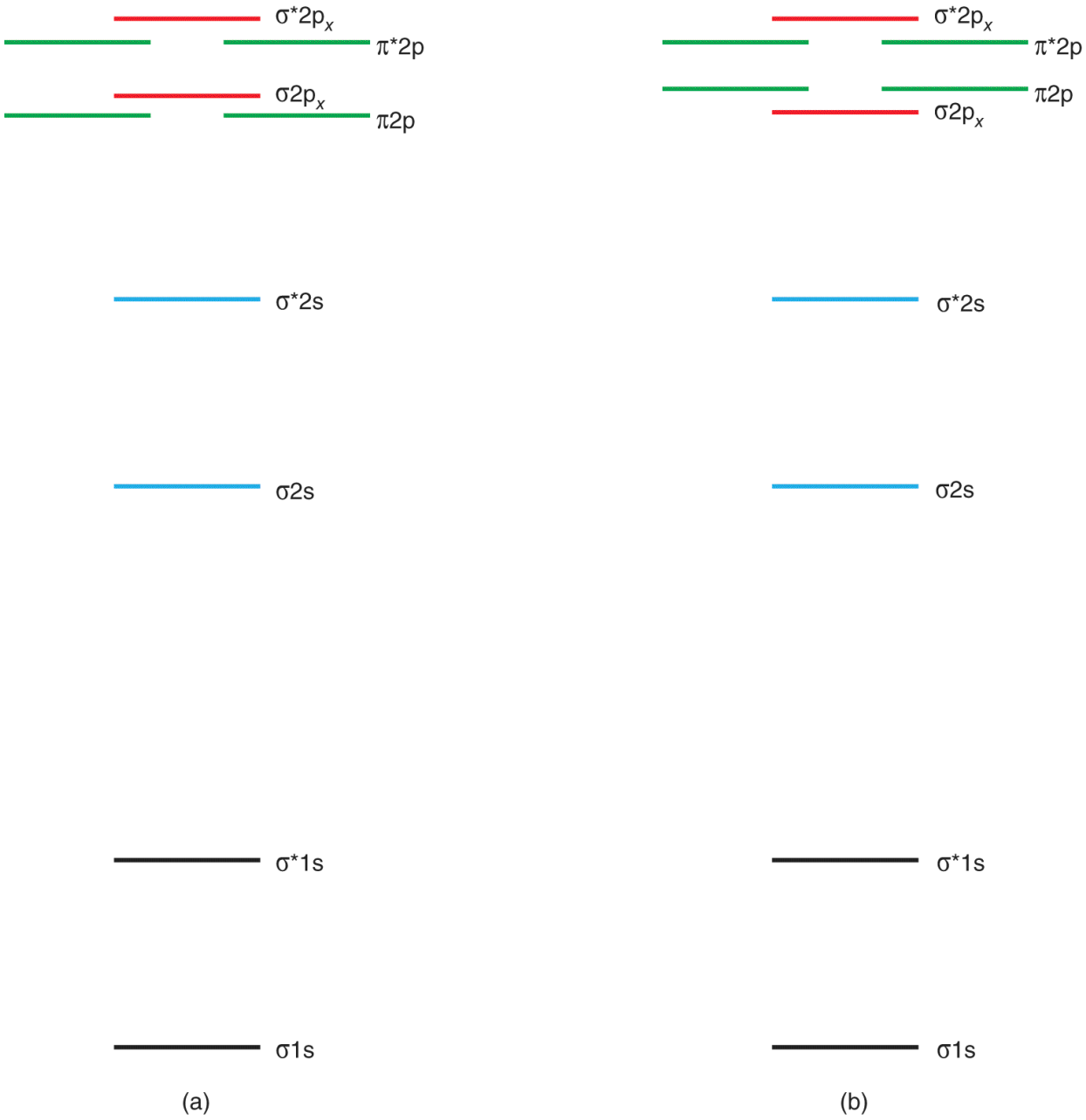


Figure 1.13 Schematic molecular orbital energy level diagram for homonuclear diatomic molecules; (a) H_2 to N_2; (b) O_2 to Ne_2.

Table 1.4 The electron configurations of some homonuclear diatomic molecules.

Molecule	Ground state configuration	Bond length (nm)	Bond energy (kJ mol^{-1})
Li_2	$[He_2]\ (\sigma 2s)^2$	0.267	101
Be_2	$[He_2](\sigma 2s)^2(\sigma^* 2s)^2$	—	—
B_2	$[Be_2](\pi 2p)^2$	0.159	289
C_2	$[Be_2](\pi 2p)^4$	0.124	599
N_2	$[Be_2](\pi 2p)^4(\sigma 2p_x)^2$	0.110	941
O_2	$[Be_2](\pi 2p)^4(\sigma 2p_x)^2\ (\pi^* 2p)^2$	0.121	494
F_2	$[Be_2](\pi 2p)^4(\sigma 2p_x)^2\ (\pi^* 2p)^4$	0.142	154
Ne_2	$[Be_2]$ $(\pi 2p)^4(\sigma 2p_x)^2(\pi^* 2p)^4(\sigma^* 2p_x)^2$	—	—

[He] is shorthand for $(\sigma 1s)^2\ (\sigma^* 1s)^2$.

$[Be_2]$ is shorthand for $(\sigma 1s)^2(\sigma^* 1s)^2\ (\sigma 2s)^2(\sigma 2s^*)^2$.

When a molecular orbital is formed between atoms of two different elements, A and X, then the energy levels of the initial atomic orbitals will differ, as will their extensions in space. The difference between the atoms can be thought of in terms of *electronegativity*, which is a measure of the power to attract electrons during chemical bonding. Atoms with a low electronegativity are called *electropositive* elements. Atoms with a high electronegativity are called *electronegative* elements. Despite differences in electronegativity, the molecular orbital diagram remains similar (Figure 1.14). The bonding energy E_b is now computed with respect to the average energy of the isolated A and X atoms, $\frac{1}{2}(E_A + E_X)$. In the case where element A is more metallic (electropositive) in character than element X, it is found that the X atom contributes most to the bonding molecular orbital and the atom A more to the antibonding molecular orbital.

The bonding molecular orbitals are then often said to be X-like in character and the antibonding orbitals A-like in character.

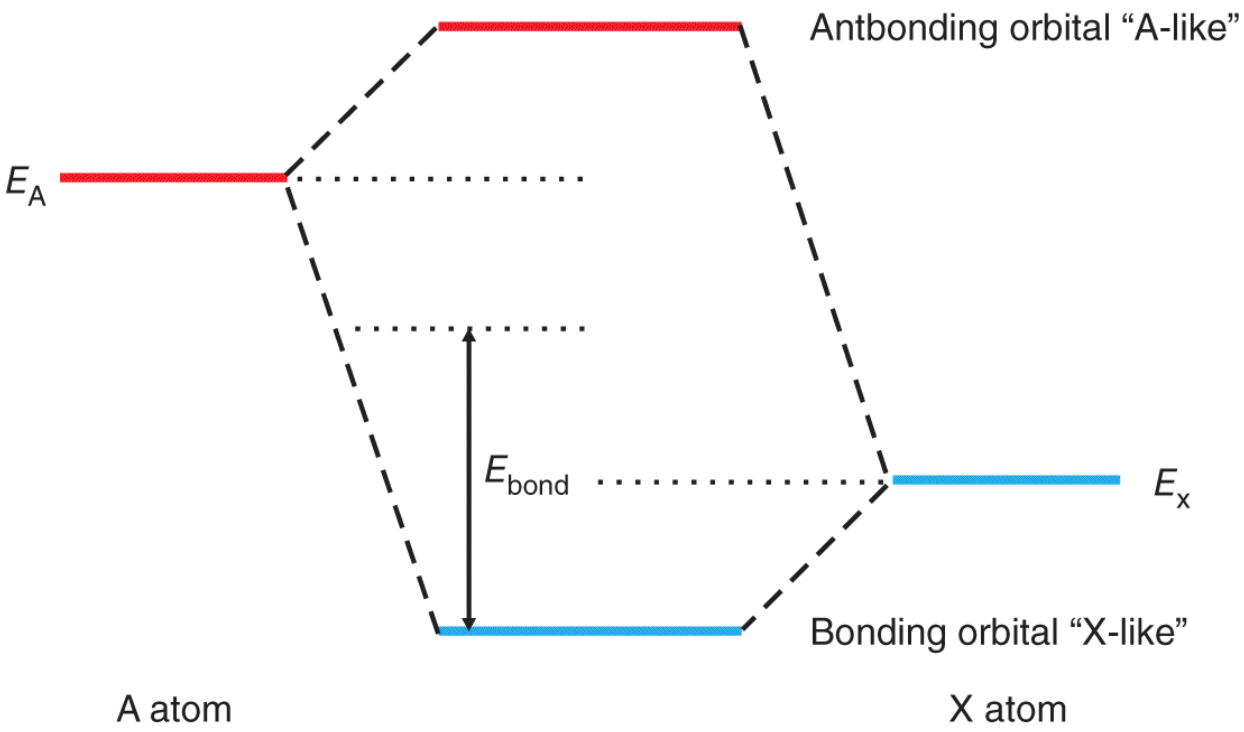


Figure 1.14 Molecular orbitals formed by a more metallic atom A and a less metallic atom X.

When the two nuclei are different, the symmetrical build-up of electron density found for identical atoms will become distorted and the electron density will pile up in the region of the non-metal atom and be depleted towards the metal atom. A covalent bond in which the electron pair is distributed unevenly is sometimes called a *polar covalent bond*. The bond will have one end that carries a small positive charge, written δ+, and the other end with a small negative charge, δ−. The charge separation gives rise to an *internal electric dipole* and such molecules are called *polar molecules*. For example, water is a polar molecule as the two O□H bonds form dipoles pointing from the oxygen towards the hydrogen atoms. However, not all molecules containing several internal dipoles show an overall dipole moment as, depending upon the molecular symmetry, they may sum to zero. These internal electric dipoles result in many important properties of molecules and molecular solids, including piezo- and ferroelectric behaviour (Section 9.2).

1.4 Metallic Bonding

In both ionic and covalent bonding, all electrons are immobile. The characteristic of metals and semiconductors is that at least the outer electrons are mobile and able to move freely through the solid under an applied voltage. As well as this, metallic bonds must act between both identical and different metallic atoms, as is revealed by the formation of numerous alloy structures as well as the structures of the elements. The bonds act between many atoms, as metal atoms in crystals often have 8 or 12 neighbours. The bonds are maintained in the liquid state, as liquid metals retain the distinguishing properties of crystalline metals.

The present-day understanding of metals and metallic bonding has arisen from the combination of two different approaches. In the first of these, electrons travel more or less freely through the structure, akin to gas atoms in a volume, and are not really associated with any particular atom, giving rise to the *free-electron gas* theory. A second approach, more chemical in nature, considers metallic bonds as a delocalisation of covalent bonds throughout the solid. The application of these ideas, called *band theory*, successfully explains the detailed physical properties of metals.

1.4.1 Molecular Orbitals and Energy Bands

The molecular orbital model of chemical bonding indicates that when two atoms approach, the outer atomic orbitals on each atom interact to form a pair of molecular orbitals, one of lower energy, the bonding orbital, and one of higher energy, the antibonding orbital. As further atoms are added to the pair, each new atomic orbital is transformed into a new molecular orbital. Calculation shows that the separation between the higher energy (antibonding) and lower energy (bonding) orbitals rapidly becomes constant as the number of atoms in the solid increases, and the new molecular orbitals fit into the energy space between them. When a large number of atoms are present, so many orbitals are incorporated that the energy space is essentially spanned by a continuum, and *energy bands* have been created (Figure 1.15). The upper part of the band will correspond to antibonding orbitals and is said to have an *antibonding character*. The lower part of the

band will correspond to bonding orbitals and have a *bonding character*.

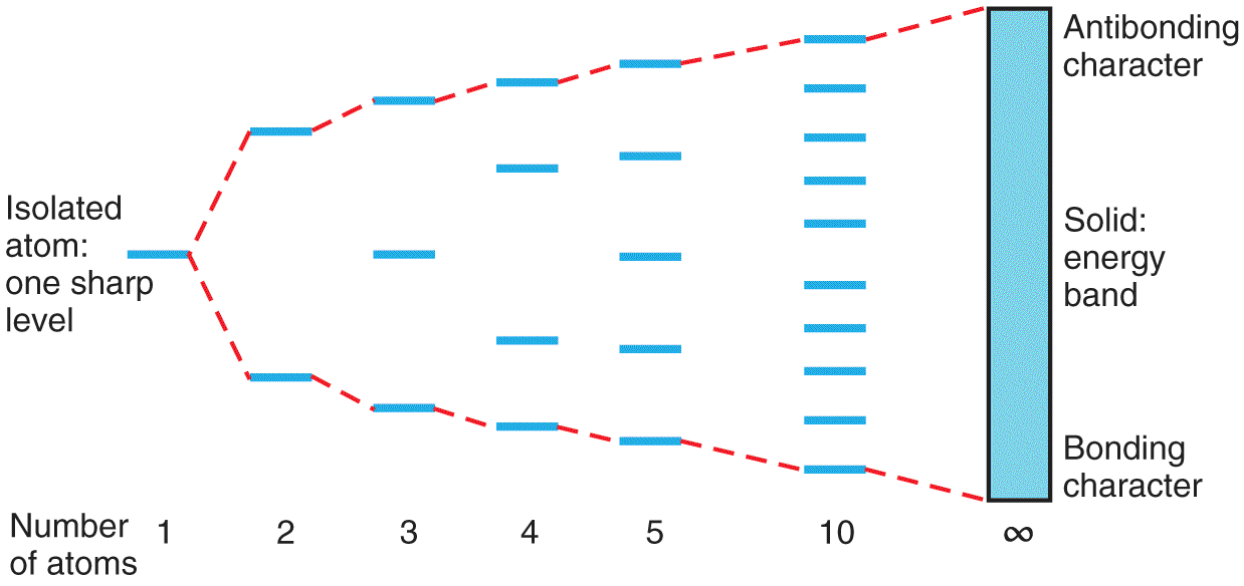


Figure 1.15 The schematic development of energy bands from delocalised molecular orbitals.

Although broadening of sharp atomic orbitals into energy bands will happen to each of the atomic orbitals on metallic atoms as they come together to form a crystal, the energy spread in an energy band is related to the extension of each orbital. The filled core electron orbitals are compact and are shielded from strong interactions with orbitals on other atoms by the valence electron orbitals. No significant overlap of these orbitals with orbitals on other atoms is possible. Thus, core electron orbitals hardly broaden and form very narrow energy bands little different from the energy levels a free atom. On the other hand, spreading of the energy bands associated with the outer orbitals is significant, more so for the heavier elements, simply because of their larger size and consequently greater interaction. Because of this, the outer orbitals on most of the heavier elements in the periodic table are transformed into rather wide bands in the solid state. Moreover, outer s and p bands are so broad that they usually overlap (Figure 1.16). Any d and f orbitals present are always shielded beneath outer s and p orbitals. They only interact weakly with orbitals on other atoms and form narrow d or f bands.

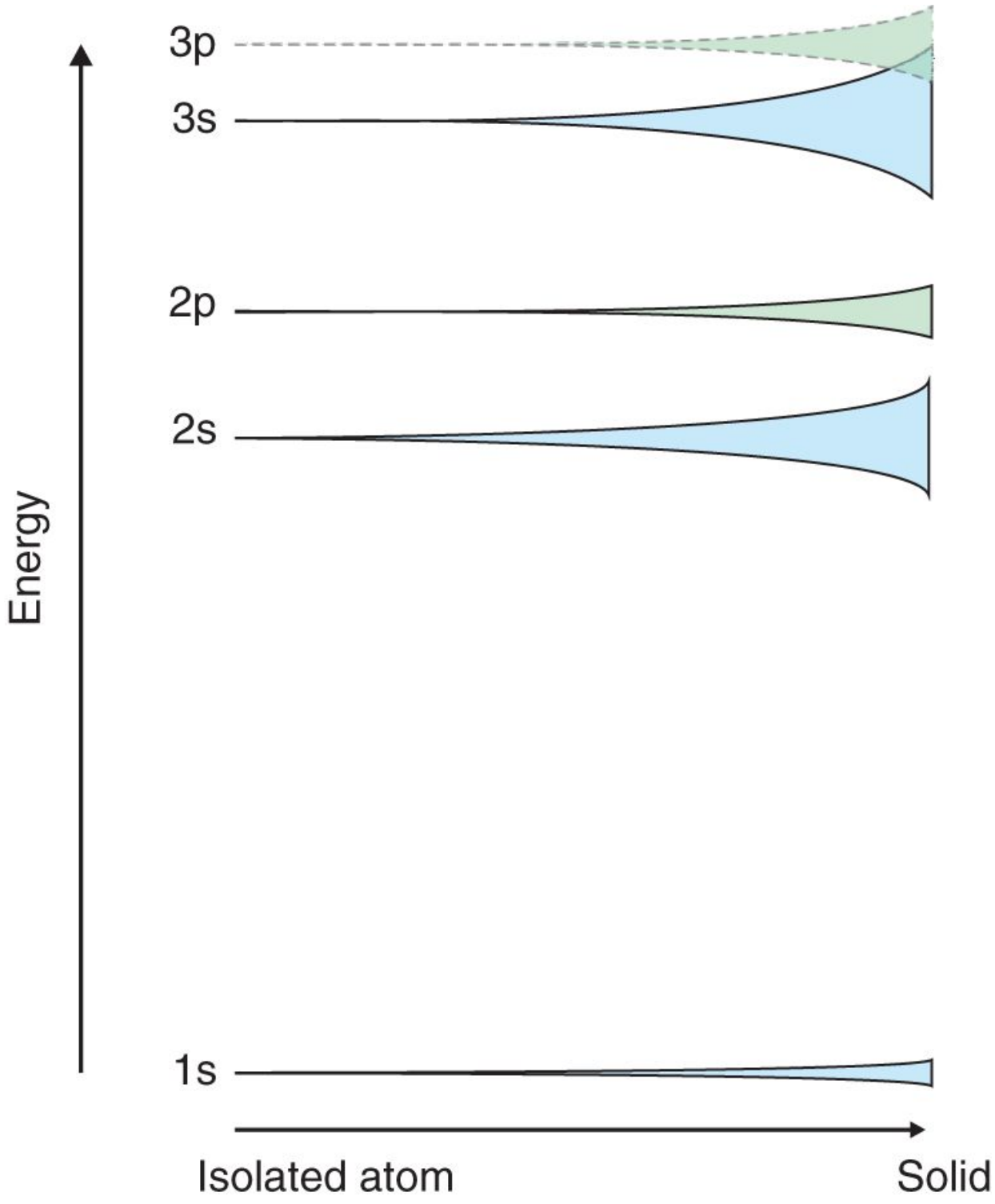


Figure 1.16 Schematic development of energy bands from atomic orbitals for magnesium metal as the atoms approach each other to form a crystal. The outermost filled 3s and empty 3p bands overlap in the metal.

1.4.2 The Free Electron Gas

The free-electron gas theory of metals started with a gross approximation: the idea that a metal contained a 'gas' of electrons, which were moving about at random in an analogous way to gas molecules in the atmosphere. The model, an adaptation of the kinetic theory of gases, was very successful. The imposition of a voltage caused the random motion to be replaced by a drift in average motion that was equated with the electric current that flowed. This simple model predicted that a metal should obey Ohm's law and that the resistivity of a metal should increase slightly with temperature. However, it could not explain some important properties. In particular, the theory was unable to account for the fact that metals had a similar specific heat to insulators. If an electron gas occurs in metals, but not in insulators, it could be proved that metals should have a molar specific heat of about $4.5R$, while insulators would have a molar specific heat of only $3R$, where R is the gas constant (Section 13.1).

The advent of quantum theory suggested that the electrons in the metal, although still free, should be treated like waves. If this was so, the Schrödinger equation could be used to determine the energy of an electron in a metal. The only variables to be specified were the dimensions of the solid and the potential energy experienced by the electron. As a first step, suppose that the potential energy is constant throughout the metal and a single electron is confined to a line of length a. (This is often referred to as the *particle in a box* model.) The potential energy is set as zero on the line (or in the box) and infinity at the extremities, so that the electron wave is completely trapped. The solutions in such a case show that the only allowed electron waves are standing waves with nodes at the fixed ends of the line (Figure 1.17). The allowed wavelengths, λ, are given by:

$$\lambda = 2a/n$$

where n is a quantum number that can take integer values 1, 2, 3, ... The wave equation describing this situation is:

$$\psi = \sqrt{(2/a)}\ \sin k\,x$$

where k is called the *wavenumber*. This equation represents a series of *plane waves*. It has quantised values given by:

$$k = 2\pi/\lambda = n\pi/a$$

(In three dimensions, $\boldsymbol{k}$ is a vector quantity, called the *wave vector*.)

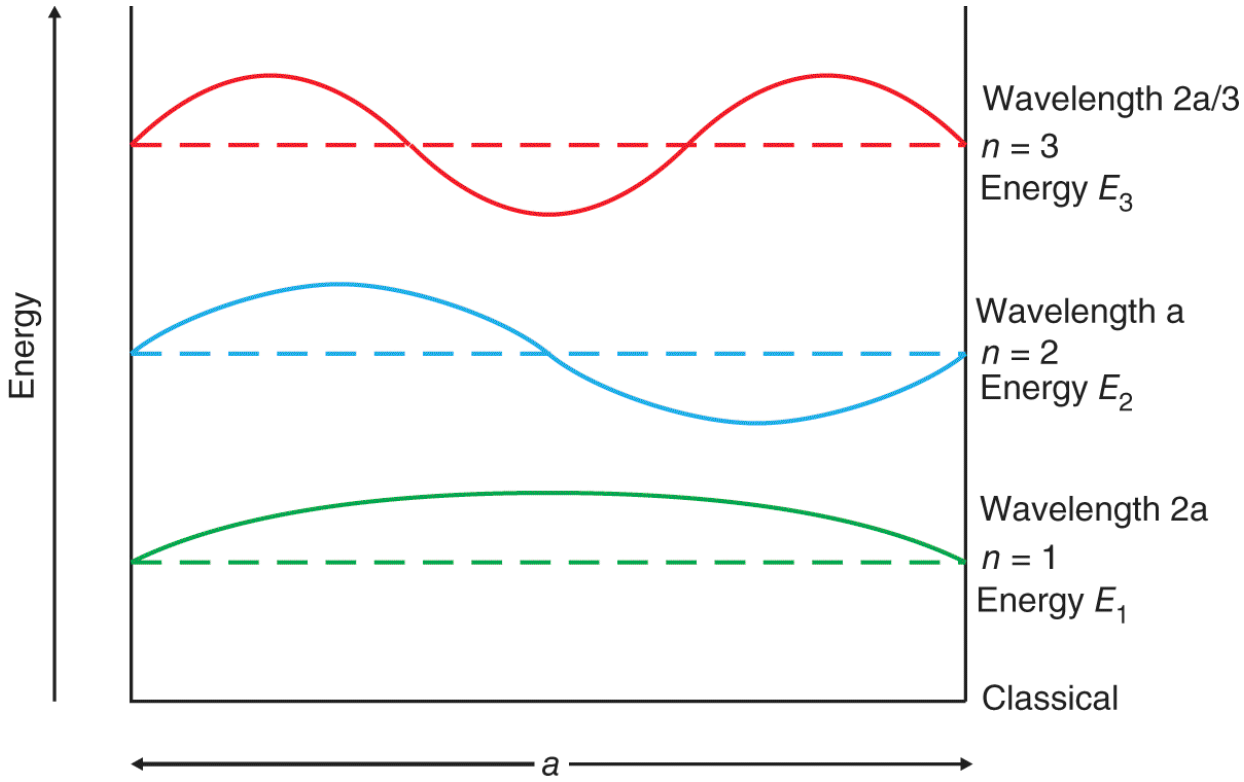


Figure 1.17 The energy levels and allowed wavelengths of an electron wave confined to a line of length a.

The electron is confined not only to certain wavelengths but also to quantised energies:

$$E_n = k^2h^2/8\pi^2m$$

where E_n is the energy of the wave associated with the quantum number n and m is the mass of the electron. The form of the energy vs. k graph is parabolic (Figure 1.18). Extending this model to three dimensions shows that the energy of a single electron in a rectangular block of metal with sides a, b, and c, lying along the three axes x, y, and z, is:

$$E(n_x, n_y, n_z) = [h^2/8m][n_x^2/a^2 + n_y^2/b^2 + n_z^2/c^2]$$

where n_x, n_y, and n_z are the quantum numbers along the x-, y-, and z-axes. For a cubic block of side a, this reduces to:

$$E(n_x, n_y, n_z) = [h^2/8ma^2][n_x^2 + n_y^2 + n_z^2]$$

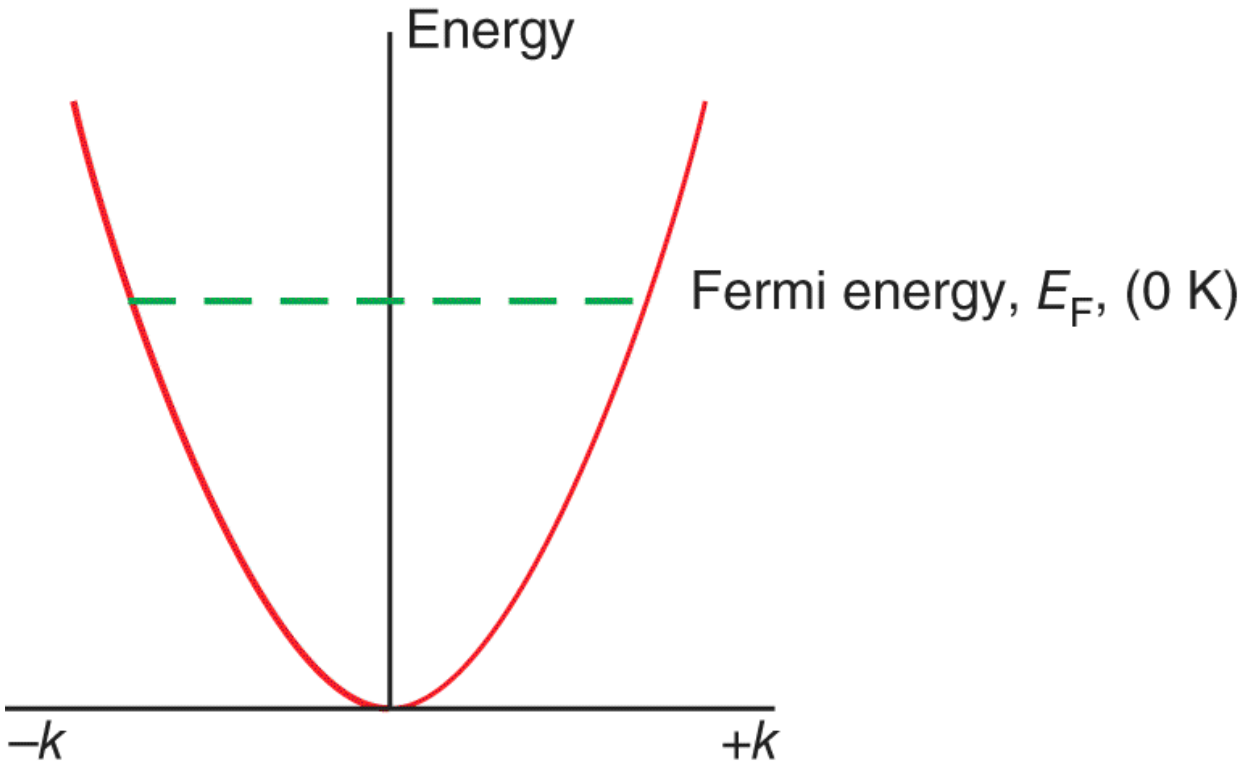


Figure 1.18 The relationship between energy, E and wavenumber, k, for an electron confined to a line. The Fermi energy, E_F, is the uppermost energy level filled at 0 K.

Wavefunctions with the same energy are said to be *degenerate*. Thus, for a cubic container, the solutions ($n_x = 1$, $n_y = 1$, $n_z = 2$), ($n_x = 1$, $n_y = 2$, $n_z = 1$), and ($n_x = 2$, $n_y = 1$, $n_z = 1$) are degenerate. The degeneracy means that the number of energy levels within a particular energy range is not a constant, but increases with increasing values of k. In order to solve the problem of the electron contribution to the specific heat of a metal, it is necessary to discover how many energy levels $N(E)$ occur in any particular range energy dE

between the energies E and $E + \delta E$. The result for a cube of volume V is:

$$N(E) = [2\pi V][8m/h^2]^{3/2}E^{1/2}$$

This equation is called the *density of states function*. This curve, like the E vs. k curve, is also parabolic in form. The number of energy levels in a small range of energy increases sharply as the energy increases. As each energy level can accommodate two electrons, the number of electrons in an energy interval between E and $E + \delta E$ is double that of the density of states.

The simplest extension of these ideas to many electrons employs a similar strategy to the orbital approximation for atoms. The energy levels calculated earlier, derived for one electron, are populated with the additional electrons, following the Pauli exclusion principle and the aufbau principle. Each level can hold just two electrons, with differing spins. To determine the overall distribution, electrons are allocated, two at a time to the energy levels, starting with the lowest energy, up to a maximum energy governed by the number of electrons available. At absolute zero, the uppermost-occupied level is at the *Fermi energy* E_F (Figure 1.18). In three dimensions this highest filled energy level takes the form of a surface, the *Fermi surface*.

At temperatures above absolute zero, the electrons are liable to gain energy. However, it is not possible for electrons in lower levels to move into slightly more energetic levels as these are already filled. Only those electrons with the highest energies, near to the Fermi surface, can move to higher energy levels. The vast body of electrons remains untouched by the rise of temperature because no empty energy levels are available to them and so do not contribute appreciably to the specific heat. Thus, the electronic heat capacity is almost negligible.

The statistics that govern the distribution of electrons between the energy levels are called *Fermi–Dirac statistics*. Fermi–Dirac statistics applies to particles with a half-integral spin, such as electrons, protons, and neutrons. Such particles are called *fermions*. No two identical fermions can occupy a single (quantised) energy level.

Fermi–Dirac statistics specify the probability, P_i, that an energy level, E_i, will be occupied is given by:

$$P_i = 1/\exp[(E_i - \mu)/k_B T] + 1$$

where μ is the chemical potential of the fermions and k_B is the Boltzmann constant. The chemical potential of electrons in a metal is equal to the Fermi energy (strictly speaking, at absolute zero), E_F, hence:

$$P_i = 1/\{\exp[(E_i - E_F)/k_B T] + 1\}$$

The distribution of electrons at temperatures other than 0 K is given by the *Fermi function*:

$$P(E) = 1/\{\exp[(E - E_F)/k_B T] + 1\}$$

where $P(E)$ is the probability than an energy level E is occupied by an electron at absolute temperature T (Figure 1.19). At T = 0 K, $P(E)$ = 1 for $E < E_F$ and 0 for $E > E_F$ and the curve has a sharp cut off. At higher temperatures the curve changes smoothly from $P(E)$ = 1 to 0. The equation also shows that at $E = E_F$, $P(E)$ = ½. Thus, the Fermi level in metals above 0 K can be defined as the energy for which $P(E)$ = ½.

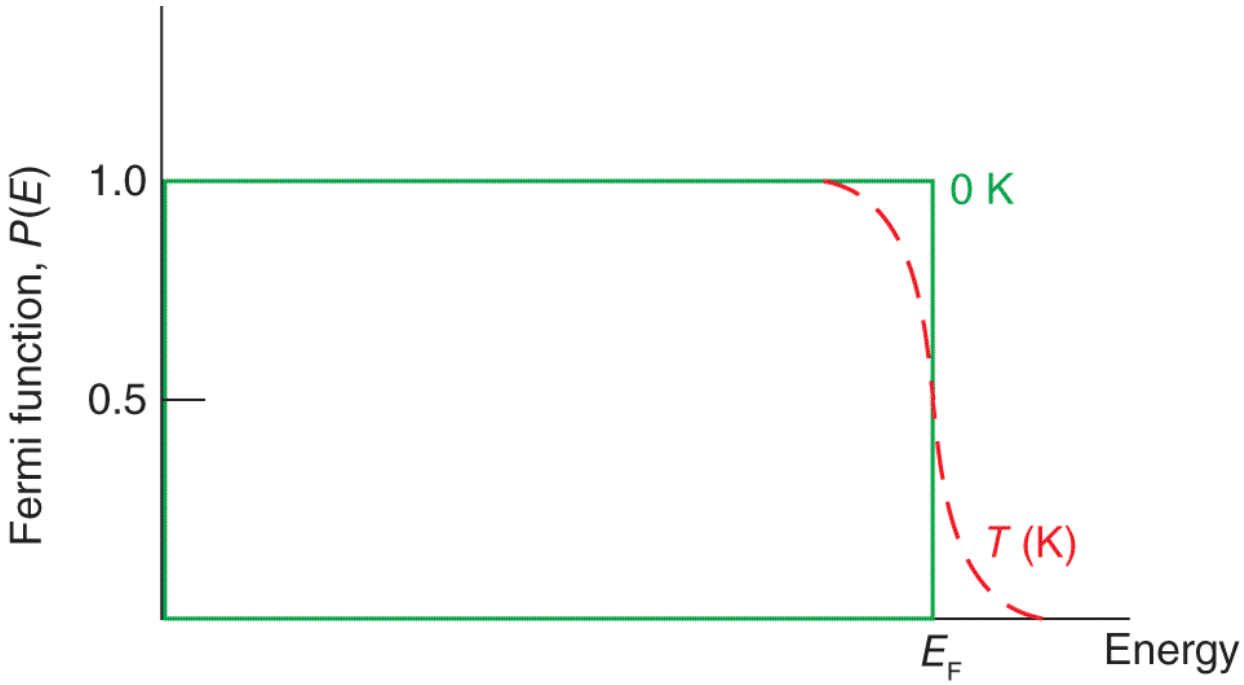


Figure 1.19 The Fermi function, $P(E)$.

1.4.3 Energy Bands

In the free-electron approach a single electron moves in a constant potential. In a solid the electron moves in a *periodic potential* that is representative of the crystal structure. The first solution to the way in which a periodic potential modifies the free electron model was given by Bloch in 1928. The solution of the Schrödinger equation was found to consist of the free electron waves multiplied by a function with the same periodicity as the crystal structure. These wave equations are called the *Bloch functions*. Two generalisations are found. Firstly, if the potential is weak the wavefunctions are similar to the free electron wavefunctions. Secondly, regardless of the potential, electron waves with a long wavelength and low energy also have wavefunctions that are almost the same as the free electron wavefunctions. In this extreme, the energy levels are closely spaced and vary with wavenumber in a parabolic fashion as described previously. The electron is 'not aware' of the atoms in the solid.

As the wavelength approaches the same dimensions as the atom spacing in the crystal, the electron waves in a solid behave in a similar way to any other waves in any medium. When a wave encounters an object of the same dimensions as the wavelength, a considerable

interaction occurs and the wave is *diffracted* (Section 12.6). In the present situation, the conditions for diffraction to occur are given by Bragg's law. For an electron wave normally incident upon planes of atoms, diffraction of the wave occurs at values of the wavenumber:

$$k = \pm n\pi/a$$

where n is an integer taking values 1, 2, 3, and so on and a is the spacing of the planes of atoms. Waves with these wavenumbers will not be able to pass through the crystal and the one-dimensional E vs. k curve is broken at these values (Figure 1.20a). Propagation will occur again when the wavenumber increases slightly. Substitution of the appropriate values into the Schrödinger equation reveals that the change in wave vector is accompanied by an energy jump from a lower value to a higher one. It is not possible for an electron to have an energy value lying between the two extremes. This is described by saying that an electron can exist within a band of allowed energies. These bands are separated from each other by bands of forbidden energies (Figure 1.20b).

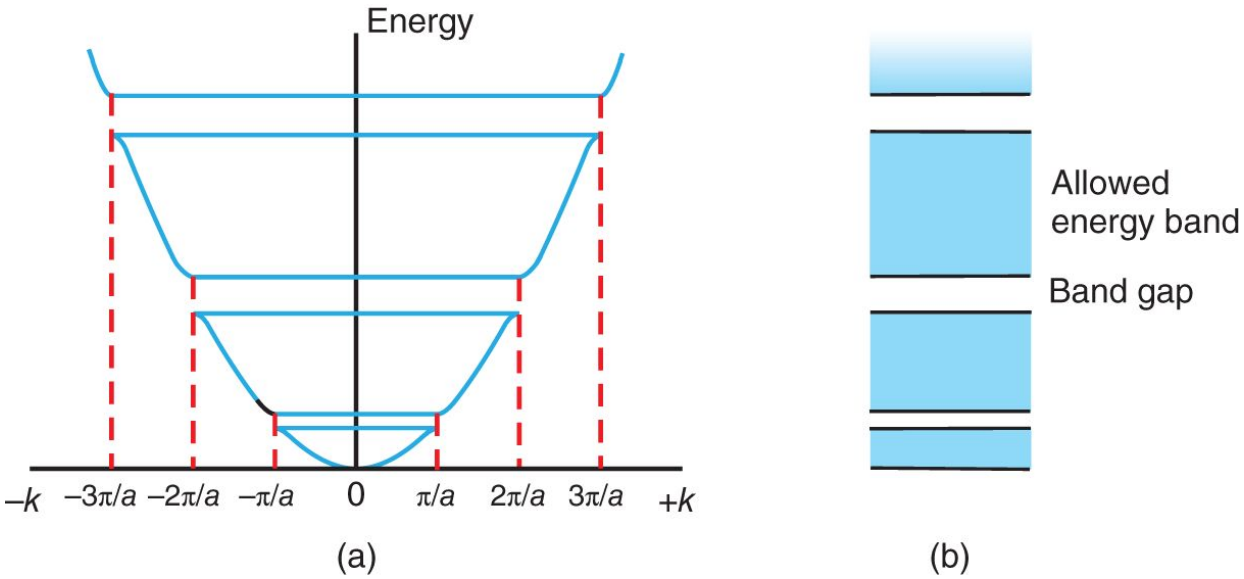


Figure 1.20 (a) The one-dimensional free-electron E vs. k curve broken into energy bands due to the periodic potential of atomic nuclei of separation a. (b) Schematic representation of energy bands.

The discontinuities in the energy vs. k curve due to electron diffraction mark the boundaries of *Brillouin zones*. The first Brillouin

zone in the one-dimensional case (Figure 1.20a) extends from k values *of* $-\pi/a$ to π/a. The second Brillouin zone lies between k values of π/a to $2\pi/a$ on the positive side of the graph and from $-\pi/a$ to $-2\pi/a$ on the negative side. The third, fourth, and so on Brillouin zones can be similarly located.

The band model accounts for the overall properties of metals. In terms of the cohesive energy of a metal, the upper energy levels of a band are antibonding in character, while the lower energy levels are bonding in character. The electrons will feed into the energy levels in a band, following the aufbau principle, with two electrons of opposing spins in each level. The core levels will be filled and be neutral in terms of bonding. Provided that the upper band is only partly filled, the lower energy bonding levels will be populated while the higher energy antibonding levels will be empty. In this case the atoms will experience an overall bonding interaction that would be expected to be roughly similar to the bonding energy of a diatomic molecule. For such a diatomic M_2 molecule, bonding is due to a filled bonding orbital combined with an empty antibonding orbital. In a solid metal, M, bonding energy is due to preferential filling of the lower bonding character part of the energy band while leaving the upper antibonding character energy band empty (Figure 1.21). Moreover, as the electrons are delocalised metallic bonds are expected to be *non-directional*.

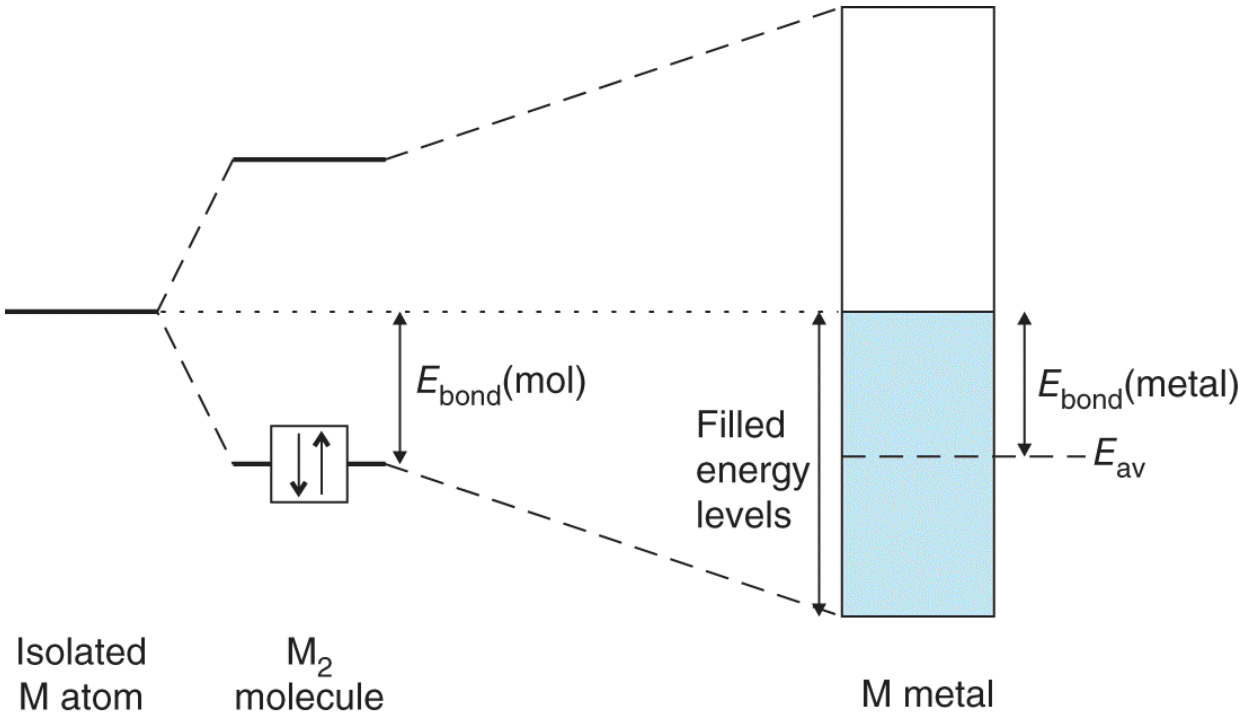


Figure 1.21 The bonding in a molecule, M_2, and a solid metal.

The fact that the uppermost band is only partly occupied also accounts qualitatively for electrical conductivity. The electrons in the highest energy band are delocalised over all the atoms in a metallic solid, and the imposition of an external voltage is able to cause the electrons to move through the material provided that there are empty energy levels available.

The loss of the outer electrons has two other consequences. As the remaining core electron orbitals are of a spherical shape, the crystal structures of many metals can be considered in terms of the packing of spherical atoms. Spheres can pack most efficiently when each is surrounded by a large number, 8–12, nearest neighbours so the delocalisation concept explains the high apparent valence of metals in crystals. Delocalisation also negates the main chemical distinguishing feature of an element and individual chemical and physical variations will be suppressed so that one metallic element is similar to another and alloys would be expected to form readily.

These metallic properties are more dependent upon the close approach of the atoms than on the relative configuration of the atoms. The geometry of the nearest neighbour atoms does not change greatly in a liquid compared to a crystal. For example, the number of nearest

neighbours to any sodium atom in a crystal of sodium metal, eight, is similar to the number in the liquid state. This means that a liquid metal should have similar properties to the solid. For example, liquid mercury shows metallic conductivity because this property is determined by the delocalisation of the outer electrons. However, Brillouin zones will not occur in liquids or glasses, as they are features of crystalline arrays.

1.4.4 Bands in Ionic and Covalent Solids

In the discussion of ionic and covalent bonding, the spreading of atomic orbitals into bands has been ignored. It is reasonable to re-examine ionic and covalent solids in band terms. Consider the archetypal ionic compound, sodium chloride. When the atoms are isolated, the lowest energy corresponds to Na and Cl atoms in their respective ground states. As the interatomic separation decreases to approximately 1.0 nm, energy is gained by the transfer of an electron from sodium to chlorine to form Na^+ and Cl^- ions. The energy of the Cl^- ions now falls below that of the Na^+ ions and the ions order on a lattice. However, even at this spacing the energy levels are still similar to those on isolated ions. As the interatomic spacing decreases further, the electron orbitals overlap. Electrons with opposed spins tend to lose energy while those with parallel spins tend to gain energy. These are the familiar bonding and antibonding interactions, and cause the narrow energy levels to broaden into bands. However, the ionic electron distribution is not changed. The band that develops on the Cl^- ions, the 3p band, is full and the band that develops on the Na^+ ions, the 3s band, remains empty (Figure 1.22). Thus, the ionic model is still a good model for the bonding in NaCl.

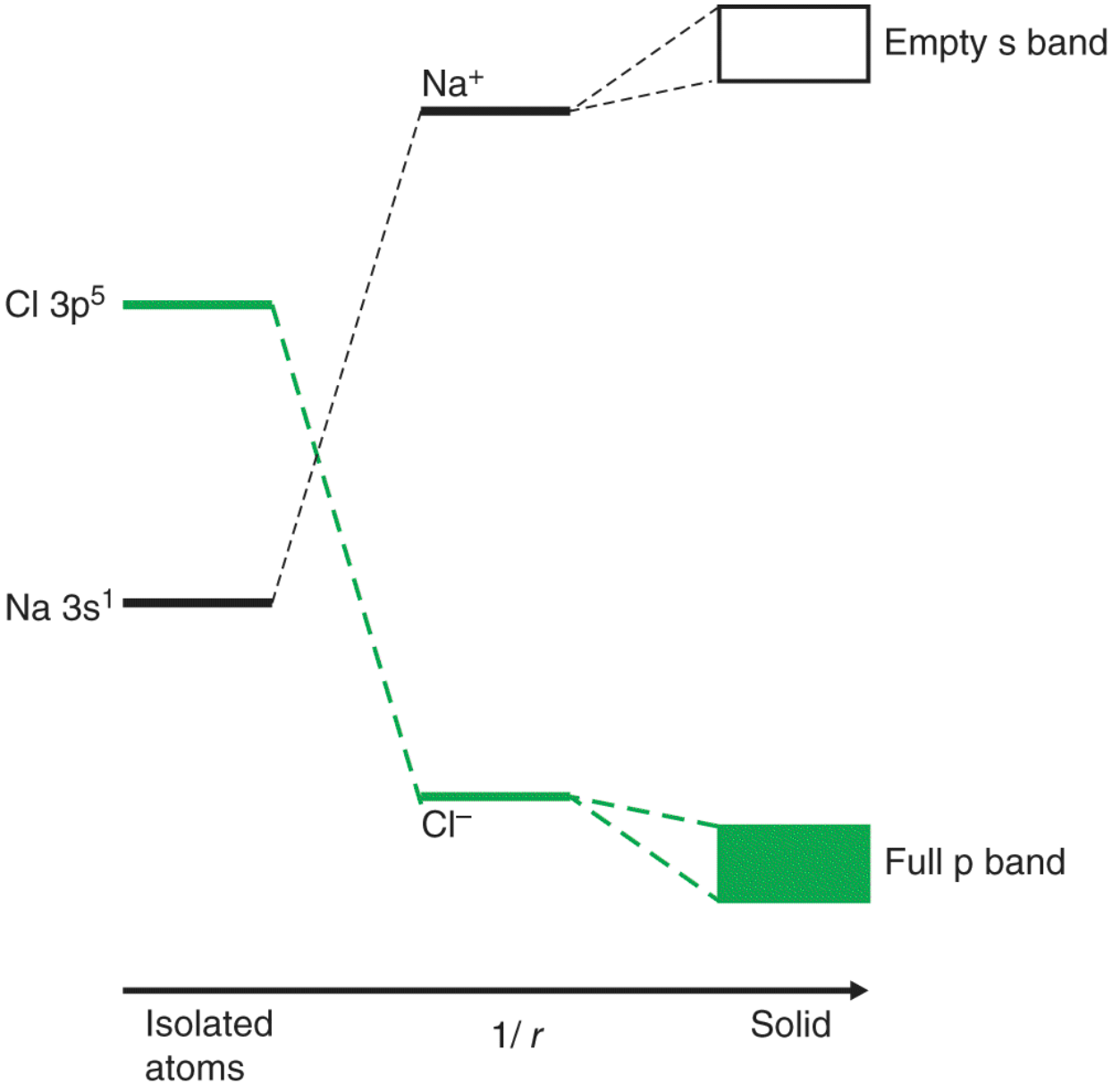


Figure 1.22 Schematic development of energy bands from isolated atoms for sodium chloride, NaCl. For clarity, the core energy levels of both atoms have been ignored and only relative energy changes are included.

In the case of a covalently bonded crystal such as germanium, which is composed of a tetrahedral array of germanium atoms linked by sp^3-hybrid bonding, a modification of the molecular orbital model is again needed. Isolated Ge atoms have an outer electron configuration s^2p^2, which is replaced by four degenerate sp^3 hybrid orbitals as the atoms are brought closer (Figure 1.23). Each sp^3 orbital on an atom

contains one electron and overlap with adjacent orbitals causes strong bonds with a tetrahedral geometry to form. As the atoms approach further, each single sp^3 energy level widens into a band, the lower half of which is bonding and the upper part antibonding. The band is able to contain a maximum of eight electrons. As each atom donates one electron per sp^3 energy level, the band is half full. Continued approach causes the antibonding part of the band to increase in energy due to the repulsion between the parallel electrons and the bonding part to decrease in energy due to the favourable interaction between the spin-paired electrons. At a critical separation, the single band is split into two bands separated by an energy gap. The lower of these bands is usually called the *bonding band* and the upper the *antibonding band* when the chemistry of the material is discussed, or the *valence band* and *conduction band* in semiconductor physics. The electrons, one in each sp^3 orbital, that combine in a covalent bond end up in the lowest bonding band, which will be completely full. The upper antibonding band will be completely empty. The cohesive energy in these crystals is, in fact, due to the appearance of the energy gap between the upper (empty) and lower (filled) bands.

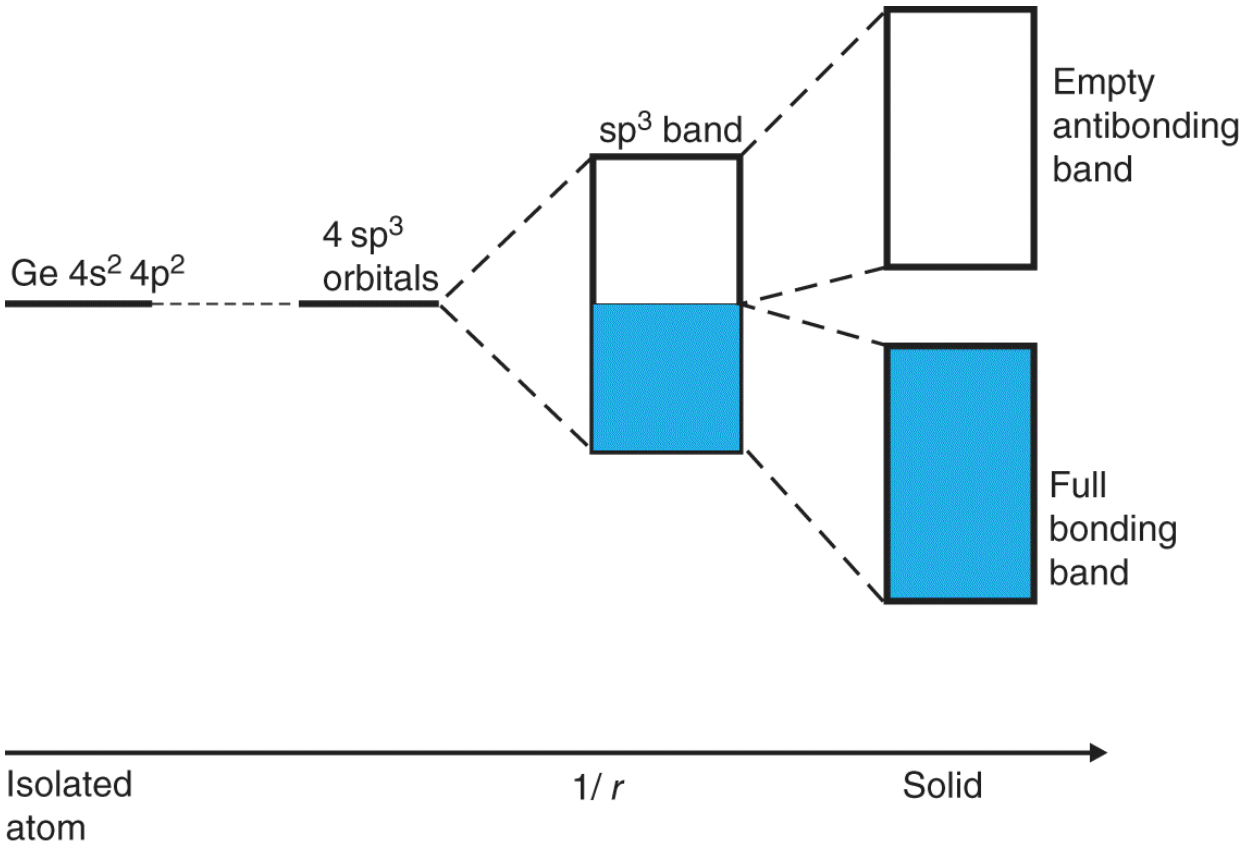


Figure 1.23 Schematic development of energy bands from isolated atoms for germanium, Ge. For clarity, the core energy levels of both atoms have been ignored and only relative energy changes are included.

The elements carbon (diamond), C, silicon, Si, germanium, Ge, and one form of tin, α-Sn, all crystallise with the same structure. The size variation of this series of atoms modifies the band formation in a predictable way. The smallest atom, carbon, C, has the smallest degree of orbital interaction and hence the narrowest bands. The energy gap is large and diamond has a very high cohesive strength. As the atoms increase in size, the orbital interactions increase, the bands widen, the energy gap narrows, and the cohesive energy falls. Unlike diamond, α-tin is a weak solid, easily crushed.

While this description is quite different than that of covalent bonding, the separation of the energy bands in the solid is similar to the separation of the bonding and antibonding molecular orbitals that would be found on a diatomic C_2, Si_2, Ge_2, or Sn_2 molecule with a similar interatomic separation to the atoms in the crystal. The

covalent bond picture is, therefore, closely related to the band picture. In the covalent model, the bond energy is related to the separation of the molecular orbitals. In the solid, the cohesive energy is related to the separation of the energy bands. With this limitation in mind, the covalent bonding model is adequate for many solids.

1.5 Weak Chemical Bonds

Although the strong chemical bonds described earlier play a dominant role in determination of the overall properties of solids, weak interactions are often of vital importance, especially in living organisms. For example, noble gas atoms experience weak interatomic forces that lead to liquefaction and, except for helium, solidification at low temperatures, while hydrogen bonding is vital in the functioning of the molecules central to life, such as proteins and DNA (Table 1.5).

Table 1.5 Bonds between atoms, ions, and molecules.

Type of bond	Approximate energy (kJ mol^{-1})	Species involved
Covalent	350	Atoms with partly filled orbitals
Ionic	250	Cations and anions
Metallic	200	Metal atoms
Ion–dipole	15	Ions and polar molecules
Dipole–dipole	2	Stationary polar molecules
Dipole–dipole	0.3	Rotating polar molecules
Dispersion	2	All atoms and molecules
Hydrogen bond	20	N, O, or F plus H

The strongest of the weak bonds are *hydrogen bonds*. These form when a hydrogen atom lies between two highly electronegative atoms, either fluorine, oxygen, chlorine, or nitrogen. The bond results from the interaction of the small positive charge, δ+, found in dipolar molecules containing hydrogen with the partial charge of δ– located on the electronegative partner, especially the exposed lone pair electrons on atoms such as oxygen and nitrogen. The hydrogen atom has an ambiguous position in the bond. At low temperatures it adopts a position nearer to one or other of the electronegative atoms, while at high temperatures it is found, on average, midway between them. In

general the two links making up the bond, O□H and H ... O, are not in the same straight line. The angle between them commonly deviates from 180° by 10° to 20°, and sometimes by much more.

Because hydrogen bonds are comparatively weak, it follows that they are not only easily ruptured but they are also formed with equal ease. Thus, they form in all appropriate materials at normal temperatures and are important in compounds such as water, H_2O, hydrogen fluoride, HF, and potassium hydrogen fluoride, KHF_2. The existence of hydrogen bonds dramatically changes many of the properties of the material. For example, HF, H_2O, and ammonia, NH_3, are characterised by melting points, boiling points, and molar heats of vaporisation that are abnormally high in comparison with those of similar molecules unable to link in this way. The fact that water is liquid on the Earth at normal temperatures is largely due to hydrogen bonding. In living organisms, hydrogen bonding is of great importance in controlling the folded (*tertiary*) structure of proteins. The tertiary structure largely determines the biological activity of the molecule and mistakes in the folding can lead to serious illness. Hydrogen bonding also endows solids with significant physical properties, such as ferroelectric behaviour (Section 9.3).

Permanent dipoles are usually found on molecules containing two atoms with very different electronegativities. For example, a molecule of HCl has the covalent bond slightly deformed to give rise to a region of depleted electron density (enhanced positive charge, δ+), associated with the hydrogen atom, and a region of increased electron density (augmented negative charge, δ−), associated with the chlorine atom. The resultant dipole has a dipole moment of 3.60×10^{-30} C m. Water, an angular molecule, has two internal dipoles, with the δ+ region residing on each hydrogen and the δ− region on the oxygen. These add together so that the resultant is directed towards the oxygen atom, and is augmented by the negative charge associated with the two lone pairs of electrons to give a dipole moment, of 6.17×10^{-30} C m.

The charges making up the dipole can interact with ions to give *ion–dipole interaction energies* of about 15 kJ mol^{-1}. The hydration of cations, in both water solution and solid hydrates, is mainly due to ion–dipole effects. Permanent dipoles can also interact with other

dipoles in *dipole–dipole interactions*, which vary from approximately 2 kJ mol^{-1} for fixed molecules to about one tenth of this value if the molecules are free to rotate.

The noble gases are monatomic at normal temperatures. On cooling, helium, the lightest, turns to a liquid (with very curious properties) at 4.2 K, the lowest known boiling point of an element. Helium can only be turned into a solid by applying pressure. The other members of the family can be liquefied and solidified by cooling. This condensation is due to weak interactions between the outer electrons on the atoms. Fleeting fluctuations in the electron clouds create momentary dipoles, which lead to a weak attraction and, at low temperatures, condensation. The force of attraction is called the *London* or *dispersion force* and the interaction is called the *van der Waals bonding* and occurs between all atoms and molecules. Although the bond energy is only about 2 kJ mol^{-1}, it is responsible for the liquid states of many molecular species, including H_2 and benzene. The strength of the interaction increases as the size of the molecules increase. Because of this, large molecules tend to exist as solids, smaller ones as liquids, and light molecules as gases. This trend is exemplified by the smooth increase in melting and boiling points of the alkanes, which have a series formula C_nH_{2n+2} (Figure 1.24). The lightest of these are gases at room temperature (methane, ethane, propane), while the heavier ones are liquids (octane, C_8H_{18}), and then waxy solids ($C_{18}H_{38}$ and higher).

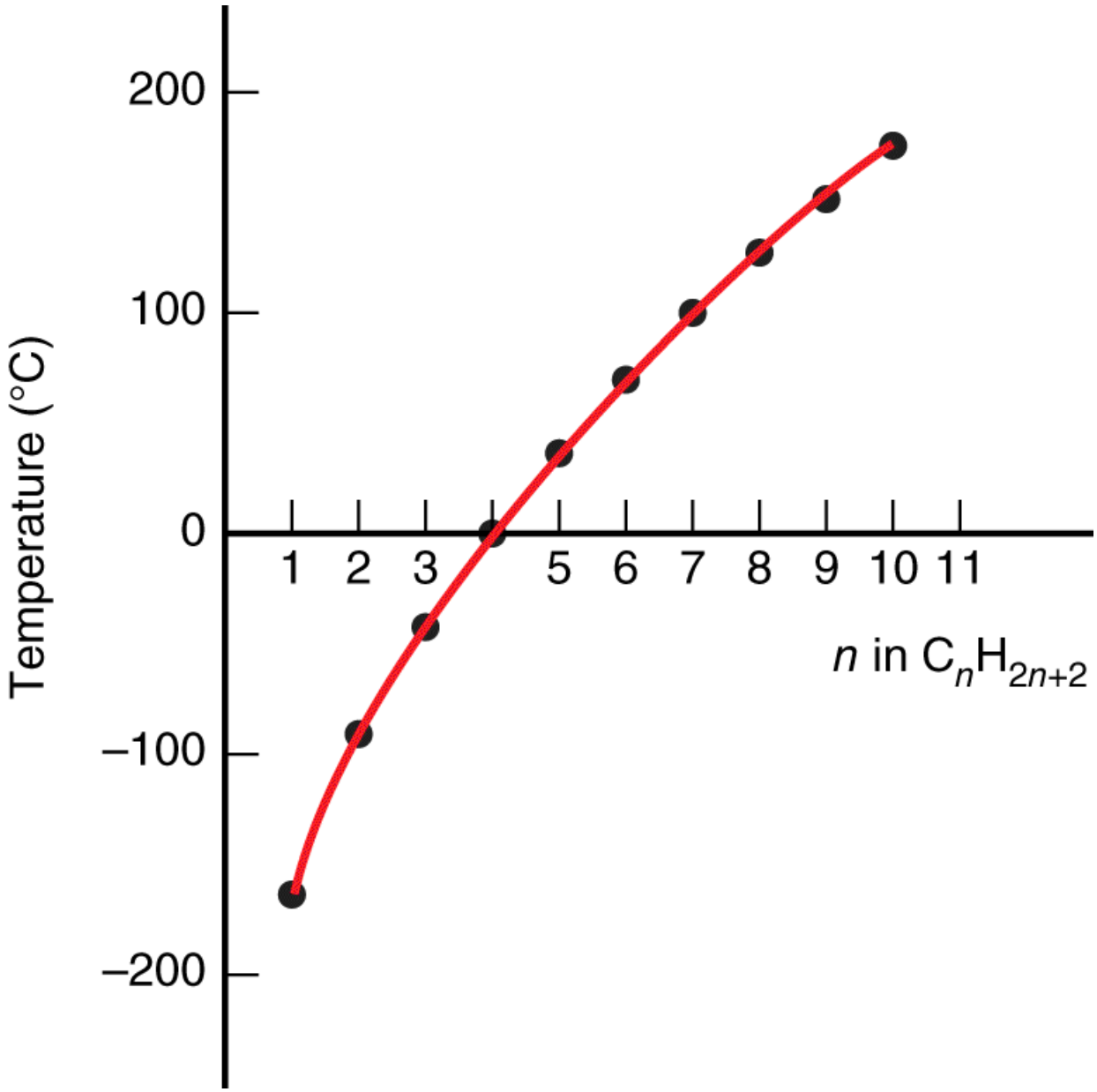


Figure 1.24 The boiling points of the alkanes, C_nH_{2n+2}, plotting against n, which is proportional to the size of the molecule.

A commonly used form of the interaction energy between a pair of atoms or molecules is the *Lennard-Jones potential*, V_{LJ}:

$$V_{LJ} = Ar^{-12} - Br^{-6}$$

where A and B are constants and r is the distance between the atoms or molecules. The first term in this equation is a repulsive energy

term and the second an attractive energy. The potential energy passes through a minimum V_m at a distance r_0. Under normal circumstances, this would represent the bonding energy of a pair of atoms or molecules, at an equilibrium separation of r_0. The Lennard-Jones potential can be written in terms of V_m as:

$$V_{LJ} = 4V_m[(r(0)/r)^6 - (r(0)/r)^{12}]$$

where $r(0)$ is the value of r when V_{LJ} is zero (Figure 1.25).

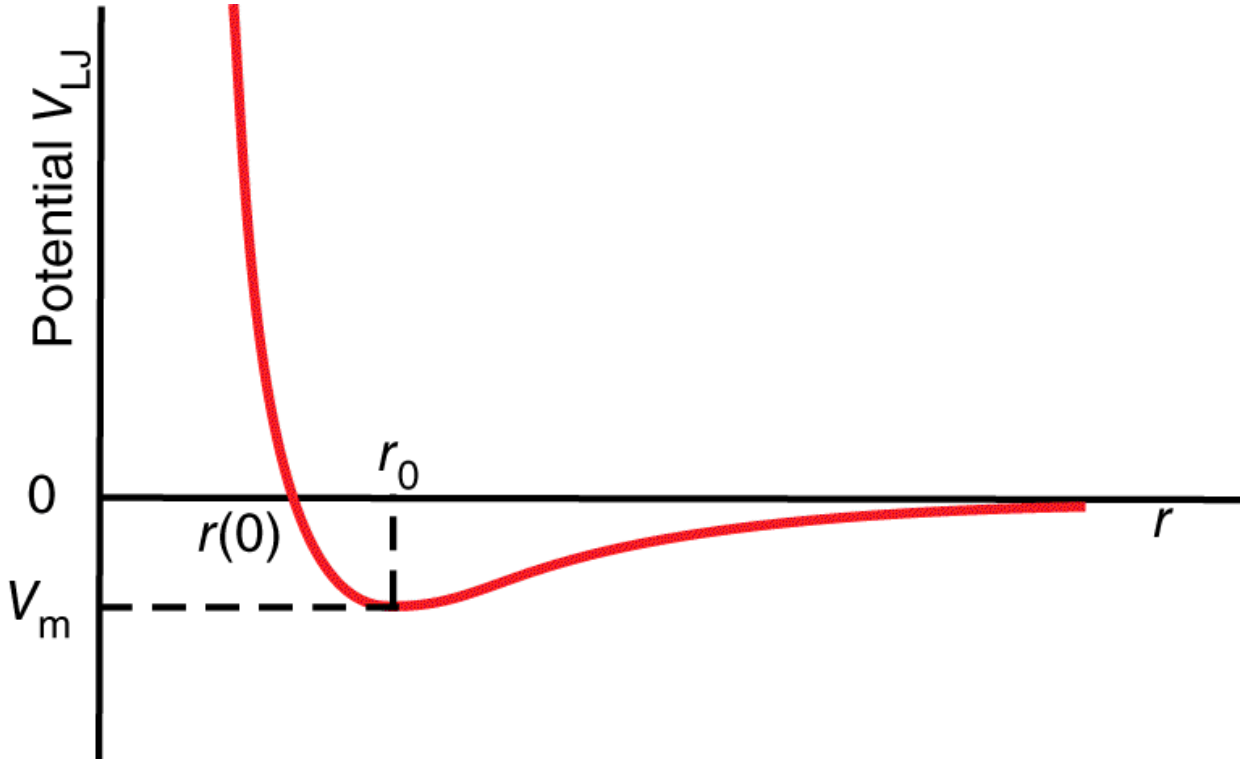


Figure 1.25 The Lennard-Jones potential between two atoms separated by a distance r.

Thermal energy is taken to be of the order of k_BT, where k_B is Boltzmann's constant and T is the absolute temperature. In cases where the energy of the bond, V_m, is greater than k_BT, one can expect pairs of atoms or molecules to be stable, and a liquid phase to condense. When V_m is less than k_BT, the bond would be expected to be too weak to hold the pair together, and a gas is likely.

The range over which bonding forces are significant varies widely. Covalent bonds act over a few nanometres only. Interactions that are essentially electrostatic in nature, as in ionic bonds, operate over larger distances, and are proportional to $1/r$, where r is the inter-ionic distance. Ion–dipole interactions decrease more rapidly, being proportional to $1/r^2$. Dipole–dipole interactions vary as $1/r^3$ for static dipoles and $1/r^6$ for rotating dipoles. Dispersion forces also decrease as $1/r^6$.

1.6 Computation of Material Properties

The calculation of materials properties using the bonding models described above has a long pedigree. The shortcomings of traditional bonding concepts are overcome by using the theoretical approach called *density functional theory*. (A functional is a function of a function. In this case, the energy of the wavefunction is a function of the electron density which is a function of electron position.) Density functional theory has been successfully applied to modelling the properties of a wide range of solids using an approach rather akin to that described for bonding in metals. The important feature of this theoretical advance is that the correlation energy, together with the exchange energy, called the *exchange–correlation* energy, can be expressed solely in terms of the electron density distribution in the material. Despite this superficially simple statement, it remains a central task of density functional calculations to find appropriate expressions for the exchange–correlation energy. One way is to use a model of a uniform electron gas occupying a volume with a uniform distribution of positive charge, the so-called *jellium model*, clearly similar to the electron gas model of a metal.

There still remains the problem of how one is to represent the wavefunctions of the system. For many materials a simple approach is satisfactory, that is to use plane waves similar to those derived as solutions to the wave equation for a free electron gas. In this case the many electron wavefunction is written as the sum of plane waves. A problem arises because near to atomic nuclei the number of plane waves needed increases enormously – too much for simple computation. This problem is removed by replacing the strong potential field of each nucleus with a weak *pseudopotential*, which reduces the number of plane waves needed, the plane wave basis set, to manageable proportions.

Further Reading

The Following References Expand the Material in this Chapter:

Atkins, P., de Paula, J., and Friedman, R. (2009). *Quanta, Matter, and Change*. Oxford: Oxford University Press, Chapters 4, 5, 6.

A Dictionary of Quantum Mechanical Language and Expressions is:

Atkins, P.W. (1991). *Quanta*, 2nde. Oxford: Oxford University Press.

Ionic Radii are Discussed and Tabulated by:

Shannon, R.D. (1976). *Acta Crystallogr. A* 32: 751–756.

Shannon, R.D. and Prewitt, C.T. (1969). *Acta Crystallogr.* B25: 925–946; Acta Crystallogr. B, 26, 1046 (1970).

The Computation of Properties is Described in:

Lesar, R. (2013). *Introduction to Computational Materials Science*. Cambridge: MRS/Cambridge University Press.

Problems and Exercises

Quick Quiz

1.1 An orbital is:

a. A bond between an electron and a nucleus.

b. A region where the probability of finding an electron is high.

c. An electron orbit around an atomic nucleus.

1.2 The configuration of an atom is:

a. The number of electrons around the nucleus.

b. The electron orbitals around the nucleus.

c. The arrangement of electrons in the various orbitals.

1.3 The outer electron configuration of the noble gases is:

a. ns^2np^6.

b. $ns^2np^6(n + 1)s^1$.

c. ns^2np^5.

1.4 The valence electron configuration of the alkali metals is:

a. ns^2.

b. np^1.

c. ns^1.

1.5 How many electrons can occupy orbitals with $n = 3$, $l = 2$?

a. 6 electrons.

b. 10 electrons.

c. 14 electrons.

1.6 How many electrons can occupy the 4f orbitals?

a. 14

b. 10

c. 7

1.7 The Russell–Saunders coupling is:

a. A procedure to obtain the energy of many-electron atoms.

b. A description of atomic energy levels.

c. A procedure to obtain many-electron quantum numbers.

1.8 A term symbol is

a. A label for a set of atomic energy levels.

b. A label for an orbital.

c. An electron configuration.

1.9 An atom has a term ^{1}S. What is the value of the spin quantum number, *S*?

a. ½.

b. 0.

c. 1.

1.10 An atom has a term ^{1}S. What is the value of the orbital quantum number, *L*?

a. 2.

b. 1.

c. 0.

1.11 An anion is:

a. An atom that has lost a small number of electrons.

b. An atom that has gained a small number of electrons.

c. A charged atom stable in solutions.

1.12 Ionic bonds are formed by which process:

a. Electron sharing.

b. Electron transfer.

c. Electron delocalisation.

1.13 The Madelung energy is:

a. The bond energy of an ionic crystal.

b. The lattice energy of an ionic crystal.

c. The electrostatic energy of an ionic crystal.

1.14 Cations are generally:

a. Smaller than anions.

b. Bigger than anions.

c. The same size as anions.

1.15 Covalent bonds are formed by which process:

a. Electron sharing.

b. Electron transfer.

c. Electron delocalisation.

1.16 Covalent bonds are:

a. Strongly directional.

b. Completely non-directional.

c. Variable in direction.

1.17 A σ bond is characterised by:

a. Being formed between identical atoms.

b. Being formed by two s orbitals.

c. Radial symmetry about the bond axis.

1.18 Bonding molecular orbitals have:

a. The same energy as antibonding orbitals.

b. A lower energy than antibonding orbitals.

c. A higher energy than antibonding orbitals.

1.19 sp^2 hybrid orbitals:

a. Point towards the vertices of a tetrahedron.

b. Point towards the vertices of a triangle.

c. Point towards the vertices of a cube.

1.20 Multiple bonds between two similar atoms:

a. Always consist of the same types of molecular orbital.

b. Always consist of the same types of hybrid orbital.

c. Always consist of different types of molecular orbital.

1.21 Metallic bonds are formed by which process:

a. Electron sharing.

b. Electron transfer.

c. Electron delocalisation.

1.22 A metallic bond is predominantly:

a. Strongly directional.

b. Completely non-directional.

c. Partly directional.

1.23 The Fermi energy is:

a. The uppermost energy level filled.

b. The highest energy in a band.

c. The energy of an electron in a band.

1.24 Energy bands can form:

a. Only in crystals.

b. Only in metals.

c. Only in electrical conductors.

1.25 Hydrogen bonding is found in:

a. Solid and liquid hydrogen.

b. Hydrocarbons.

c. Compounds containing oxygen and hydrogen.

1.26 Van der Waal's bonds are due to:

a. Dispersion forces.

b. Ion–dipole forces.

c. Dipole–dipole forces.

Calculations and Questions

1.1 What energy is required to liberate an electron in the $n = 3$ orbital of a hydrogen atom?

1.2 What is the energy change when an electron moves from the $n = 2$ orbital to the $n = 6$ orbital in a hydrogen atom?

1.3 Calculate the energy of the lowest orbital (the ground state) of the single electron hydrogen-like atoms with $Z = 2$ (He^{+}) and 3 (Li^{2+}).

1.4 What are the frequencies and wavelengths of the photons emitted from a hydrogen atom when an electron makes a transition from $n = 4$ to the lower levels $n = 1$, 2 and 3?

1.5 What are the frequencies and wavelengths of the photons emitted from a hydrogen atom when an electron makes a transition from $n = 5$ to the lower levels $n = 1$, 2 and 3?

1.6 What are the frequencies and wavelengths of photons emitted when an electron on a Li^{2+} ion makes a transition from $n = 3$ to the lower levels $n = 1$ and 2?

1.7 What are the frequencies and wavelengths of photons emitted when an electron on a He^{+} ion makes a transition from $n = 4$ to the lower levels $n = 1$, 2 and 3?

1.8 Sodium lights emit yellow colour, with photons of wavelength 589 nm. What is the energy of these photons?

1.9 Mercury lights emit photons with a wavelength 435.8 nm. What is the energy of the photons?

1.10 What are the possible quantum numbers for an electron in a 2p orbital?

1.11 Titanium has the term symbol ^{3}F. What are the possible values of J? What is the ground state level?

1.12 Phosphorus has the term symbol ^4S. What are the possible values of J? What is the ground state level?

1.13 Scandium has a term symbol ^2D. What are the possible values of J? What is the ground state level?

1.14 Boron has a term symbol ^2P. What are the possible values of J? What is the ground state level?

1.15 What is the splitting g_J, for sulfur, with ground state $^3P_2$?

1.16 What is the splitting g_J, for iron, with a ground state $^5D_4$?

1.17 Draw a diagram equivalent to Figure 1.6, for the *ground state* of a chlorine atom, with a ground state $^2P_{3/2}$.

1.18 Write out the electron configuration of the ions: Cl^-, Na^+, Mg^{2+}, S^{2-}, N^{3-}, Fe^{3+}.

1.19 Write out the electron configuration of the ions: F^-, Li^+, O^{2-}, P^{3-}, Co^{2+}.

1.20 Write the symbols of the ions formed by oxygen, hydrogen, sodium, calcium, zirconium, and tungsten.

1.21 Write the symbols of the ions formed by iron, chlorine, aluminium, sulfur, lanthanum, and tantalum.

1.22 Calculate the lattice energy of the ionic oxide CaO, which has the halite (NaCl) structure. $\alpha = 1.75$, $n = 9$, $r_0 = 0.24$ nm.

1.23 Calculate the lattice energy of the ionic halides NaCl and KCl, which have the halite (NaCl) structure. $\alpha = 1.75$, $n = 9$, r_0 (NaCl) = 0.281 nm, r_0 (KCl) = 0.314 nm.

1.24 Calculate the lattice energy of the ionic halides NaBr and KBr, which have the halite (NaCl) structure. $\alpha = 1.75$, $n = 9$, r_0 (NaBr) = 0.298 nm, r_0 (KBr) = 0.329 nm.

1.25 Determine the number of free electrons in gold, assuming that each atom contributes one electron to the 'electron gas'. The

molar mass of gold is 0.196 97 kg mol^{-1}, and the density is 19 281 kg m^{-3}.

1.26 Determine the number of free electrons in nickel, assuming that each atom contributes two electrons to the 'electron gas'. The molar mass of nickel is 0.058 69 kg mol^{-1}, and the density is 8907 kg m^{-3}.

1.27 Determine the number of free electrons in copper, assuming that each atom contributes one electron to the 'electron gas'. The molar mass of copper is 0.063 55 kg mol^{-1}, and the density is 8933 kg m^{-3}.

1.28 Determine the number of free electrons in magnesium, assuming that each atom contributes two electrons to the 'electron gas'. The molar mass of magnesium is 0.024 31 kg mol^{-1}, and the density is 1738 kg m^{-3}.

1.29 Determine the number of free electrons in iron, assuming that each atom contributes two electrons to the 'electron gas'. The molar mass of iron is 0.055 85 kg mol^{-1}, and the density is 7873 kg m^{-3}.

1.30 Calculate the lowest energy level of a free electron in a cube of metal 1 cm on edge and compare it with thermal energy at room temperature, given by $k_B T$, where k_B is Boltzmann's constant and T is the absolute temperature.

1.31 Estimate the Fermi energy of silver. The molar mass of silver is 0.1079 kg mol^{-1}, the density is 10 500 kg m^{-3}, and each silver atom contributes one electron to the 'electron gas'.

1.32 Estimate the Fermi energy of sodium. The molar mass of sodium is 0.022 99 kg mol^{-1}, the density is 966 kg m^{-3}, and each sodium atom contributes one electron to the 'electron gas'.

1.33 Estimate the Fermi energy of calcium. The molar mass of calcium is 0.044 08 kg mol^{-1}, the density is 1530 kg m^{-3}, and each calcium atom contributes two electrons to the 'electron gas'.

1.34 Estimate the Fermi energy of aluminium. The molar mass of aluminium is 0.026 98 kg mol^{-1}, the density is 2698 kg m^{-3},

and each aluminium atom contributes three electrons to the 'electron gas'.

1.35 Draw the wavefunctions and the probability of locating an electron at a position x for an electron trapped in a one-dimensional potential well, with internal potential 0 and external potential infinity.

1.36 Make accurate plots of the Fermi function, which expresses the probability of finding an electron at an energy E, for temperatures of 0, 300, 1000, and 5000 K.

1.37 37 The Lennard-Jones constants for argon are $A = 1.78 \times 10^{-134}$ J m^{12} and $B = 1.08 \times 10^{-77}$ J m^{6}.

a. Plot the attractive and repulsive potential energies;

b. Estimate the minimum potential energy of the pair, which can be considered the bonding energy of a molecule of argon;

c. The equilibrium interatomic separation of this molecule.

1.38 By equating the thermal energy ($k_B T$, where k_B is the Boltzmann constant), with the bonding energy, estimate the temperature at which argon atoms are likely to start to form pairs, and so form a liquid, using the data in Question 1.37. Compare this to the boiling point of argon (–185.9 °C).

1.39 If the atoms in liquid argon are surrounded by 12 nearest neighbours on average, estimate the energy of evaporation of the liquid.

1.40 40 The Lennard-Jones constants for neon are $A = 4.39 \times 10^{-136}$ J m^{12} and $B = 9.30 \times 10^{-79}$ J m^{6}. Calculate:

a. The bonding energy;

b. The equilibrium separation; of a pair of neon atoms.

1.41 The Lennard-Jones constants for helium are $A = 4.91 \times 10^{-137}$ J m^{12} and $B = 4.16 \times 10^{-80}$ J m^{6} and for xenon are $A = 2.54 \times 10^{-133}$ J m^{12} and $B = 5.66 \times 10^{-77}$ J m^{6}. Using these values and those in Questions 1.37 and 1.40, estimate the Lennard-Jones constants for krypton, Kr.

1.42 Derive the relationship

$$V_{LJ} = 4V_{min}[(r(0)/r)^6 - (r(0)/r)^{12}]$$

from $V_{LJ} = Ar^{-12} - Br^{-6}$.

1.43 The melting points and boiling points of the noble gases are given in the table. Explain these trends, and suggest why the melting and boiling points are so close.

Element	Melting point (°C)	Boiling point (°C)
Helium	—	−268.9
Neon	−248.6	−246.1
Argon	−189.4	−185.9
Krypton	−157.4	−157.4
Xenon	−111.8	−108.0
Radon	−71	−61.7

Note

1 The unit of energy *electron volt*, eV, is frequently used for atomic processes. $1\,\mathrm{eV} = 1.602 \times 10^{-19}\,\mathrm{J}$.

2
Microstructures and Phase Relationships

- What is the difference between a crystal and a glass?
- What is a binary phase diagram?
- What is a eutectic point?

2.1 Macrostructure, Microstructure, and Nanostructure

2.1.1 Crystalline Solids

Crystals are solids in which all of the atoms occupy well-defined locations, ordered across the whole of the material. The structure is defined in terms of a *unit cell*, a brick-like shape that contains all of the atomic species in the crystal, and that, when stacked in three-dimensions, simply by translation, without rotation or reflection, reproduces the macroscopic crystal. The simplest crystal structures are that of the metallic elements, which contain just a small number of atoms and have a unit cell side of a few nanometres: tungsten, for example, adopts a body-centred unit cell of side 0.3160 nm (Figure 2.1a) and copper a face-centred cubic unit cell of side of 0.3610 nm (Figure 2.1b). At the other extreme, protein crystals may contain many thousands of atoms. Crystalline solids have sharp melting points and the chemical and physical properties are usually dependent upon the direction that they are measured with respect to the underlying crystal structure. Crystals often show *cleavage* on certain planes, indicating that some sheets of atoms are linked by weaker bonds. A significant number of crystals, especially organic crystals, have several sorts of bonding involved. The atoms forming the molecular building blocks are linked by strong covalent bonds, but these molecules are often linked by the weaker bonds described earlier (Section 1.5). Such materials tend to have low melting points and to be rather frail.

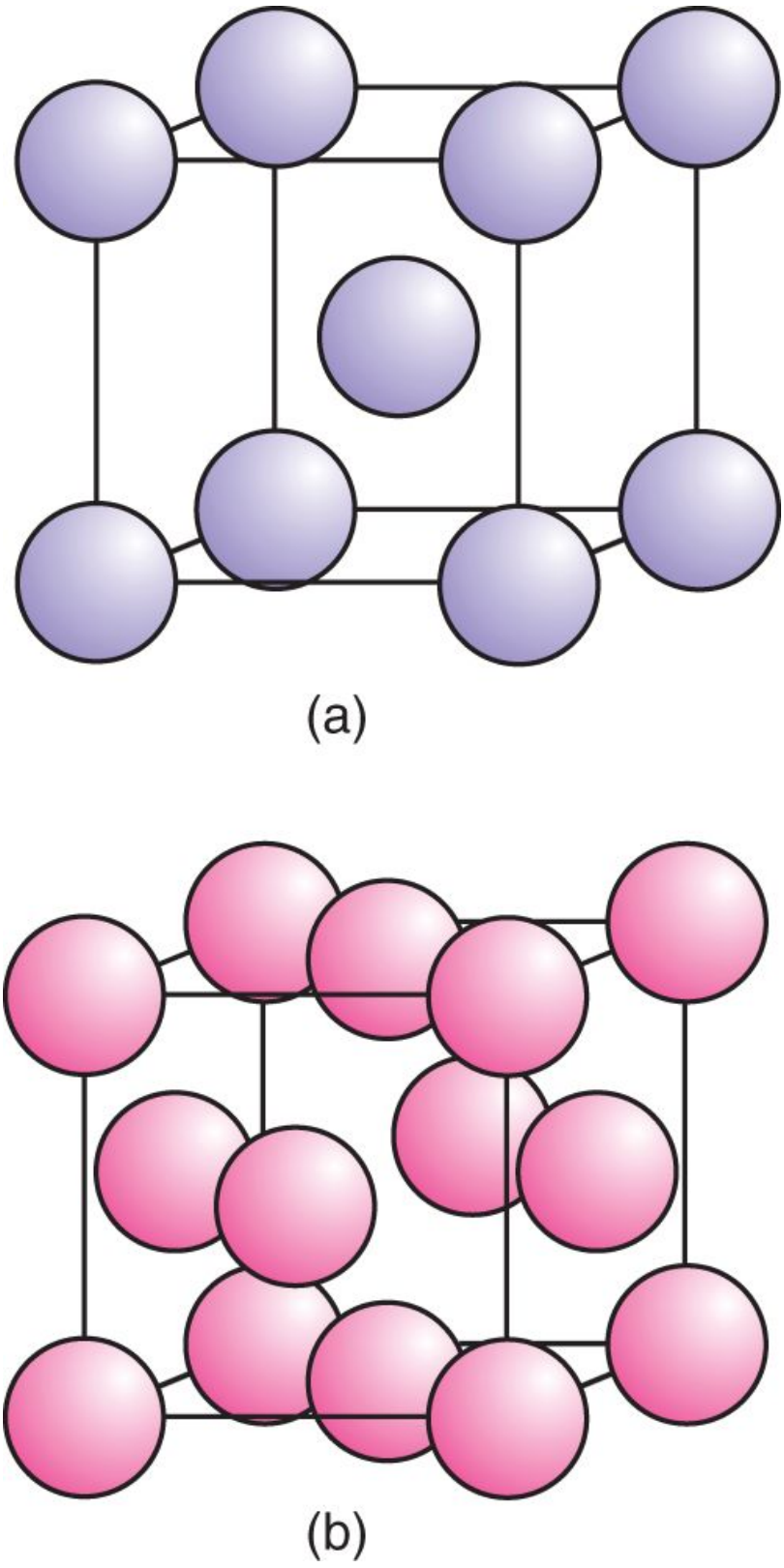


Figure 2.1 The cubic unit cells of (a) tungsten; (b) copper.

Polycrystalline solids are composed of many interlocking small crystals. Most metals and ceramics in their normal states are polycrystalline. The small crystals are often called *grains*, especially in metallurgy. The properties of polycrystalline materials are frequently dominated by the boundaries between the crystallites, *grain boundaries*.

2.1.2 Non-crystalline Solids

Non-crystalline solids do not have long-range order of the atoms in the structure and cannot be described in terms of a unit cell. There is often some degree of short-range order, extending over a few atom radii, but no correlation of atom positions at longer distances. There are four types of non-crystalline solid of most importance, glasses, polymers, amorphous solids, and aerogels.

A *glass* is formed when a high temperature liquid state is *undercooled* or *supercooled* below the solidification point of a potential crystalline solid and then solidifies without crystallising. Glasses are often described as *supercooled liquids*. To some extent, glass can be thought of a product of kinetics, and the structure of a glass can depend upon the rate at which the liquid is cooled. Glasses do not have a sharp melting point. Instead, they continually soften from a state that can be confidently defined as solid to a state that can be defined as a viscous liquid. Glasses containing several components are often found to be inhomogeneous at a scale of about 10^{-6} m. Composition variations arise in the melt, when the various components of the system do not mix completely, rather like, but not as extreme as, oil and water. They can also arise on cooling a homogeneous solid, when some components separate by a process called *spinodal decomposition*.

The best-known glasses are derived from silica, SiO_2. Cooling pure silica melts usually results in a glass made up of randomly oriented SiO_4 tetrahedra (Figure 2.2a,b) joined by corner-sharing without any long-range order (Figure 2.2c) Commercial silicate glasses are manufactured from silicon dioxide mixed with other large metal oxides, typically alkali metal or lead. These are made up of SiO_4 tetrahedra, joined as before, with the large cations present arranged at random, between the networks of SiO_4 tetrahedra (Figure 2.2d). Silicate glasses tend to be transparent because the component atoms do not absorb visible light. Coloured glasses are made by adding light scattering or absorbing impurities.

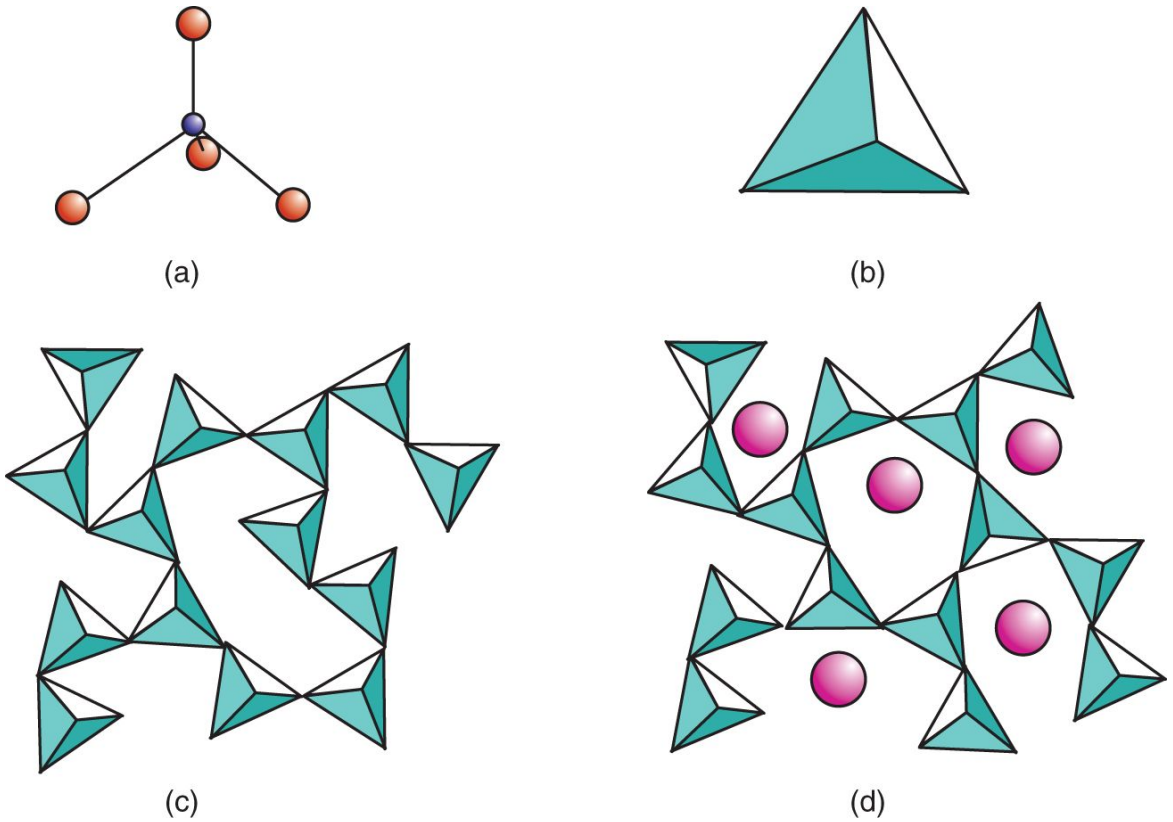


Figure 2.2 Silicate glasses, schematic: (a) structure of an SiO_4 tetrahedron; (b) SiO_4 tetrahedron as a solid; (c) pure silica glass; (d) silicate glass containing large alkali metal cations.

Liquid metals, because of their simple crystal structures, tend to crystallise easily on cooling and the first metallic glasses, with compositions $Au_{75}Si_{25}$ and $Pd_{77.5}Cu_6Si_{16.5}$, were not fabricated until the 1960s, using ultra-rapid cooling of small volumes of molten metal. Metallic glasses are generally alloys of dissimilar elements: the dissimilarity impeding the tendency to order into crystals as the temperature of the liquid drops (Section 4.1.4).

Many organic compounds can solidify as a glass. Boiling and cooling crystalline sugar, an organic molecular compound, will form glasses called 'boiled sweets' or toffee, depending on the other ingredients included. If additives are used to make the melt partly crystallise during cooling, the product is fudge.

Polymers are a class of substances that consist of very large molecules, *macromolecules*, built up from many multiples of small molecules, *monomers*. They can be synthetic (polythene, nylon) or natural (protein, rubber) and occur widely in nature as vital components of living organisms. A sheet of a solid transparent polymer such as methyl

methacrylate (Plexiglass or Perspex), is very difficult to tell from a sheet of window glass by sight alone because the structure of these polymeric solids is non-crystalline, and, as with silicate glasses, made up of molecules that do not interact with incident light.

Most polymers, both natural and synthetic, have a framework of linked carbon atoms. These are strong because the carbon atoms are linked by covalent bonds. Polyethylene (polythene), for example, is a polymer of ethylene, C_2H_4. The polymerisation process involves opening the C□C bond in a single molecule and then joining these fragments into long chains. The long molecules themselves are linked by weak bonds and are usually disordered. Dependent upon the method of production, the polymer molecules can result in mostly very long unbranched molecules (Figure 2.3a) or more branched forms (Figure 2.3b). Polythene molecules are of the order of 10^4 monomer units long and resemble thin strings. In these molecules, each carbon is tetrahedrally coordinated to two hydrogen atoms and two carbon atoms to form long chains (Figure 2.3c). Chains that are largely unbranched tend to pack together in parallel arrangements, resulting in a higher density material (Figure 2.3d), while the branched forms do not pack together as well and produce lower density material.

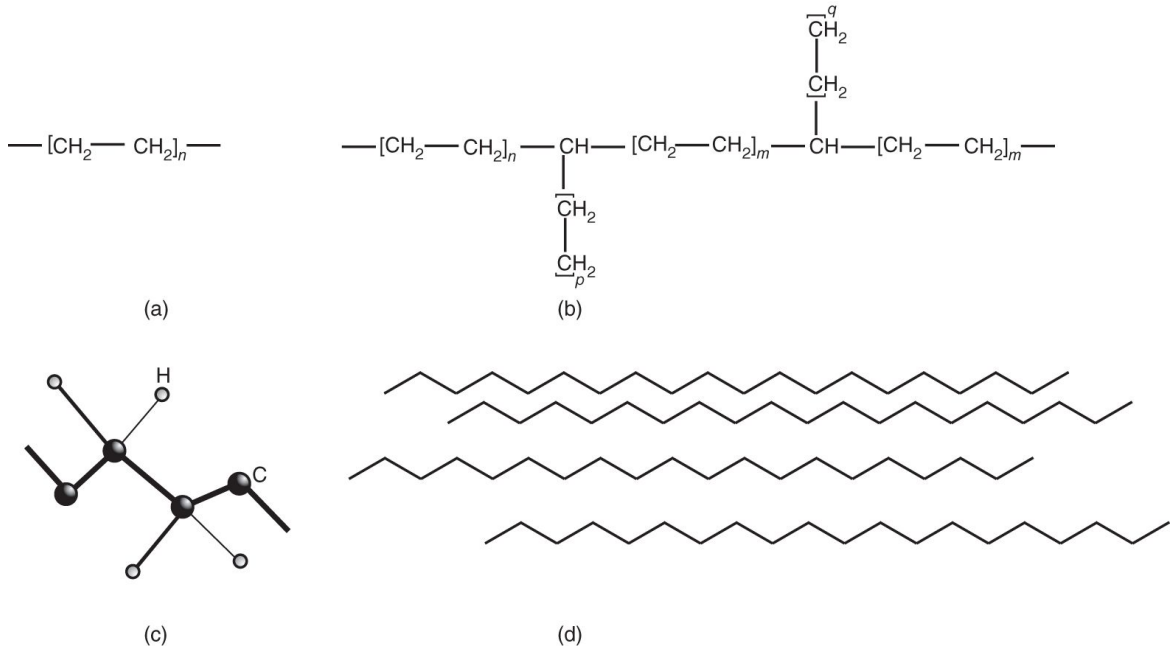


Figure 2.3 Polythene: chemical formula representation of (a) linear polythene molecules; (b) branched polythene molecules; (c) tetrahedrally coordinated C atoms in a polythene chain, schematic; (d) linear chains of polythene molecules in a solid, schematic.

Solids evaporated and then condensed rapidly onto cool surfaces usually do not crystallise and are said to be in an *amorphous* state. Amorphous coatings of this type are widely used in the electronics and optics industries. Such compounds will generally transform into a crystalline state if sufficient energy is supplied to allow crystallisation to occur.

Aerogels are ultra-low-density solids that have a microstructure of highly porous foam. The interconnected pores have a size of less than 100 nm and the structure can be described as of a *fractal* nature, with the smallest characteristic dimension being of the order of 10 nm. Aerogels have been made from many materials, but silica aerogels are the best known. These have extremely low densities, with porosities of over 99.9%. Because of this, the physical properties of aerogels are quite different to that of the parent phase. The thermal conductivity of a silica aerogel is 10^{-2}–10^{-3} that of ordinary silica glass, its refractive index varies from 1.002 to 1.3, compared with 1.5 for silica glass, and the speed of sound drops to 100–300 m s^{-1} compared with 5000 m s^{-1} in silica glass, with all values dependent upon the porosity. These materials find uses ranging from thermal insulation to high-energy nuclear particle detectors.

2.1.3 Partly Crystalline Solids

Although most solids are crystalline, there are important groups that are partly crystalline and partly disordered. For example, glasses are not stable thermodynamically. Given enough time a glass will crystallise. The process of glass crystallisation is called *devitrification*. Opal glass is a silicate glass prepared so that it has partly recrystallised to give a glassy matrix containing a dispersion of small crystallites. These crystallites reflect light from their surfaces to create the opacity of the solid. Glass ceramics are deliberately recrystallised during processing to give a material with the formability of glass and the enhanced mechanical properties of a polycrystalline ceramic. Porcelain is a material consisting of a glassy matrix in which small crystals of other oxides are embedded.

Polymers also show a natural tendency to crystallise, which is thwarted to a greater of lesser extent by the structure of the polymer molecules. For example, polythene often shows a partly crystalline structure. If the liquid is cooled reasonably quickly, the chains remain in an extended form. The material has a low density and a low refractive index and is very flexible. It resembles a glass. However, if the polythene is cooled slowly from the melt, some chains can fold up into crystalline regions 10–20 nm thick. These crystalline regions are of higher density and refractive index

compared with the non-crystalline parts. Most polythene is a mixture of crystalline and amorphous regions, which is why it appears milky.

2.1.4 Nanoparticles and Nanostructures

Nanoparticles and nanostructures are, in a general sense, solids that show chemical and physical properties that are observably different from the normal properties of bulk solids, occasioned by small size. Despite the large research effort, the definition of a nanostructure remains slightly imprecise, and is equated with a dimension, in the solid state, of between 1 and 100 nm. The unique properties of these materials stem from changes in the electronic energy levels of the material as the size diminishes. The energy levels of isolated atoms are sharp, while in a solid these coalesce into an energy band. As a solid is imagined to fragment into smaller and smaller units, the energy bands must eventually revert to more atom-like sharp levels. At the nanoscale, this transformation has partly occurred, giving rise to unique properties that differ significantly from those of the bulk solid. These properties have made research and application of nanostructures of considerable importance for many fields of endeavour, including physics, especially optics and electronics, as well as chemistry, biology, and medicine.

Nanosheets, which are isolated atomically thin sheets, have one small dimension. Graphene, a single sheet of hexagonal rings of carbon atoms (Figure 2.4a), is derived from the carbon compound graphite. Graphene shows several remarkable electronic properties, including, when two sheets are stacked and given a small orientation, misfit, just 1.1°, superconductivity. Other sheet-like materials have been fabricated for device use, including boron nitride, BN (Figure 2.4b), and a number of transition metal chalcogenides, especially MoS_2 and WS_2 (Figure 2.4c).

All of these materials, like graphite, are built up from layers of strongly bonded atoms lying in the plane of the sheet, these being held together by weak van der Waals interactions. Because of this, these solids are frequently called *van der Waals materials*. More complex structures composed of stacked chemically different nanosheets are being explored.

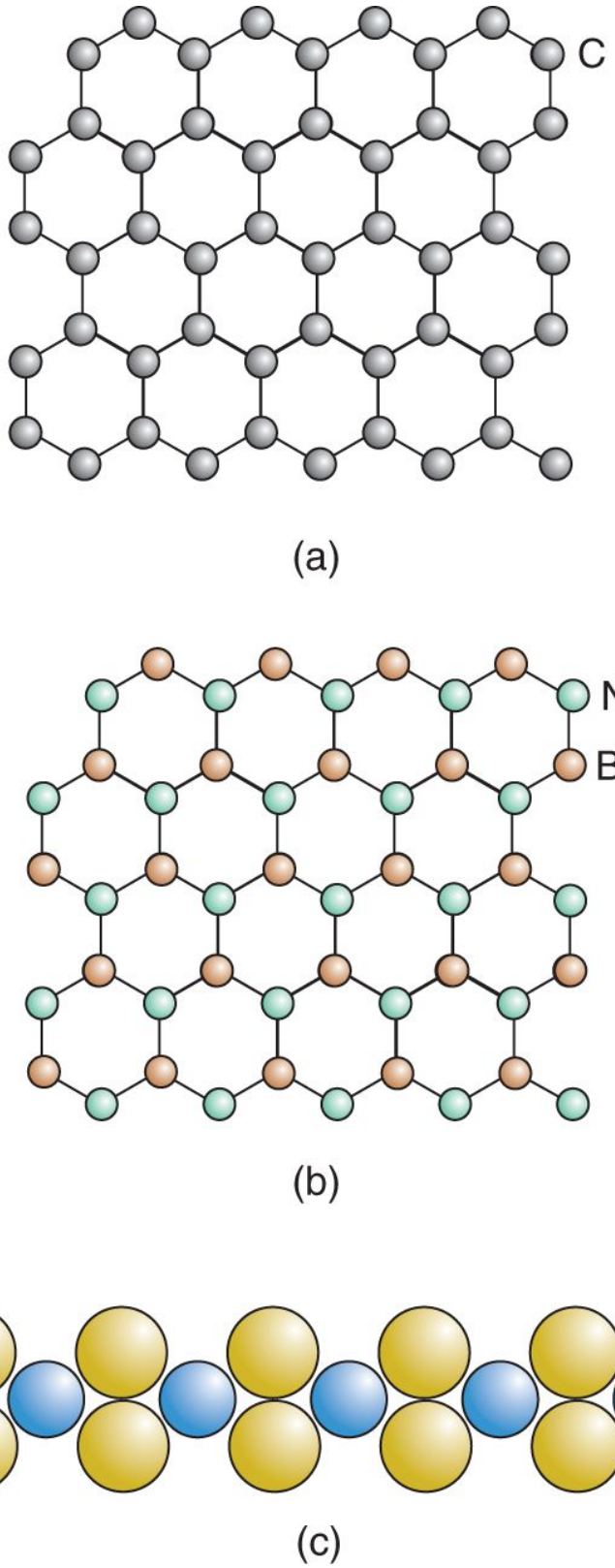
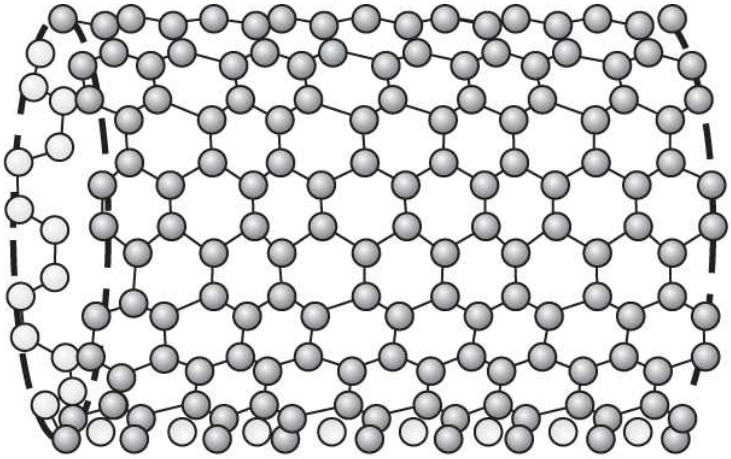
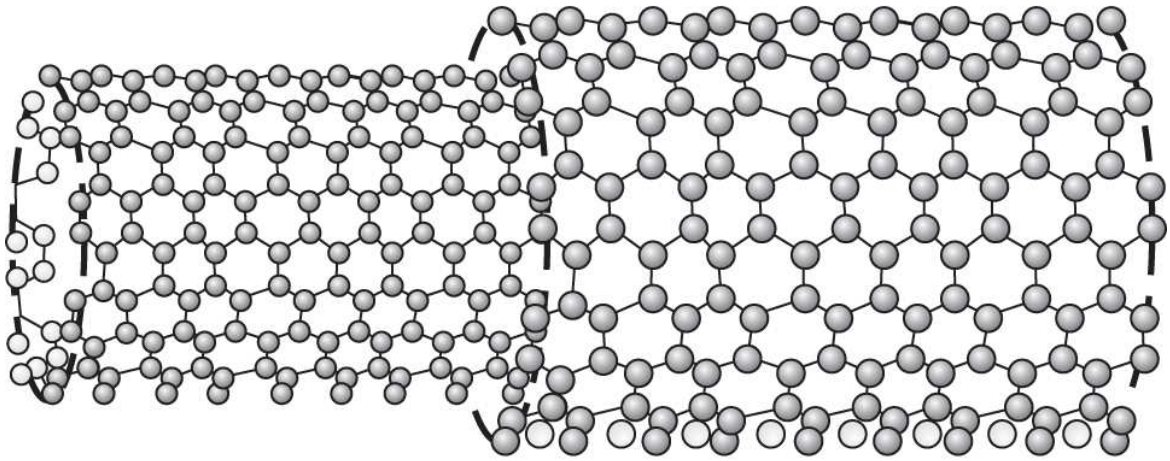


Figure 2.4 Sheet-like nanomaterials: (a) graphene; (b) a single boron nitride layer; (c) a single MoS_2 layer.

Nanotubes, nanowires, and nanorods have two short dimensions. Among the first of these to be investigated were carbon nanotubes, which can be likened to rolled up sheets of graphene (Figure 2.5a). There are a number of ways in which this (imaginary) join can be made, and carbon nanotubes have a number of differing geometries, depending on how this linkage is made. The utility of such nanotubes is being expanded by the inclusion of fluorescent molecules within the tube or attached to the outer surface, making them valuable for a number of sensor applications, particularly in biomedicine. In addition, multiwalled tubes have been fabricated that consist of concentric C, BN, and MoS_2 tubes, which also find use as sensors (Figure 2.5b).



(a)



(b)

Figure 2.5 (a) C nanotube; (b) double tube arrangement, schematic.

Nanoparticles have three short dimensions. As before, the first of these to be described were carbon structures, in particular the *fullerenes*, which are roughly spherical assemblies of carbon atoms linked by strong covalent bonds. The first example to be characterised, C_{60}, was called *buckminsterfullerene* as the structure (Figure 2.6a), resembled the geodesic dome structure developed by Buckminster Fuller. Later these structures were also called *buckyballs* or *fullerenes*. The structure of C_{60} is a truncated icosahedron and is built of faces made up of pentagons and hexagons. A carbon atom is found at each vertex of the structure. More recently large numbers of nanoparticles have been fabricated from insulating materials such as oxides, chalcogens, and metals (Figure 2.6b,c). The properties of these particles are extremely varied and are described in Sections 8.9, 9.3.9, 10.6, 11.5.6, and 12.9.

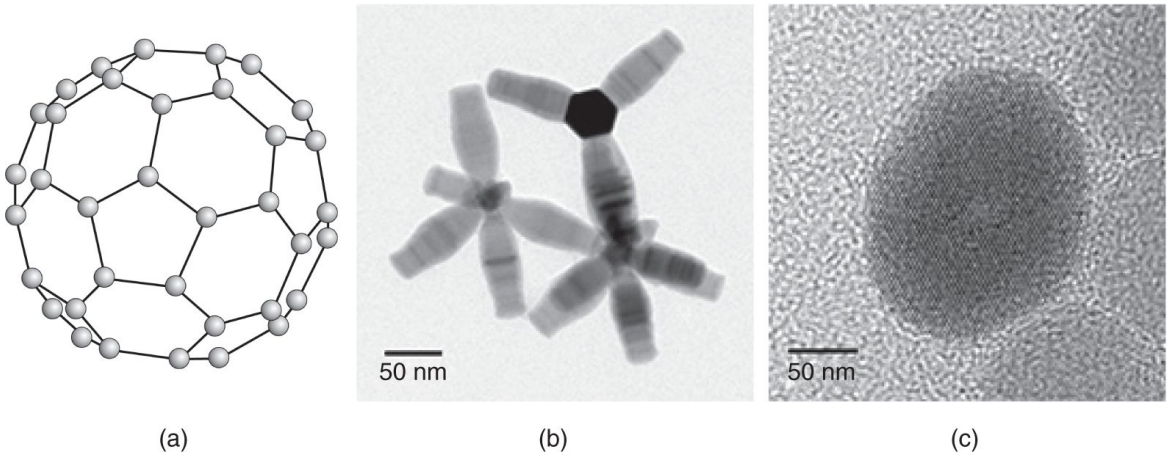


Figure 2.6 Nanoparticles: (a) C_{60} buckyball; (b) Au core nanoparticle with Ni branches; (c) $ZnFe_2O_4$ nanoparticle.

Source: (b) and (c) Courtesy of Cheong et al. (n.d.).

In many cases nanoscale properties are retained if the nanostructure is an integral part of a larger solid. For example, a thin layer of a semiconductor sandwiched between layers of a different semiconductor is called a *quantum well* (Figure 2.7a). In this structure, the electrons are essentially confined to the plane of the layers and are regarded as *two-dimensional* from the point of view of behaviour. Similar devices built up from several alternating thin layers of materials are called *multiple quantum well structures*, or *superlattices* (Figure 2.7b). A thin strip of semiconducting material is called a *quantum wire* (Figure 2.7c) and a small nanoscale volume of crystal is called a *quantum dot* (Figure 2.7d).

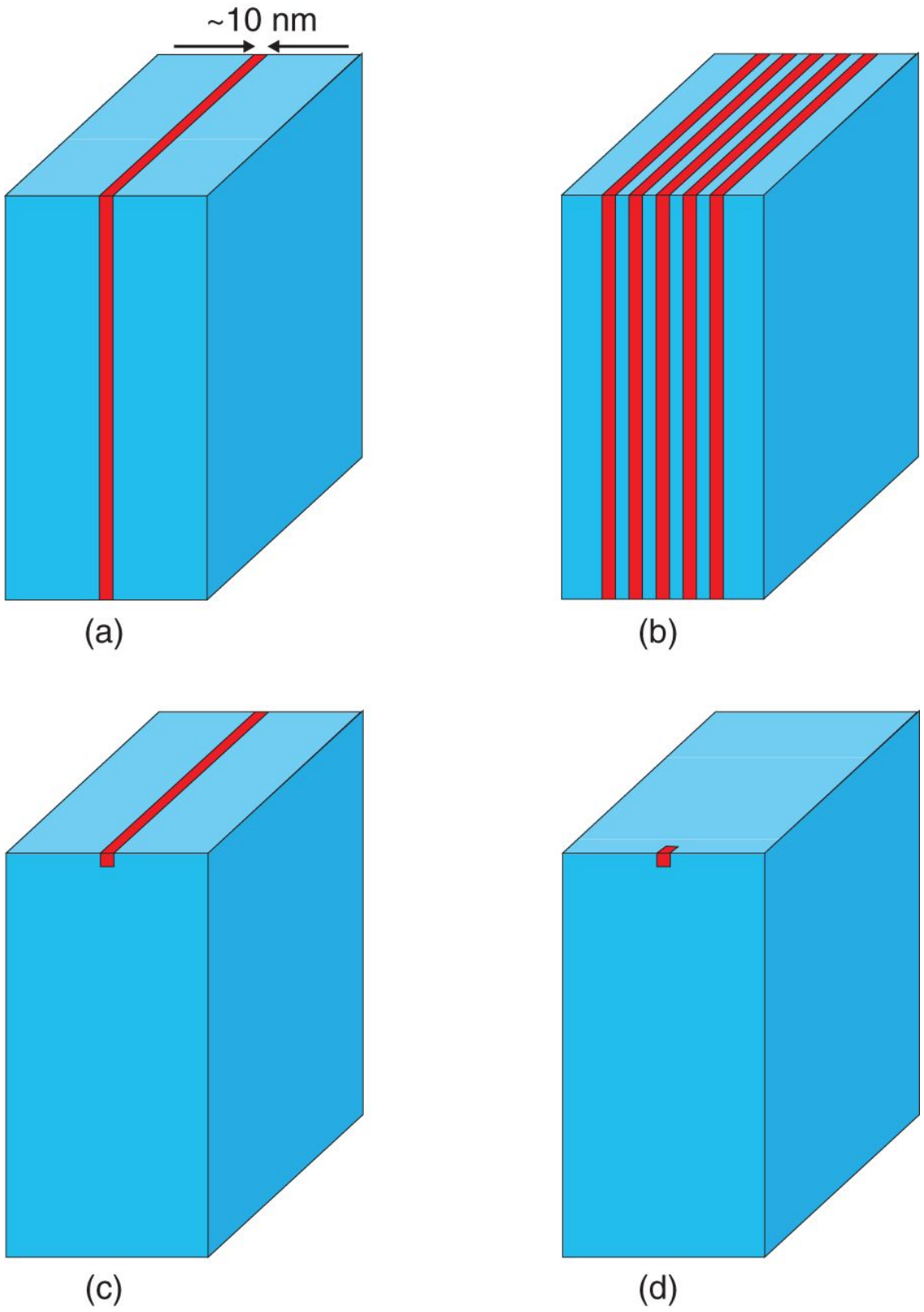

~10 nm
(a)
(b)
(c)
(d)

Figure 2.7 Electronic semiconductor nanostructures: (a) a quantum well; (b) a multiple quantum well superlattice; (c) a quantum wire; (d) a quantum dot.

2.2 The Development of Microstructures

The development of the correct microstructure is of prime importance in many manufacturing processes, and an increasing mastery of this ability has marked out the progression of both ancient and modern civilisations. In early times, control was achieved by trial and error. The resulting recipes were then closely guarded by tradesmen or trade guilds. The control of microstructures in modern times has come to depend upon a precise knowledge of the science behind the chemical and physical changes that are taking place. This is typified by the rise of metallurgy concurrent with the development of the modern steel industry, some 100 or so years ago.

2.2.1 Solidification

Many solids, especially metals, are produced from liquid precursors, and control of solidification is important in the development of the appropriate microstructure. There are two important steps involved in solidification. *Nucleation* refers to the initial formation of tiny crystallites. This occurs especially at mould edges and dust particles, which act as *sites* for nucleation. If only one nucleus forms a single crystal is produced, whereas if many nuclei form, a polycrystalline solid results. In many systems, cooling produces nuclei of more than one phase and impurities will add to the complexity. Nucleating agents can be added deliberately if nucleation throughout the volume of the melt is desired. On the other hand, the formation of nuclei must be suppressed during glass formation.

Crystal growth follows nucleation and contributes greatly to the development of microstructure. The resulting solid may contain crystals of different compounds, as in the rock granite, which is mainly composed of interlocking crystallites of mica, quartz, and feldspars (Figure 2.8a). Pure metals and alloys are also normally polycrystalline (Figure 2.8b). Many crystals grow from the melt with a branching shape or morphology that resembles a tree in form, with each branch consisting of another single crystal. These are called *dendrites* and the growth is called *dendritic growth* (Figure 2.8c). The shape of the dendritic crystal reflects the internal symmetry of the crystal structure. Cubic metals usually have side arms perpendicular to the long growth axis, while in hexagonal crystals the side arms are at angles of 60°. It is this symmetry dependence that gives snowflakes and frost, which are dendritic ice crystals, their definitive form.

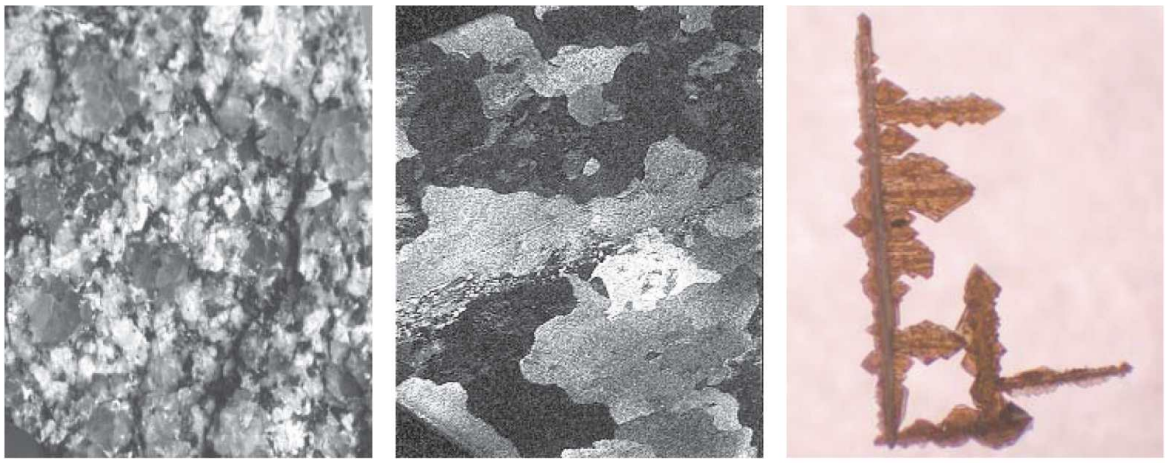


Figure 2.8 Polycrystalline solids: (a) granite; mica (black), quartz and feldspars (white/grey), grain size ~2 mm; (b) aluminium, grain size ~1 cm; (c) dendritic rutile (TiO_2), branch crystals ~5 mm.

The microstructure of a solid will also depend upon how quickly different crystal faces develop. This controls the overall shape of the crystallites, which may be needle-like, blocky, or dendritic. The shapes will also be subject to the constraint of other nearby crystals so that the product will be a solid consisting of a set of interlocking grains. The rate of cooling will also affect the size distribution of the crystallites and rapid chilling of liquid in contact with the cold outer wall of a mould can lead to amorphous or poorly crystallised products. Liquid within the centre of the sample may crystallise slowly and produce large crystals.

2.2.2 Processing

Processing refers to the treatment of a solid to alter the microstructure and external form. It is a subject of considerable importance in industry.

Working and *heat treatment* are techniques mainly applied to metals. When a metal is hammered, rolled, or deformed, it is referred to as working. Cold metals get harder on working as the process introduces large numbers of defects and strain energy into the sample. If the metals are heated to about half of their melting point (called *annealing*), they can partly recrystallise and release the strain energy. This causes the metal to become softer and more *ductile*.

Thermoplastic polymers can easily be melted and moulded into flexible shapes. The rigidity and strength of the product can be improved by *cross-linking* between the polymer chains. One of the first deliberate cross-

linking processes is the vulcanisation of rubber, which was originally used in car tyre manufacture. The process transforms sticky soft rubber into a hard, flexible material.

The *devitrification* of glass to produce glass ceramics is typical of processing in the glass and ceramic industries. Here the processing aim is to overcome the brittleness typical of glasses while retaining good chemical inertness and easy formability.

Sintering is widely used to make polycrystalline ceramic bodies. A powder is compressed and heated at a temperature below the melting point to produce a strong polycrystalline solid. Sintering comes about by solid state diffusion and may be helped by the presence of traces of liquid. Many electrical and electronic components are produced by sintering. Some metal parts are also made via this method, and the subject area is called *powder metallurgy*. The main aim of sintering is to produce a high-density solid with little porosity (also see Section 6.1).

Dehydration, or more exactly, fluid phase removal, is used to form the ultra-porous microstructure of aerogels. Normally, when a gel (for example, ordinary gelatine) is dehydrated, the material shrinks and collapses. As fluid in the pores evaporates, a meniscus forms, which generates large surface tension forces. These cause the pore structure to disintegrate. The formation of aerogels is a typical processing problem; how is it possible to remove fluid while maintaining the porous microstructure. In original work, high pressures and temperatures were used to take the fluid in the material above its critical temperature (see Section 1.2.1). In this state, the fluid does not exert surface tension, and it is possible to remove it without collapse of the solid framework.

2.3 Phase Diagrams

A *phase* is a part of a system that is chemically uniform and has a boundary around it. Phases can be solids, liquids, or gases, and on passing from one phase to another, it is necessary to cross a *phase boundary*. Liquid water, water vapour, and ice are the three phases found in a water system. In a mixture of water and ice it is necessary to pass a boundary on going from one phase, say ice, to the other, water.

Phase diagrams are diagrammatic representations of the phases present in a system consisting of a number of *components* (i.e. chemical species) under specified conditions of composition, temperature, and pressure. Phase diagrams mostly relate to *equilibrium conditions*. If a diagram represents non-equilibrium conditions, it is called an *existence diagram*. The conditions limiting the existence and coexistence of phases are given by the (*Gibbs*) *Phase Rule*, described in the following text.

2.3.1 One-Component (Unary) Systems

A *one-component*, or *unary*, system is made up of a single chemical species, for example, Fe, H_2O, or C_2H_5OH. The phases that exist in a one-component system are limited to vapour, liquid, and solid. Phase diagrams for one-component systems are specified in terms of two variables, *temperature*, normally specified as °C, plotted along the abscissa (*x*-axis), and *pressure*, specified in atmospheres, plotted along the ordinate (*y*-axis) (Figure 2.9).

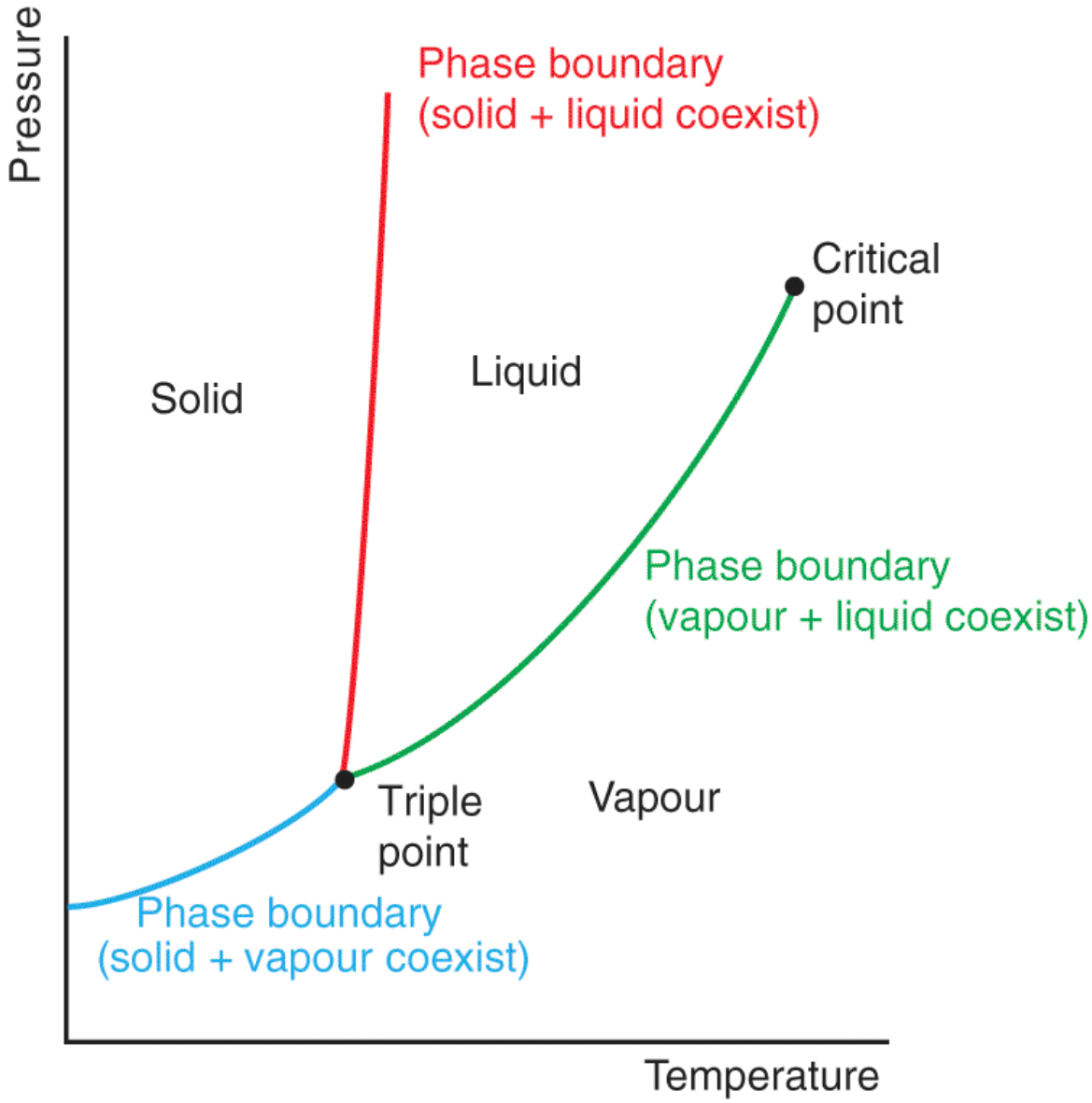


Figure 2.9 The generalised form of a one-component phase diagram.

The areas on the diagram within which a single phase exists are labelled with the name of the phase present. The areas over which single phases occur are bounded by lines called phase boundaries. In a single-phase region, *both* the pressure and the temperature can be changed independently of one another without changing the phase present. On a phase boundary, *two phases coexist*. If the phase boundary between liquid and vapour in a one-component system is followed to higher temperature and pressures, ultimately, it ends. At this point, called the *critical point*, located at the *critical temperature* and the *critical pressure*, liquid and vapour *cannot be distinguished*. A gas can be converted to a liquid by applying pressure only if it is below the critical temperature. At one

unique point, the *triple point*, three phases coexist at equilibrium. If there is any change at all in either the temperature or the pressure, three phases will no longer be present. The triple point is an example of an *invariant point*.

The characteristics of one-component systems will be illustrated with respect to the phase diagram for water (Figure 2.10). The three phases found are ice (solid), water (liquid), and vapour. On the phase boundaries, two phases coexist indefinitely, ice and water, water and vapour, or ice and vapour. (The critical point of water, at 374 °C and 218 atm, is not shown.) The triple point, where ice, water, and vapour coexist, is found at 0.01 °C and 0.006 05 atm.

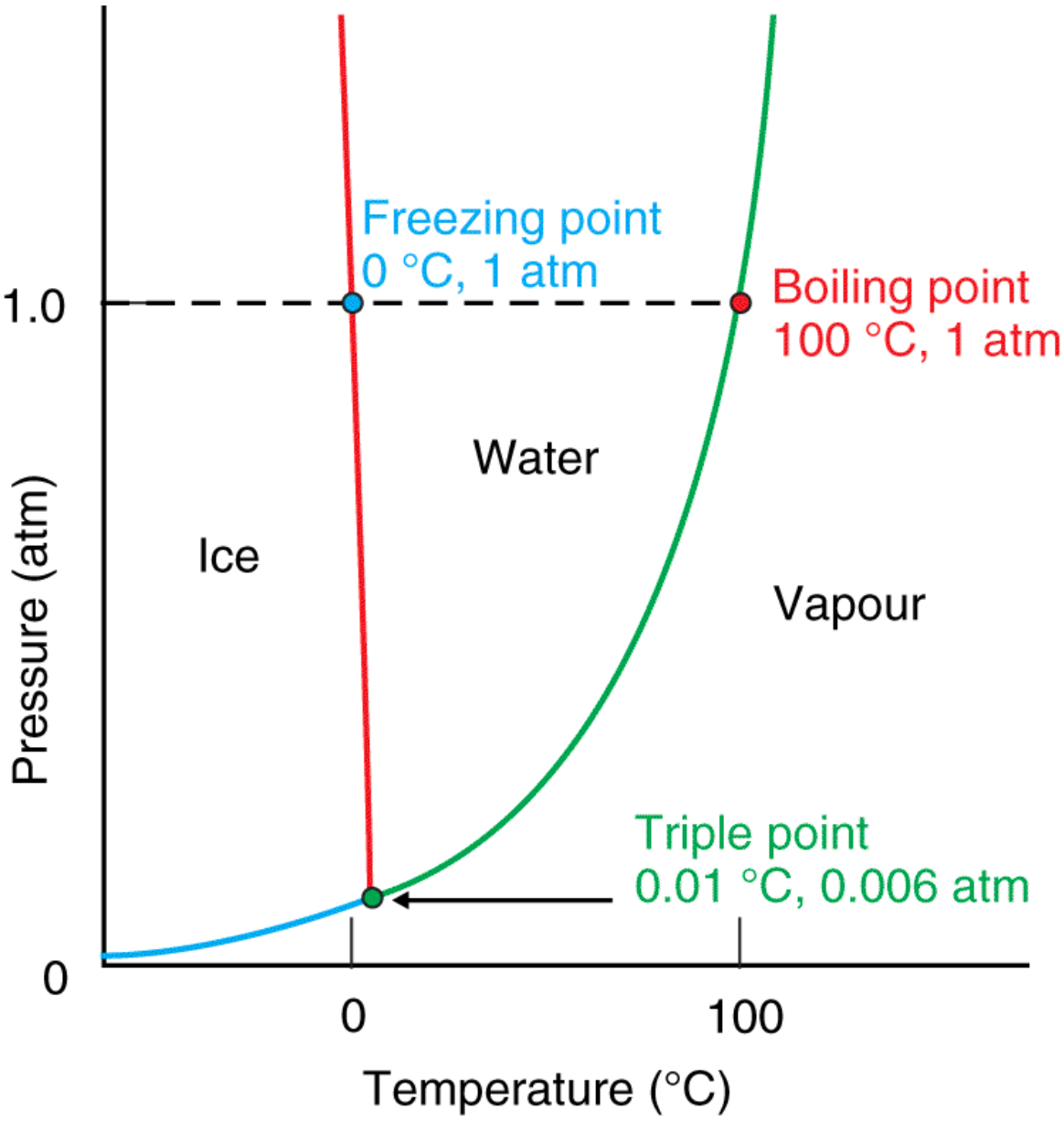


Figure 2.10 The phase diagram for water, schematic.

The slopes of the phase boundaries give some information about the change of boiling and freezing points as the pressure varies. The phase boundary between water and water vapour slopes upwards to the right. This indicates that an increase in pressure will favour liquid compared with vapour, and that the boiling point of water increases with increasing pressure. The ice–water phase boundary slopes upwards towards the left. This indicates that an increase in pressure will favour the liquid over the solid. An increase in pressure will cause the water to freeze at a lower temperature, or ice to melt. This is one reason in supposing that liquid water might occur as a lubricating layer under ice skates.

A phase diagram can be used to explain the pattern of temperature changes observed as a substance cools. For example, a sample of water at a temperature T_s will cool steadily until the water–ice phase boundary, at the melting point T_m, is reached. Between T_s and T_m the temperature vs. time plot, the *cooling curve*, will change smoothly (Figure 2.11a). Further cooling will cause ice to form and two phases will be present. The temperature will now remain constant and more and more ice will form until all of the water has become ice. Thereafter, the ice will then cool steadily and a smooth cooling curve will again be obtained between T_m and the final temperature T_f. An abrupt change in slope of a cooling curve is an indication that the system is passing across a phase boundary.

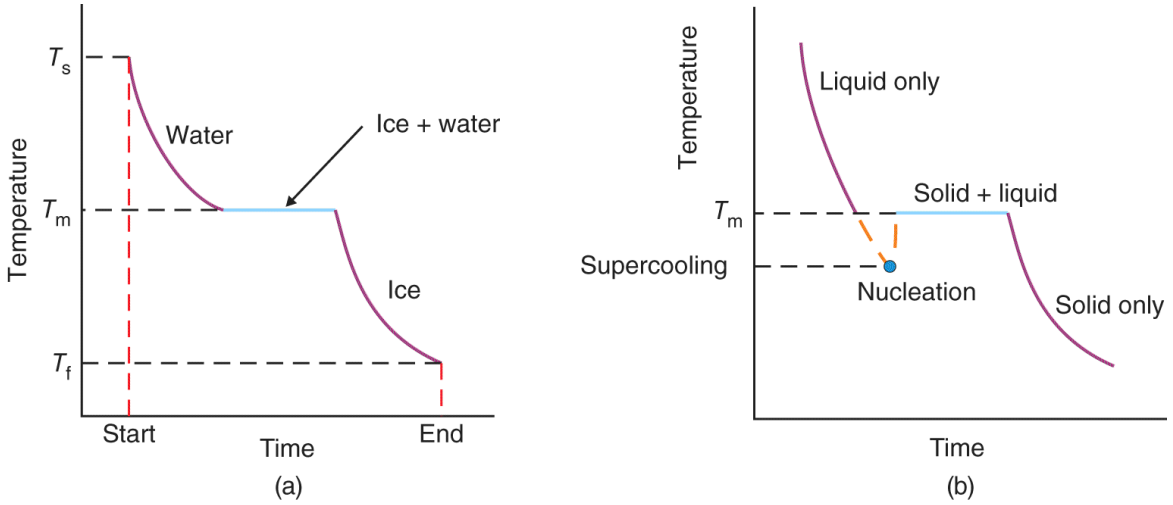


Figure 2.11 (a) The cooling curve generated as a sample of water is slowly cooled from a temperature T_s to a temperature T_f; (b) a cooling curve showing supercooling.

It frequently happens that normal rates of cooling are fast, and experimental cooling curves for liquids often continue below the break expected at a phase boundary (Figure 2.11b). This overshoot is called *supercooling* or *undercooling*. Supercooling reflects the fact that energy is needed to cause a microscopic crystal nucleus to form. In a very clean system, in which dust and other nucleating agents are absent, supercooling can be appreciable. Glasses form in systems in which nucleation is difficult or prevented so that the liquid solidifies before crystallisation has occurred.

The number of phases that can coexist at equilibrium, P, is specified by the Phase Rule:

$$P + F = C + 2 \tag{2.1}$$

where C is the number of *components* needed to form the system and F is the number of *degrees of freedom* or *variance*. The variance specifies the number of variables in the system, such as composition, temperature, and pressure, that can be altered independently without changing the state of the system.

In a one-component system the phase rule becomes:

$$P + F = 3 \tag{2.2}$$

The conditions limiting the existence and coexistence of phases can now be assessed. With reference to the phase diagram for water (Figure 2.10), when only one phase is present, for example, ice, Eq. (2.2) reduces to $F = 2$. That is, the region over which ice is stable has to be specified in terms of two variables, temperature and pressure, or, alternatively, ice can exist over a range of temperatures and pressures. At the triple point, water, ice, and vapour occur together and $P = 3$, so from Eq. (2.2) $F = 0$. There are no degrees of freedom available. That is to say, it is not possible to change either the temperature or the pressure and still keep three phases present. At this point, the system is said to be *invariant*. The phase diagram shows that a change in either temperature or pressure is most likely to lead to the formation of a single phase, and the system will change to all solid, all liquid, or all vapour. Along the phase boundary separating ice and water, only two phases are present and $P = 2$, so from Eq. (2.2) $F = 1$. This phase boundary is characterised by one degree of freedom. This means that one variable, either pressure or temperature, can be changed at will but to maintain two phases in equilibrium the other variable must also change appropriately. A change in one variable automatically determines the value of the other if two phases are to remain present.

The pattern of temperature changes observed as a substance cools follows directly from the Phase Rule. When two phases are present, there is only one degree of freedom available, and the temperature can vary only if the pressure also varies. If the pressure is constant at 1 atm, the temperature cannot change until the phase change is completed. A mixture of ice and water will therefore have a constant temperature that will not change until either the ice has melted or the water freezes.

2.3.2 Two-Component (Binary) Systems

Binary systems contain two components, for example, Fe + C, $NaNbO_3$ + $LiNbO_3$, Pb + Sn. The added component means that three variables are needed to display a phase diagram. The variables are usually chosen as *temperature*, *pressure*, and *composition*. A binary phase diagram thus needs to be plotted as a three-axis figure and a single phase will be represented by a *volume* on the diagram. Phase boundaries form *two-dimensional surfaces* in the representation, and three phases will co-exist along a *line*. However, as most studies are carried out at atmospheric pressure, planar diagrams, using temperature and composition as variables, called *isobaric* phase diagrams, are sufficient. A point in such a binary phase diagram defines the temperature and composition of the system. In metallurgical phase diagrams, compositions are usually expressed as *weight percentages*. In chemical work, *atom percentages* or *mole percentages* are used. Note that the composition gives the amount of the components present, not the chemical composition of any phases present. A single phase occurs over an area in the figure, phase boundaries are drawn as lines, and three phases occur at a point.

The phase rule for a binary system is given by Eq. (2.3):

$$P + F = 4 \tag{2.3}$$

When only one phase is present, $P = 1$, so from Eq. (2.3) $F = 3$. A single phase will thus be represented by a volume in the diagram. On a phase boundary, two phases are present ($P = 2$), and from Eq. (2.3) $F = 2$, so that phase boundaries form two-dimensional surfaces. A two-component system containing two phases is called a *bivariant* system. In a bivariant system, it is possible to change any two of the three variables temperature, pressure, and composition independently, but the third will be fixed by the values selected for the other two. The maximum number of phases that can co-exist is three (for example, solid, liquid, and vapour). In this case, $P = 3$ and Eq. (2.3) gives $F = 1$ and there is only one degree of freedom available to the system. Three phases will co-exist along a line in the phase diagram.

2.3.2.1 Simple Binary Diagrams: Nickel–Copper as an Example

The simplest form of a binary phase diagram is exhibited when components have very similar chemical and physical properties; the nickel–copper system (Figure 2.12) being typical. At the highest temperatures, one homogeneous liquid phase occurs. In this, the Cu and Ni atoms are distributed random. At the lowest temperatures, a homogeneous solid phase, called the *α phase*, is found. Just as in the

liquid, the Cu and Ni atoms are distributed at random, and by analogy, such a material is called a *solid solution*. Because the solid solution exists from pure copper to pure nickel, it is called a *complete* solid solution. Between the liquid and solid phases, phase boundaries delineate a lens-shaped region. Within this area solid α and liquid coexist. The lower phase boundary, between the solid and the liquid plus solid region is called the *solidus*. The upper phase boundary, between the liquid plus solid region and the liquid only region is called the *liquidus*.

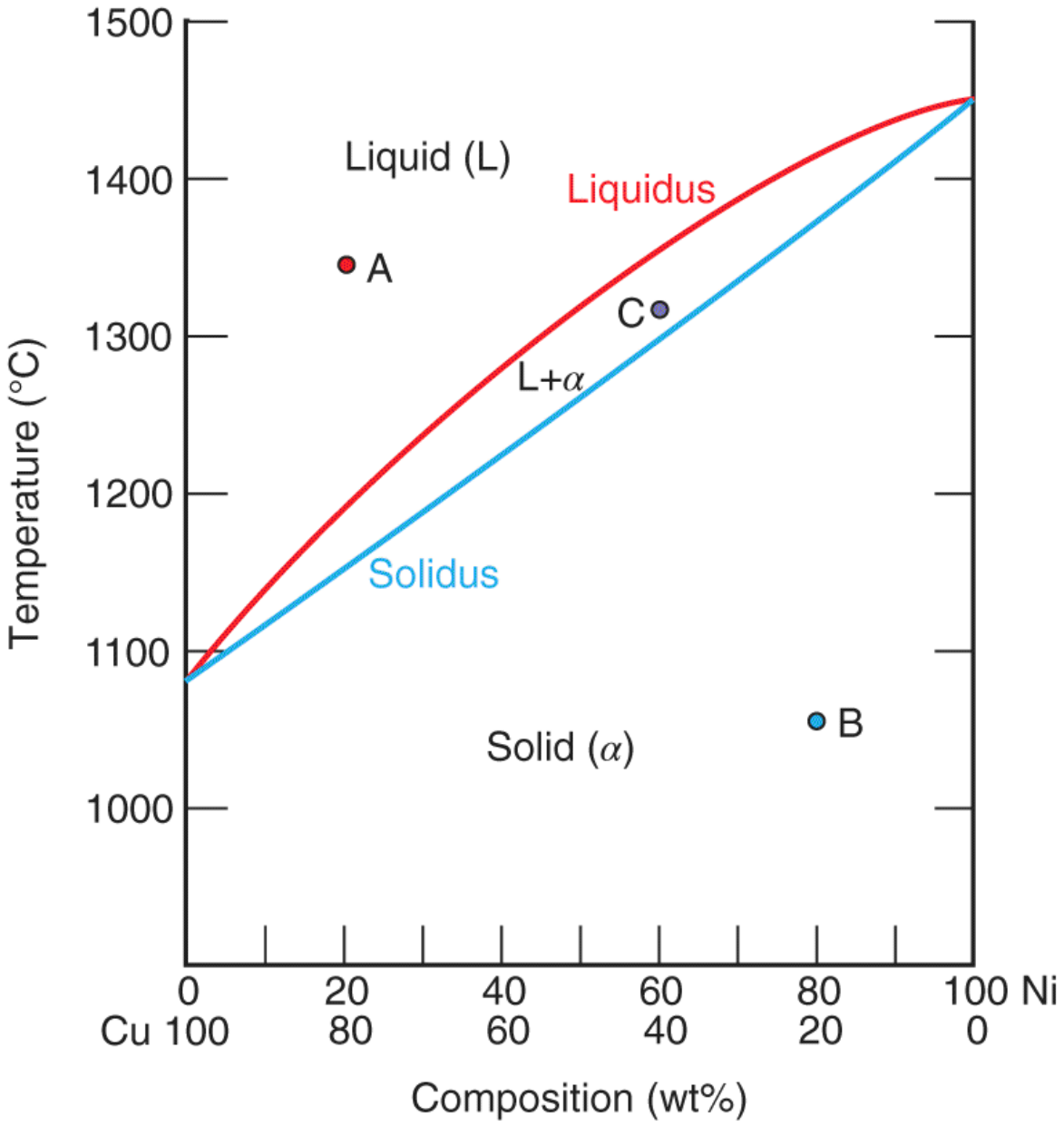


Figure 2.12 The copper (Cu)–nickel (Ni) phase diagram at 1 atm pressure.

The composition of any point in the diagram is simply read from the composition axis. Thus, point A in Figure 2.12 has a composition of 80 wt% copper, 20 wt% nickel. Point B has a composition of 20 wt% copper, 80 wt% nickel. Point C has an *average* composition of 40 wt% copper, 60 wt% nickel, but here two phases present, solid and liquid. To determine the compositions of each of these phases, draw a line parallel to the composition axis, called a *tie line*. The composition of the solid phase is read from the diagram as the composition where the tie line intersects the solidus. The composition of the liquid is read from the diagram as the composition where the tie line intersects the liquidus (Figure 2.13). The composition of the liquid phase, c_l, is ~51 wt% copper and that of the solid, c_s, is ~33 wt% copper. The *amounts* of each of the phases in a two-phase region can be calculated using the *lever rule*. The fraction of solid phase x_s, is given by:

$$x_s = (c_0 - c_l)/(c_s - c_l)$$

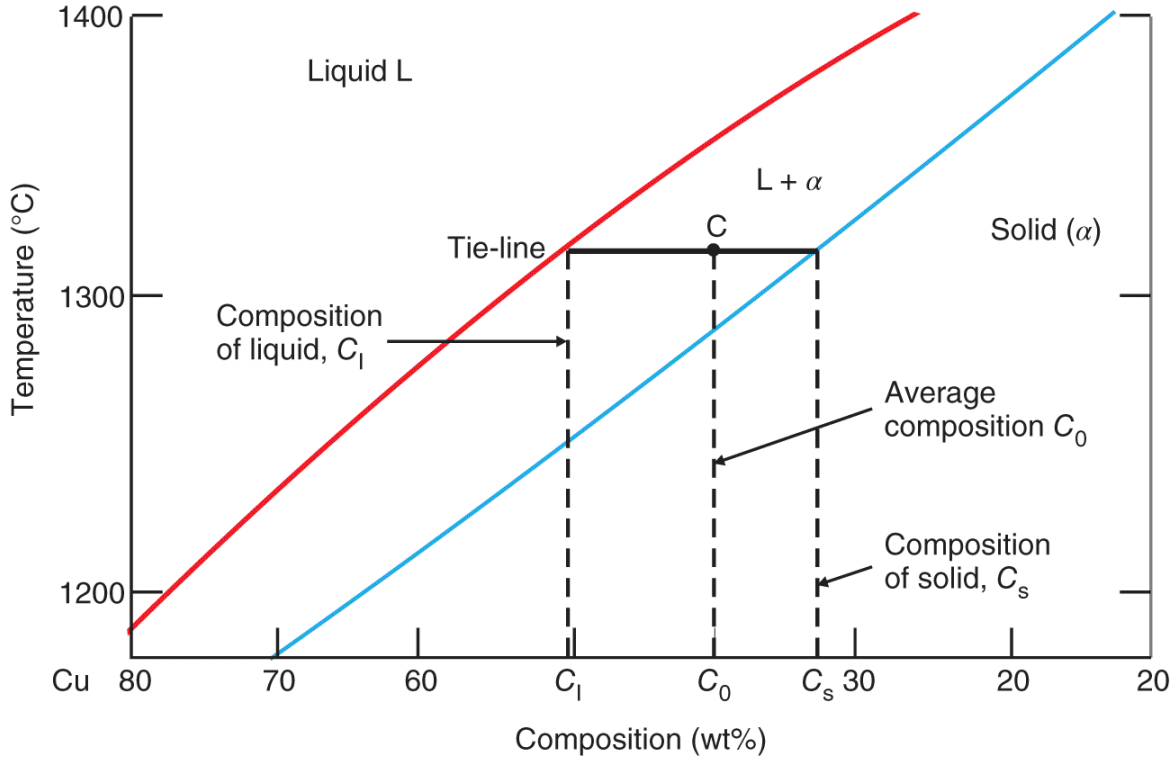


Figure 2.13 Composition determination when two phases are present.

The fraction of liquid phase, x_l, is given by:

$$x_l = (c_s - c_0)/(c_s - c_l)$$

where c_o is the average composition of the sample, c_s the composition of the solid phase present in the two-phase mixture and c_l the composition of the liquid phase present in the two-phase mixture.

The cooling curve of a liquid as it passes through the two-phase region shows an arrest, just as in a one-component system, but the change of slope is not so pronounced.

2.3.2.2 Binary Systems Containing a Eutectic Point: Tin–Lead as an Example

The majority of binary phase diagrams are more complex than the Ni–Cu example. Typical of many is the diagram of the lead–tin system (Figure 2.14). At high temperatures, the liquid phase is a homogeneous mixture of the two atom types, lead and tin. However, the mismatch in the sizes of the lead and tin atoms prevents the formation of a complete homogeneous solid solution in the crystalline state. Instead, *partial solid solutions* occur at each end of the phase range, close in composition to the pure components or *parent phases*. The solid solutions, also referred to as *terminal solid solutions*, are called α, found on the lead-rich side of the diagram, and β, found on the tin-rich side. These solid solutions adopt the crystal structure of the parent phases. Thus, the α-phase has the same crystal structure as lead, and the tin atoms are distributed at random within the crystal. The β-phase has the same crystal structure as that of tin, and the lead atoms are distributed at random within the crystal. The extent of solid solution in the α-phase is much greater than that in the β-phase, as the smaller tin atoms are more readily accommodated in the structure of the large lead atoms than vice versa. The extent of the solid solution increases with temperature for both phases.

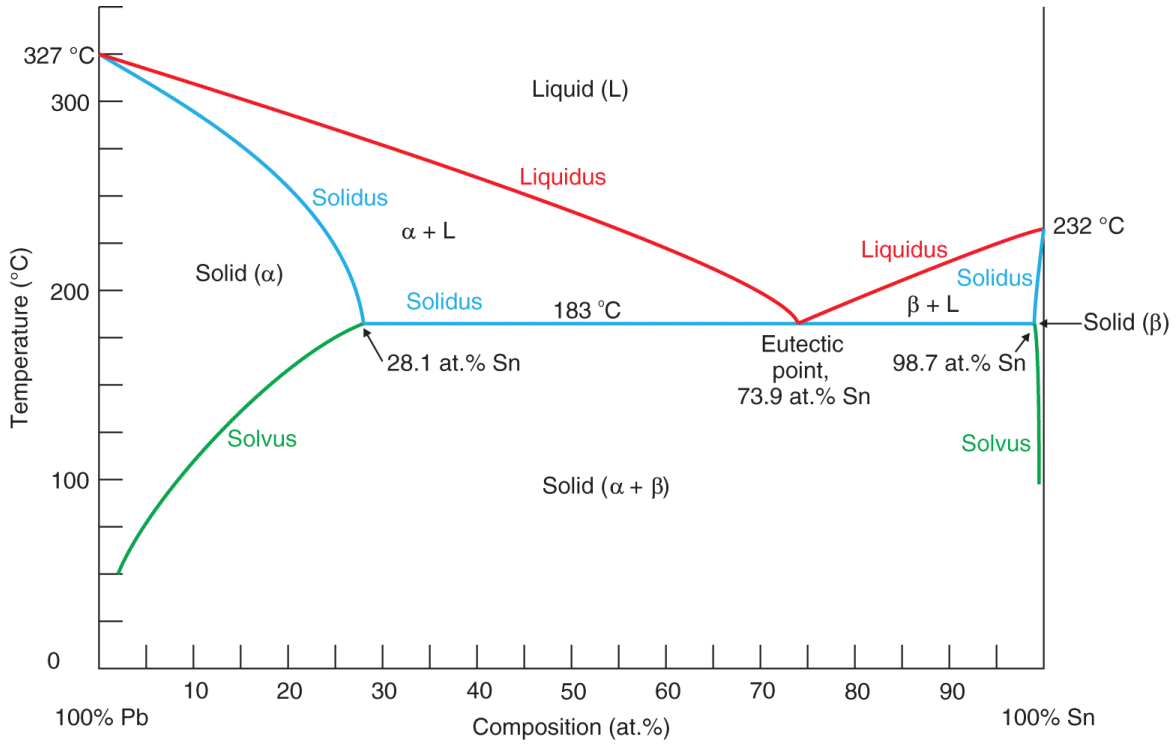


Figure 2.14 The lead (Pb)–tin (Sn) phase diagram at 1 atm pressure.

The overall composition of a crystal in the terminal solid solution regions is simply read from the composition axis, as in the nickel–copper system. The amount of the phase present is always 100%. Thus, point A (Figure 2.15) corresponds to a homogeneous α-phase solid of composition 15 at.% tin, 85 at.% lead at a temperature of 200 °C. Between the solid solutions, below 183 °C, a *two-phase* solid exists composed of α and β, in proportions depending upon the overall composition of the system. The phase boundaries between the solid solutions and the two-phase region are called the *solvus* lines. The *overall* composition of any sample is read from the composition axis. The compositions of the two phases present are given by the compositions at which the tie line intersects the appropriate solvus, drawn at the appropriate temperature. Thus, the overall composition of point B (Figure 2.15) is 40 at.% tin, Sn, 60 at.% lead, Pb. The composition of the α-phase is 18 at.% Sn, 78 at.% Pb, and the composition of the β-phase is 99 at.% Sn and 1 at.% Pb, at 150 °C. The amounts of the two phases are found by application of the lever rule, using these compositions. Thus:

$$\text{Amount of } \alpha\text{-phase} = (99 - 40)/(99 - 18) = 72.8\%$$

$$\text{Amount of } \beta\text{-phase} = (40 - 18)/(99 - 18) = 27.2\%$$

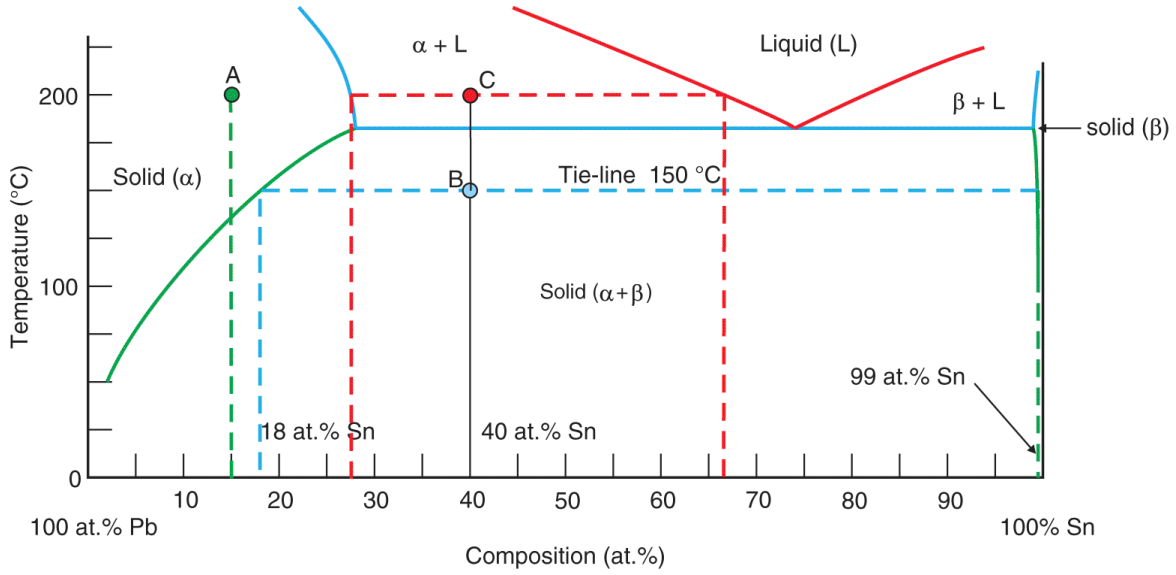


Figure 2.15 Composition determination in the Pb–Sn system.

The liquidus has a characteristic shape, meeting the solidus at the *eutectic point*. The *eutectic composition*, which is the composition at which the eutectic point is found, solidifies at the *lowest temperature* in the system, the *eutectic temperature*. The eutectic point in the lead–tin system is at a composition of 73.9 at.% tin and a temperature of 183 °C.

A eutectic point (in any system) is characterised by the co-existence of three phases, one liquid and two solids. The three phases can only be in equilibrium at one temperature and composition, at a fixed pressure. The eutectic point is therefore analogous to a triple point in a one-component system, and like a triple point, it is also an invariant point and the reaction that occurs on cooling through a eutectic point is called an *invariant reaction*. In the lead–tin system a liquid with the eutectic composition transforms directly into a solid consisting of a mixture of α and β by way of a *eutectic transformation* on cooling:

$$L \rightarrow \alpha + \beta$$

The reaction is reversible. A mixture of α and β with an overall composition equal to the eutectic composition will turn directly to a single liquid phase of the same composition on heating. A cooling curve shows a horizontal break on passing through a eutectic.

Solidification over the rest of the phase diagram involves the transition to a two-phase solid region. A liquid on the Pb-rich side of the eutectic, on passing through the liquidus, will consist of solid α plus liquid. A liquid on the Sn-rich side of the eutectic, on passing through the liquidus, will

consist of solid β together with liquid. The composition of the two solids is obtained by drawing a tie line at the appropriate temperature and reading from the composition axis. The amounts of the solid and liquid phases are obtained by noting the average composition and using the lever rule. For example, point C (Figure 2.15) corresponds to an overall of composition 40 at.% Sn. On slow cooling to 200 °C the sample will consist of liquid of composition 67.5 at.% Sn and solid α with a composition of 27 at.% Sn. The amounts of these two phases can be obtained via the lever rule.

On slowly cooling a sample from a homogeneous liquid through a two-phase region, it is seen that as the temperature falls, the composition of the solid follows the left-hand solidus and the composition of the liquid follows the liquidus. When the eutectic temperature is reached, the remaining liquid will transform to solid with a composition equal to the eutectic composition. At this stage, the solid will contain only solid α and solid β. Further slow cooling will not change this, but the compositions of the solid α and solid β will evolve, as the compositions at a given temperature always correspond to the compositions at the ends of the tie lines. The microstructure of the solid will reflect this history.

2.3.2.3 Intermediate Phases

The components of any phase diagram may react to produce new compounds that, in the context of phase diagrams, are called *intermediate phases*. For example, the phase diagram for the binary system $CaSiO_3$, calcium silicate (wollastonite), and $CaAl_2O_4$, calcium aluminate, contains a single intermediate phase gehlenite, $Ca_2Al_2SiO_7$, above 1100 °C (Figure 2.16):

$$CaSiO_3 + CaAl_2O_4 \rightarrow Ca_2Al_2SiO_7$$

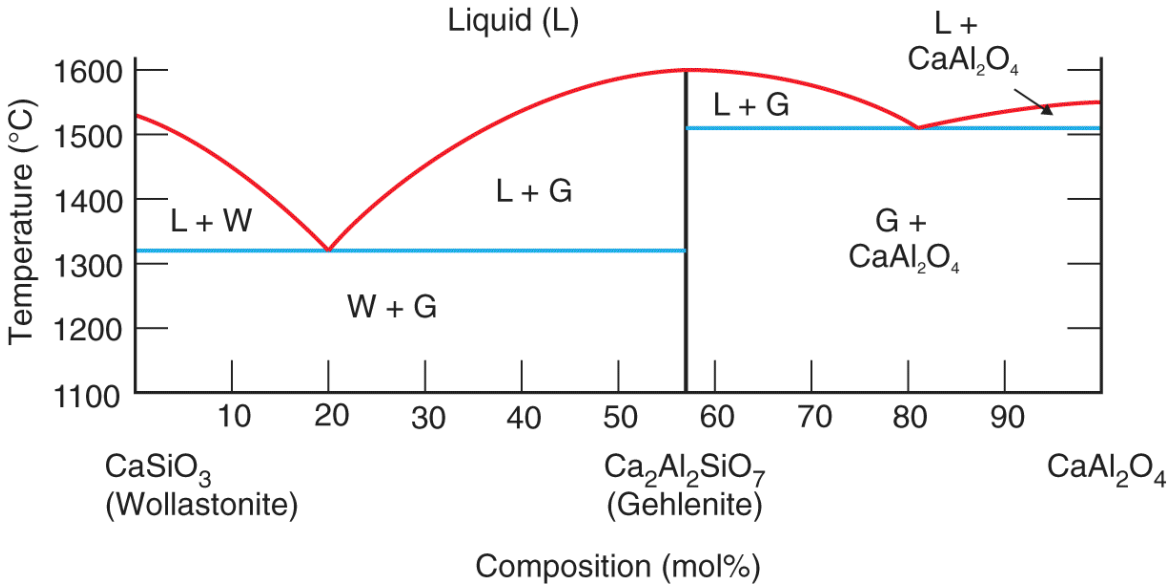


Figure 2.16 The $CaSiO_3$ (wollastonite)–$CaAl_2O_4$ phase diagram at 1 atm pressure.

None of the phases form solid solutions or show a composition range, and such compounds are often called *line phases*. A diagram containing an intermediate phase can often be treated as a composite of several simpler phase diagrams joined side by side. Thus, Figure 2.16 can be regarded as a composite of the $CaSiO_3$–$Ca_2Al_2SiO_7$ and $Ca_2Al_2SiO_7$–$CaAl_2O_4$ diagrams. The same methods described earlier can be used to obtain quantitative information in each part of the diagram.

The phase diagram shows that gehlenite melts without any compositional changes occurring. This feature is called *congruent melting*. Not all intermediate compounds show congruent melting. Many transform into a liquid plus another solid of a different composition and the solid is said to melt *incongruently*. The composition and temperature at which this occurs are called *peritectic* points and here three phases coexist. The point is thus an invariant point, and the reaction is an invariant reaction. A peritectic reaction is reversible and can be written as:

$$\text{Liquid} + \text{solid 1} \leftrightarrow \text{solid 2}$$

The V_2O_5–MgO system shows three intermediate phases MgV_2O_6, $Mg_2V_2O_7$, and $Mg_3V_2O_8$, all of which show incongruent melting (Figure 2.17). For example, the phase MgV_2O_6 melts at 756 °C to form solid $Mg_2V_2O_7$ and a liquid with a composition given by the point where the tie-line intersects the liquidus, 32 mol% MgO:

$$\mathrm{MgV_2O_5(s) \rightarrow Mg_2V_2O_7(s)} + \mathrm{L\,(32\,mol\%MgO)}$$

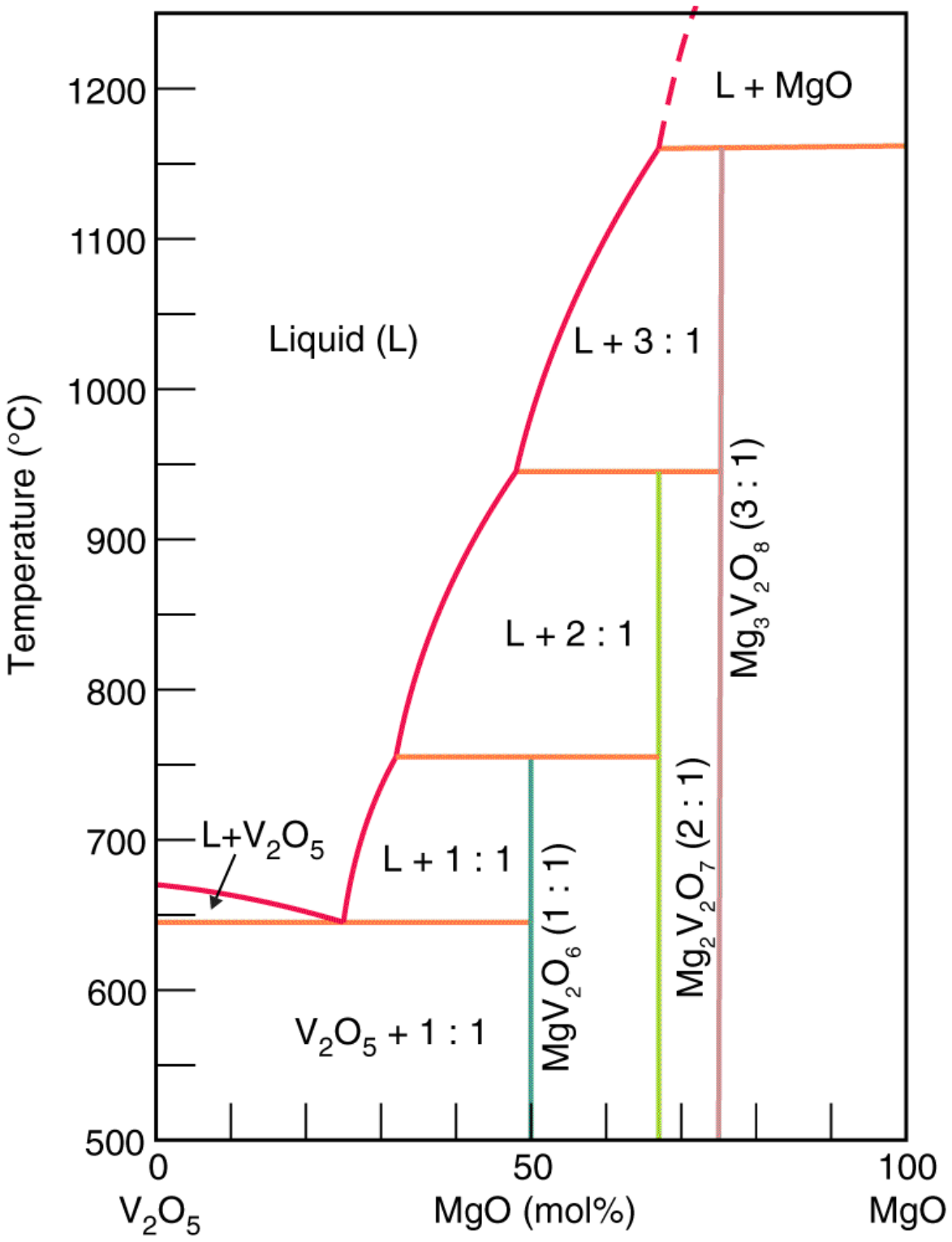


Figure 2.17 The V_2O_5–MgO phase diagram at 1 atm pressure.

As the reaction is reversible, slow cooling the resultant liquid will regenerate MgV_2O_5.

2.3.2.4 The Iron–Carbon System Close to Iron

The systematic understanding of phase relations in the low carbon region of the iron–carbon system at the end of the nineteenth century and the early years of the twentieth was at the heart of the technological advances that characterise these years. The version most used is that in which the composition axis is specified in weight % carbon (Figure 2.18). However, this is not the *equilibrium* phase diagram of the system as the intermediate compound *cementite*, Fe_3C, is metastable and slowly decomposes. The true equilibrium is between iron and graphite. However, cementite is an important constituent of steel and the rate of decomposition is slow under normal circumstances, so that the figure drawn is of most use for practical steel making. Cementite occurs at 6.70 wt% carbon, and has no appreciable composition range.

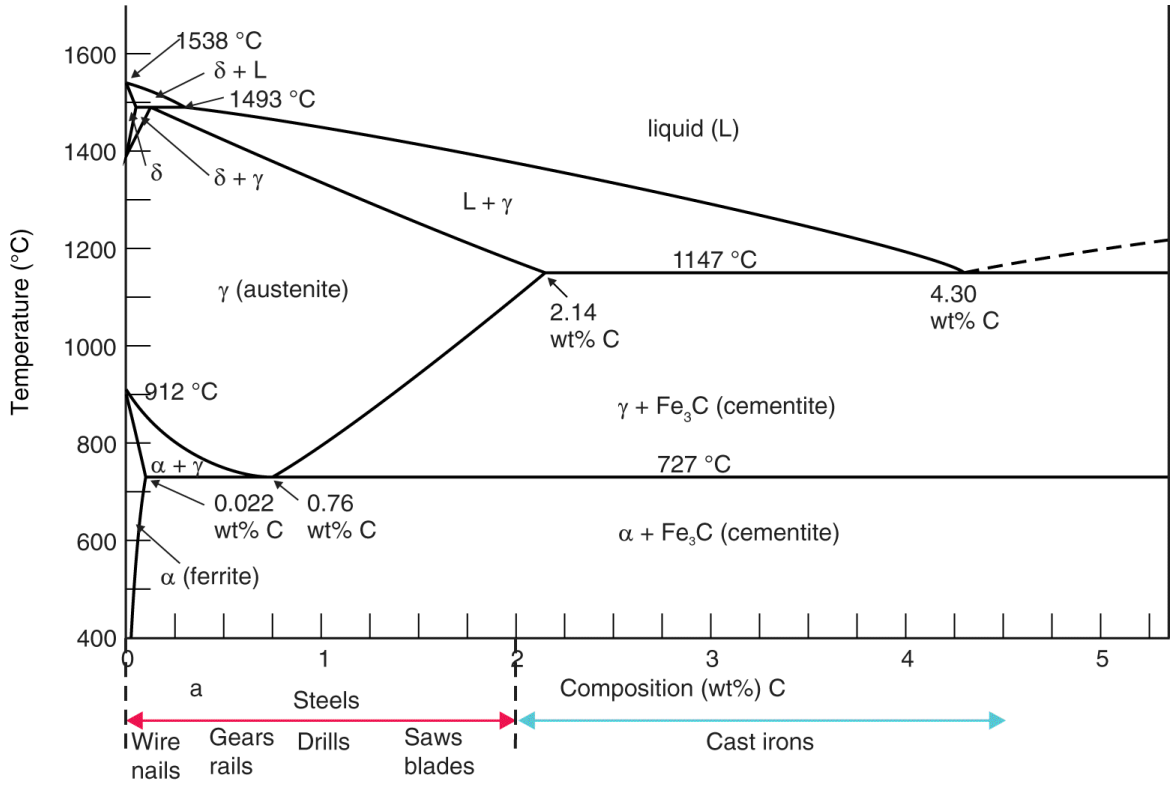


Figure 2.18 The iron-rich region of the iron–carbon existence diagram. (The α [ferrite] and δ phase fields have been slightly expanded for clarity.)

On the left-hand side of the diagram, the forms of pure iron are indicated. Below the melting point, 1538 °C, pure iron adopts one of three different crystal structures (called *allotropes*), at atmospheric pressure. Below a temperature of 912 °C, α-iron, which has a body-centred cubic structure (Section 3.2.2.2), is stable. This material can be made magnetic below a

temperature of 768 °C. The old name for the non-magnetic form of iron, which exists between temperatures of 768 and 912 °C, was β-iron, but this terminology is no longer in use. Between the temperatures of 912 and 1394 °C, the allotrope γ-iron is stable. This phase adopts the face-centred cubic structure (Section 3.2.2.1). At the highest temperatures, between 1394 and the melting point, 1538 °C, the stable phase is called δ-iron. The structure of δ-iron is the same as that of α-iron. It is rare that low temperature and high temperature polymorphs share the same crystal structure.

Between the pure iron allotropes and the intermediate phase cementite, a number of solid-solution regions occur. The extent of these depends upon the crystal structure of the iron. Only a small amount of carbon can enter the body-centred cubic form of α-iron. At 727 °C, this amounts to 0.022 wt% C. This solid solution is called *ferrite* or sometimes α-ferrite if it needs to be differentiated from the high temperature solid solution**.** More carbon can be taken into solid solution in the face-centred cubic form of γ-iron, to a maximum of 2.14 wt% C at 1147 °C to form *austenite*. The amount of interstitial carbon that can enter the body-centred cubic structure of δ-iron is slightly larger than in α-iron, amounting to 0.09 wt% C at 1493 °C. This material is also called ferrite, but generally δ-ferrite, to distinguish it from α-ferrite.

There are three invariant points in the iron–carbon diagram. A eutectic point is found at 4.30 wt% carbon and 1148 °C, at which point the liquid transforms to the two solids γ and Fe_3C on cooling. A *eutectoid point* occurs at the lowest temperature of the austenite phase field, 727 °C, and at 0.76 wt% carbon. At a eutectoid point, a solid transforms to two solids on cooling, thus:

$$\gamma \rightarrow \alpha + Fe_3C$$

This eutectoid transformation is of great importance in steel making. The phase diagram also contains a *peritectic point*, at the highest temperature of the austenite phase region, 1493 °C. At a peritectic point, a liquid plus a solid transform into a different solid on cooling.

$$\delta + L \rightarrow \gamma$$

The existence diagram allows us to understand the difference between *plain carbon steels*, alloys of carbon and iron only, and cast irons. Plain carbon steels contain less than about 2 wt% carbon, although commercial

steels rarely contain much more than 1.4 wt% C. They can be heated to give a homogeneous austenite solid solution. In this condition, they can readily be worked or formed. *Low-carbon steel*, with less than 0.15 wt% C, is ductile, not very hard, and is used for wires. *Mild carbon steel*, containing 0.15–0.20 wt% C is harder and less ductile. It is used for cables, chains, nails, and similar objects. *Medium-carbon steel*, containing 0.20–0.60 wt% C, is used for nails, girders, rails, and structural steels. *High-carbon steel*, containing 0.61–1.5 wt% C, still well inside the austenite phase region, is used in applications requiring greater hardness, such as knives, razors, cutting tools, and drill bits. *Ultrahigh-carbon steels*, containing between 1 and 2 wt% carbon, have been studied and found to accept extreme deformation before fracture, called *superelasticity* (Section 8.1).

When the carbon content is greater than about 2 wt% and less than about 5 wt%, the material cannot be heated to give a homogeneous solid solution. At all temperatures below the eutectic temperature of 1148 °C, the solid is a mixture of austenite and cementite or ferrite and cementite. The effect of this is that the materials are hard, brittle, and resist deformation. The material can be casted into the desired shape, and is referred to as *cast iron*. Commercial cast irons rarely contain much more than about 4.5 wt% C.

2.4 Ternary Systems

Ternary systems have three components. These require five-axis coordinate systems to display the phase relations, three for the compositions, one for pressure, and one for temperature. In practice, the three components are arranged at the vertices of an equilateral triangle, and the composition of each component is indicated along the sides of the triangle. The temperature axis is drawn normal to the composition plane, to form a triangular prism (Figure 2.19a). Each of the three faces of the prism is the binary phase diagram A–B, A–C, or B–C.

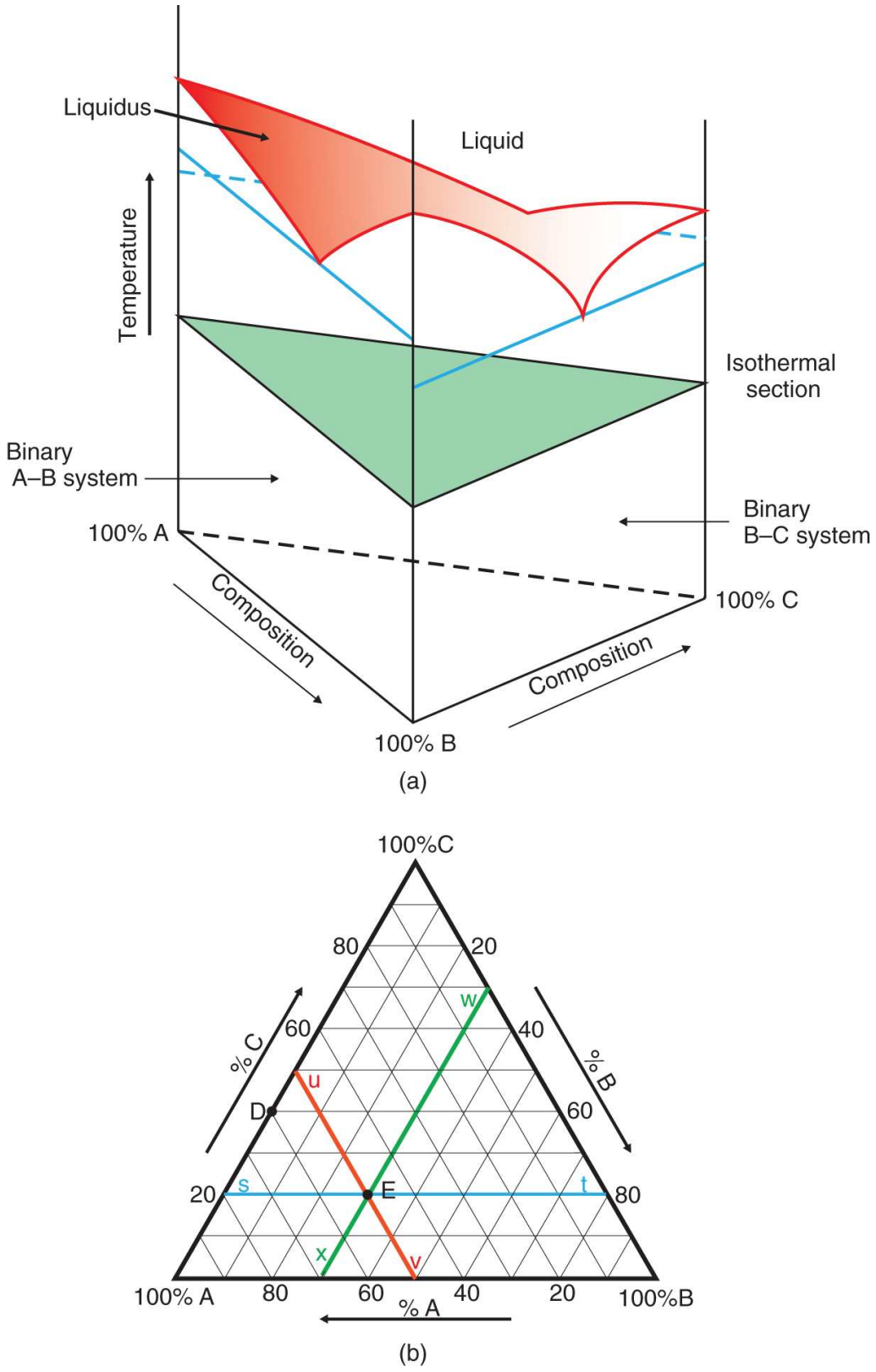


Figure 2.19 (a) The general form of an isobaric ternary phase diagram; (b) Representation of compositions on an isothermal section through (a).

Working phase diagrams are normally sections through the prism at a chosen value of temperature and a pressure of one atmosphere, and are called *isothermal sections* (Figure 2.19b). The composition of a point on one of the edges is read directly from the diagram. For example, point D (Figure 2.19b) represents a composition of 60% A and 40% C. The material consists of solid A + solid C. The amounts of the two phases can be determined via the lever rule, as explained in the following text. The composition of an internal point such as E is also found from the composition axes. The point lies on the line s–t. This line is the locus of all points with a composition of 20% C, and so E corresponds to 20% C. Similarly, it lies on line u–v, corresponding to the locus of all points containing 50% A and line w–x, corresponding to the locus of all points containing 30% B. The composition at E is therefore 50% A, 30% B, and 20% C.

Phases, with or without composition ranges, are plotted in an analogous way to those on binary phase diagrams. For example, Figure 2.20a is the diagram for the system WO_3–WO_2–ZrO_2 at approximately 1400 °C, which is part of the ternary W–Zr–O system. No phase has a composition range and none lie within the body of the phase diagram. It is seen that the area of the diagram is dived up into triangles by joining the phases with lines. A point in such a diagram will represent either three solid phases, if it lies within a triangle, two if it lies on a triangle edge, or one if it lies at a triangle vertex. Thus, point G represents a composition containing WO_2, WO_3, and ZrO_2 at 1400 °C. A composition represented by point H would consist of the phases ZrO_2 and $W_{18}O_{49}$.

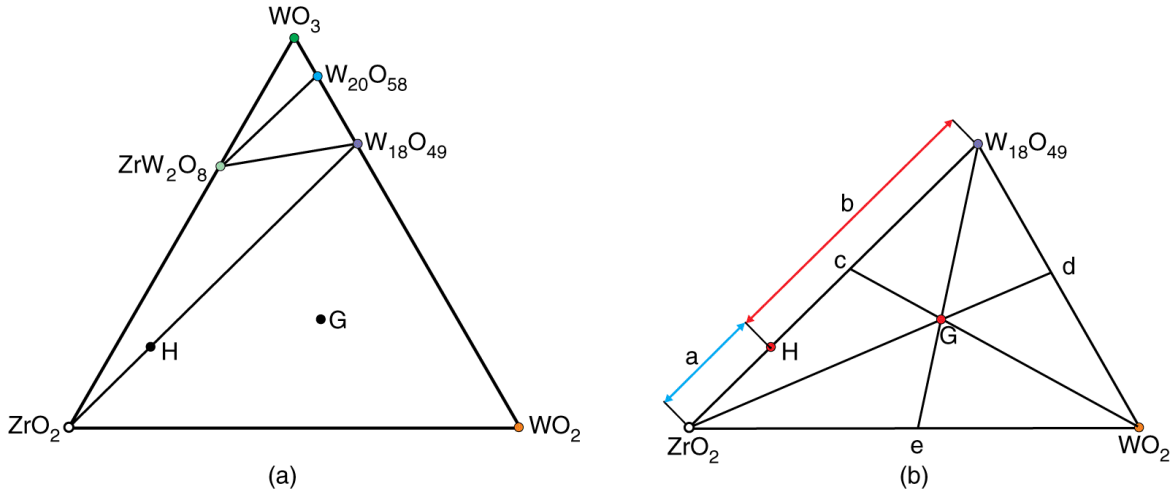


Figure 2.20 The simplified WO_3–WO_2–ZrO_2 phase diagram; (b) determination of compositions (see Section 1.4.1).

The amount of a phase present at a vertex of a triangle is 100%. The amounts of the phases present for points lying on the side of a triangle can be determined by the lever rule. For example, the amounts of the two phases present at point H in Figure 2.20b are given by:

Amount of $W_{18}O_{49}$ = a/(a + b)

Amount of ZrO_2 = b/(a + b)

The amounts of the three phases present at point G can be determined by an extension of the lever rule. The phase triangle made up of $W_{18}O_{49}$–WO_2–WO_3 is called a *tie triangle*, by analogy with the tie line of binary systems. Lines are drawn connecting the point G to the vertices of the tie triangle (Figure 2.20b). The amounts of the phases are then given by the lengths of these lines. For example,

Amount of $W_{18}O_{49}$ = distance e to G/distance e to $W_{18}O_{49}$

Amount of ZrO_2 = distance d to G/distance d to ZrO_2

Amount of WO_2 present = distance c to G/distance c to WO_2

This method is called the *triangle rule*. Note that, just as in the lever rule, we assume that the composition scales are linear. If they are not, actual compositions must be used, not distances. However, this is rarely found in ternary diagrams, which almost always use a linear scale for the composition axes.

Many phases display composition ranges, as described for some of the binary systems earlier. In these cases, the appearance of the ternary diagram is slightly modified. For example, in the FeO–Fe_2O_3–TiO_2 system at about 500 °C, Fe_3O_4 forms a complete solid solution with Fe_2TiO_4; both adopting the spinel structure (Section 3.2.2.8) with formula $Fe_{3-x}Ti_xO_8$, where x takes any value between 0 and 1. Within the phase triangle limited by FeO–Fe_3O_4–Fe_2TiO_4 a point A will then only contain two phases, FeO and corresponding spinel solid solution, while all other phase triangles will contain three phases (Figure 2.21a). At temperatures close to 700 °C, the phase $FeTiO_3$ now forms a complete solid solution with Fe_2O_3; both adopting the corundum (Al_2O_3) structure, with formula $Fe_{2-x}Ti_xO_3$, where x takes any value between 0 and 1. In this case all of the phase triangles below the $FeTiO_3$–Fe_2TiO_5 tie-line will contain just two

phases. Point A, as earlier; point B, Fe_2TiO_4–Fe_3O_4 solid solution and $FeTiO_3$–Fe_2O_3 solid solution; point C, $FeTiO_3$–Fe_2O_3 solid solution and Fe_2TiO_5 (Figure 2.21b).

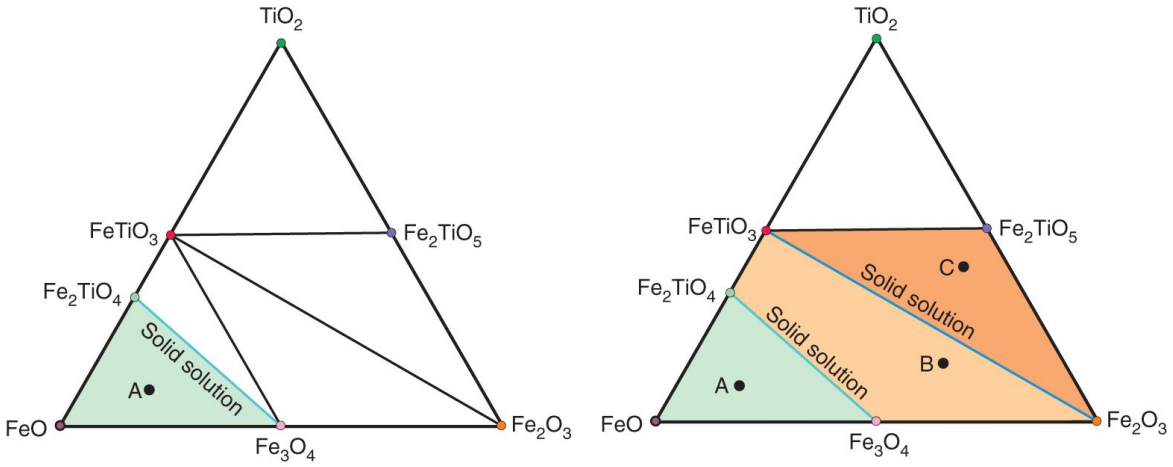


Figure 2.21 The FeO–Fe_2O_3–TiO_2 system at: (a) approximately 500 °C; (b) approximately 700 °C. Shaded areas contain two phases and open areas three phases.

Ceramic bodies are typically multiphase materials and ternary phase diagrams can indicate the likely phase make-up of these solids. The MgO–Al_2O_3–SiO_2 system contains important cordierite and steatite ceramics (Figure 2.22). Although these materials exist over considerable composition ranges, they are not single phases with a variable composition but consist of phase assemblies. The diagram shows that both may contain up to five different phases, depending upon the composition of the starting mix. (Note that both sapphirine and mullite also have composition ranges which have been denoted by representative points in the figure.)

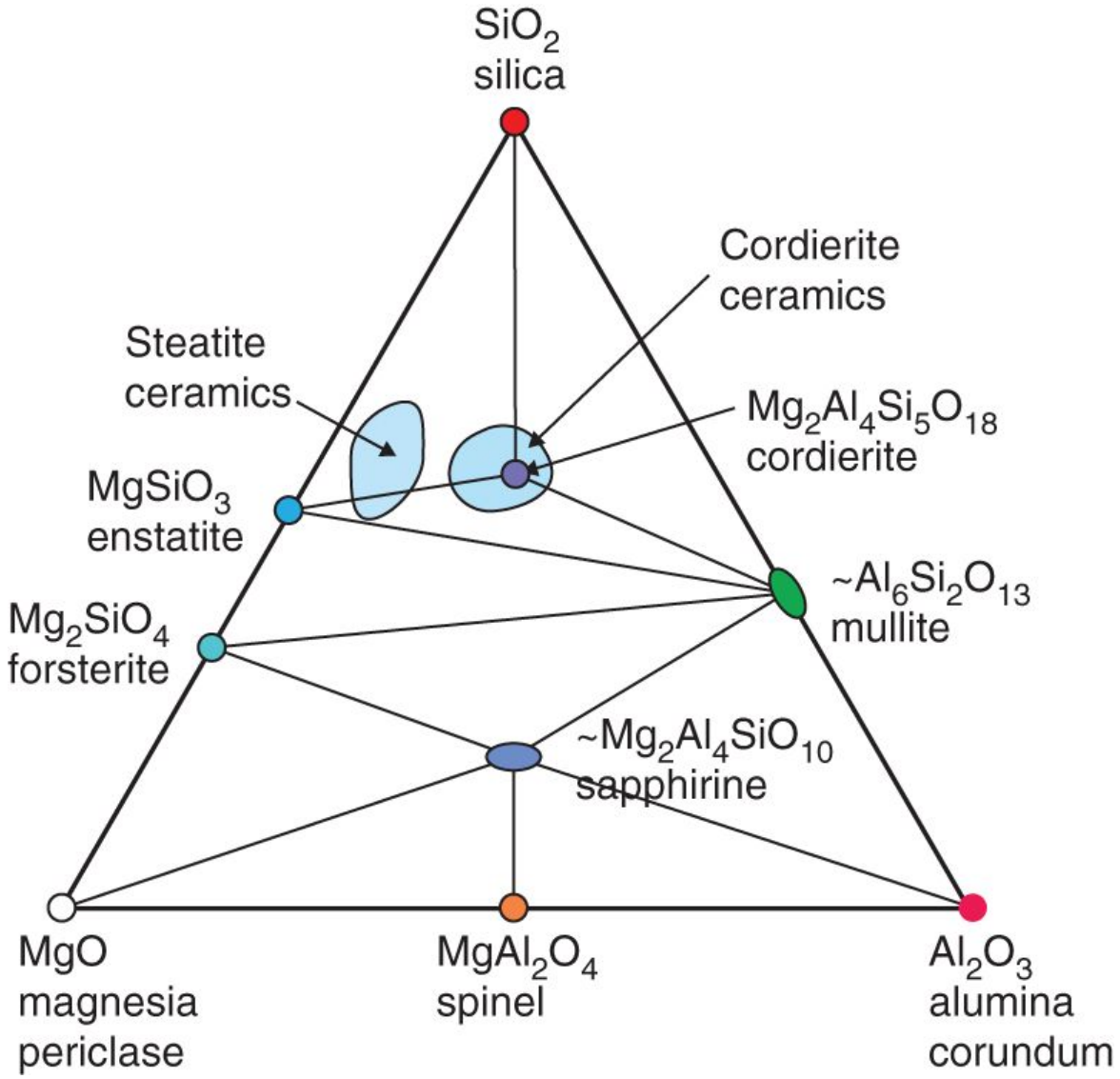


Figure 2.22 Approximate phase diagram for the $MgO–Al_2O_3–SiO_2$ system.

Reference

S. Cheong, A. Poerwoprajitno and R. D. Tilley, Electron Microscopy Centre, University of New South Wales, Australia

Further Reading

Ehlers, E.H. (1972). *The Interpretation of Geological Phase Diagrams*. San Francisco, CA: W. H. Freeman.

Lukas, H.L., Fries, S.G., and Sundman, B. (2007). *Computational Thermodynamics, the Calphad Method*. Cambridge: Cambridge University Press.

Levin, E.M., Robbins, C.R., and McMurdie, H.F. (1964, 1994), a continuing series, with changing editors). *Phase Diagrams for Ceramicists*, vol. 1, 10. Westerville, OH: American Ceramic Society.

Massalski, T.B. (ed.) (1990). *Binary Alloy Phase Diagrams*, 2nde, vol. 1–3. Materials Park, OH: ASM International.

Villars, P., Prince, A., and Okamoto, H. (1995). *Ternary Phase Diagrams*, vol. 1–10. Materials Park, OH: ASM International.

The scientific journal CALPHAD, is devoted to the computation of thermodynamic based phase properties and can be consulted for up-to date research.

Problems and Exercises

Quick Quiz

2.1 A phase is:

a. A compound in a system.

b. An element in a system.

c. A homogeneous part of a system.

2.2 The phase diagram of a one-component system is described in terms of

a. One variable.

b. Two variables.

c. Three variables.

2.3 How many phases co-exist at a triple point in the iron phase diagram?

a. One.

b. Two.

c. Three.

2.4 On a phase boundary in the sulfur system:

a. One phase exists.

b. Two phases exist.

c. Three phases exist.

2.5 At a critical point in a unary system:

a. Vapour and liquid cannot be distinguished.

b. Solid and liquid cannot be separated.

c. Freezing cannot occur.

2.6 A binary system is one in which:

a. There are two components present.

b. There are two variables needed.

c. A solid and a liquid are present.

2.7 The number of variables needed to specify phase relations in a binary system are:

a. Two.

b. Three.

c. Four.

2.8 The liquidus is a boundary that separates:

a. Two different liquids.

b. The liquid from the solid phase.

c. The liquid from the solid + liquid region.

2.9 The solidus is a boundary that separates:

a. A solid from a liquid.

b. A solid from a solid plus a liquid.

c. Two different solid phases.

2.10 A tie line is drawn:

a. Parallel to the pressure axis.

b. Parallel to the temperature axis.

c. Parallel to the composition axis.

2.11 A binary phase diagram in which the pressure is always constant is:

a. An isobaric section.

b. An isostatic section.

c. An isothermal section.

2.12 The solvus line on a binary phase diagram separates:

a. A solid from a solid plus liquid region.

b. A solid solution from a two-solid region.

c. A solid from a solid solution region.

2.13 A eutectic point on a binary phase diagram is:

a. An invariant point.

b. A triple point.

c. A critical point.

2.14 Cooling a homogeneous liquid sample through a eutectic point produces:

a. A liquid plus a solid.

b. Two solid phases.

c. A homogeneous solid.

2.15 A line phase on a binary phase diagram is a phase that:

a. Exists along phase boundary.

b. Has a composition range along a tie line.

c. Has no apparent composition range.

2.16 Solid phases that melt congruently melt to form:

a. A liquid of the same composition and a solid of a different composition.

b. A solid of the same composition and a liquid of a different composition.

c. A liquid of the same composition.

2.17 Heating a solid through a peritectic point produces:

a. A solid with a different composition and a liquid.

b. A solid with the same composition and a liquid.

c. A liquid with a different composition.

2.18 Pure iron has:

a. Two allotropes.

b. Three allotropes.

c. Four allotropes.

2.19 Steel is an alloy of iron and carbon in which the carbon occupies:

a. Substitutional sites.

b. Interstitial sites.

c. Vacancies.

2.20 Ferrite is:

a. An allotrope of iron

b. An intermediate phase.

c. An iron–carbon alloy.

2.21 Austenite has:

a. No appreciable composition range.

b. A narrow composition range compared to ferrite.

c. A wide composition range compared to ferrite.

2.22 Austenite has the same crystal structure as:

a. α-iron.

b. β-iron.

c. γ-iron.

2.23 Steel is an alloy of iron that has a composition less than:

a. The maximum austenite composition.

b. The maximum ferrite composition.

c. The maximum cementite composition.

2.24 Cast irons generally have compositions of iron and

a. Exactly 2 wt% carbon.

b. More than 2 wt% carbon.

c. Less than 2 wt% carbon.

2.25 On cooling a homogeneous solid phase through a eutectoid point, it forms:

a. Two solid phases.

b. A solid and a liquid phase.

c. A homogeneous single solid phase.

2.26 A ternary system is one in which there are:

a. Three components.

b. Three variables.

c. Three phases present.

2.27 To represent all possible phase relations, a ternary system needs:

a. Five axes.

b. Four axes.

c. Three axes.

2.28 Within a tie triangle on a ternary phase diagram, a point will usually correspond to:

a. Two solid phases.

b. Three solid phases.

c. Four solid phases.

Calculations and Questions

2.1 A copper–nickel alloy is made up with 28 g of copper in 100 g alloy. What is the atom % Ni in the alloy?

2.2 A solid solution of aluminium oxide (Al_2O_3), and chromium oxide (Cr_2O_3), has a composition $Al_{0.70}Cr_{1.30}O_3$. What mass of Cr_2O_3 needs to be weighed out to prepare 100 g of sample?

2.3 A copper–zinc alloy is made up with 35 g Cu in 100 g alloy (35 wt%). What is the atom percentage Zn in the alloy?

2.4 A solder contains 50 wt% Sn and 50 wt% Pb. What is the atom % of Sn and Pb in the solder?

2.5 An intermetallic compound in the Ti–Al system is found at 78 wt% Ti and 22 wt% Al. What is the approximate formula of the compound?

2.6 An equilibrium sample of a copper–nickel alloy with a composition of 65 wt% Ni is prepared. With reference to the Cu–Ni phase diagram (Figure 2.12):

a. What is the atom percentage of copper present in the alloy?

b. On heating the alloy from room temperature, at what temperature does liquid first appear?

c. What is the composition of the liquid?

d. At what temperature does the solid disappear?

e. What is the composition of the final solid?

2.7 For the sample in the previous question, the alloy is held at a temperature of 1340 °C.

a. What phase(s) is(are) present?

b. How much solid is present, if any?

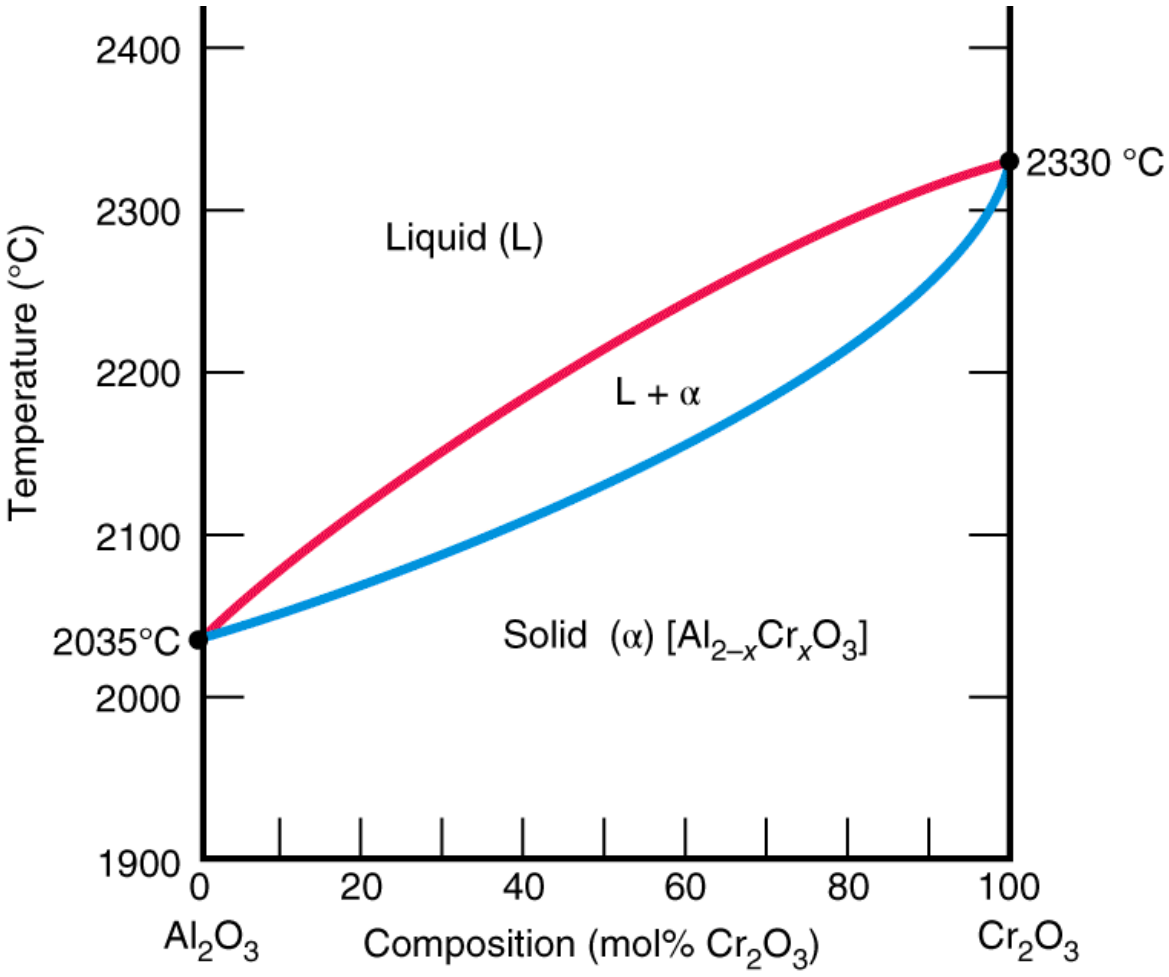
c. How much liquid is present, if any?

2.8 With reference to the Cu–Ni phase diagram (Figure 2.12), an equilibrium sample of a copper–nickel alloy with a composition of 37 wt% Ni is held at a temperature of 1250 °C.

a. What are the amounts of solid and liquid present?

b. If the density of the solid is 8.96×10^3 kg m^{-3}, and the density of the liquid is 90% of that of the solid, calculate the volume percentage of the solid and liquid present.

2.9 An equilibrium sample of composition 50 mol% Cr_2O_3 is prepared using the phase diagram of the Al_2O_3–Cr_2O_3, system:



a. What is the weight percentage of Al_2O_3 present?

b. The sample is held at 2200 °C. What is the composition of the solid phase present?

c. How much liquid phase is present?

d. At what temperature will the last of the solid disappear?

e. What will the composition of this solid be?

2.10 With reference to Question 2.9, a sample of composition 30 mol% Cr_2O_3 is held at 2300 °C and then slowly cooled.

a. At what temperature does solid first appear?

b. What is the composition of the solid?

c. At what temperature does the liquid finally disappear?

d. What is the composition of the last drop of liquid?

2.11 With reference to Question 2.9, a sample of composition 30 mol% Cr_2O_3 is held at 2080 °C.

a. What is the composition of any solid phase present?

b. What is the composition of any liquid phase present?

c. How much of each phase is present?

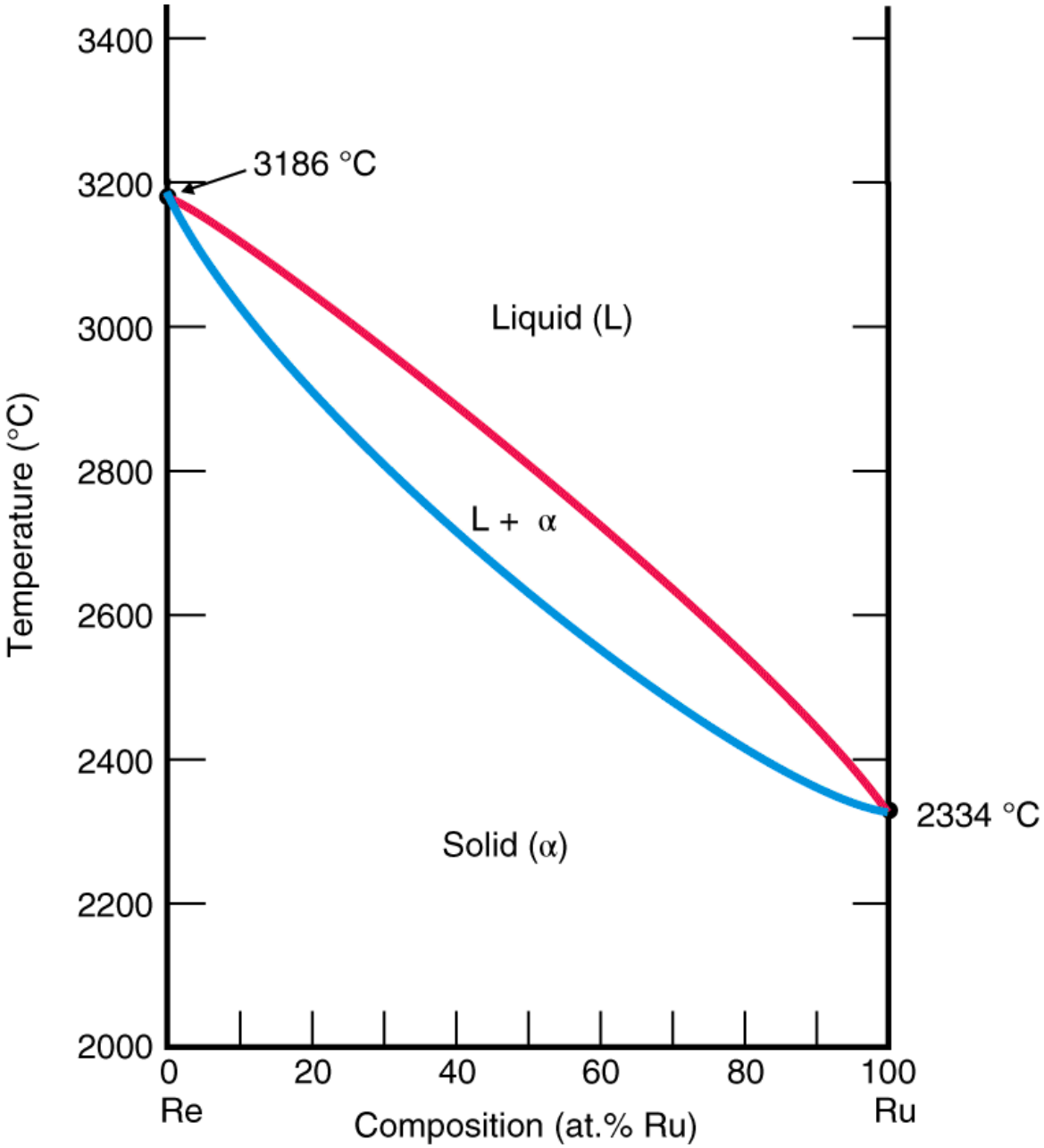
2.12 With reference to Question 2.9, a sample of composition 30 mol% Cr_2O_3 is held at 2100 °C.

a. What is the composition of any solid phase present?

b. What is the composition of any liquid phase present?

c. How much of each phase is present?

2.13 The phase diagram of the ruthenium (Ru), rhenium (Re), system is:



a. What are the melting points of pure Re and pure Ru?

b. A sample of composition 70 at.% Ru is made up. What weights have to be added to prepare 100 g of sample?

c. This alloy is held at a temperature of 2200 °C. What phases are present and what are their compositions?

d. The alloy is held at 3000 °C. What phases are present and what are their compositions?

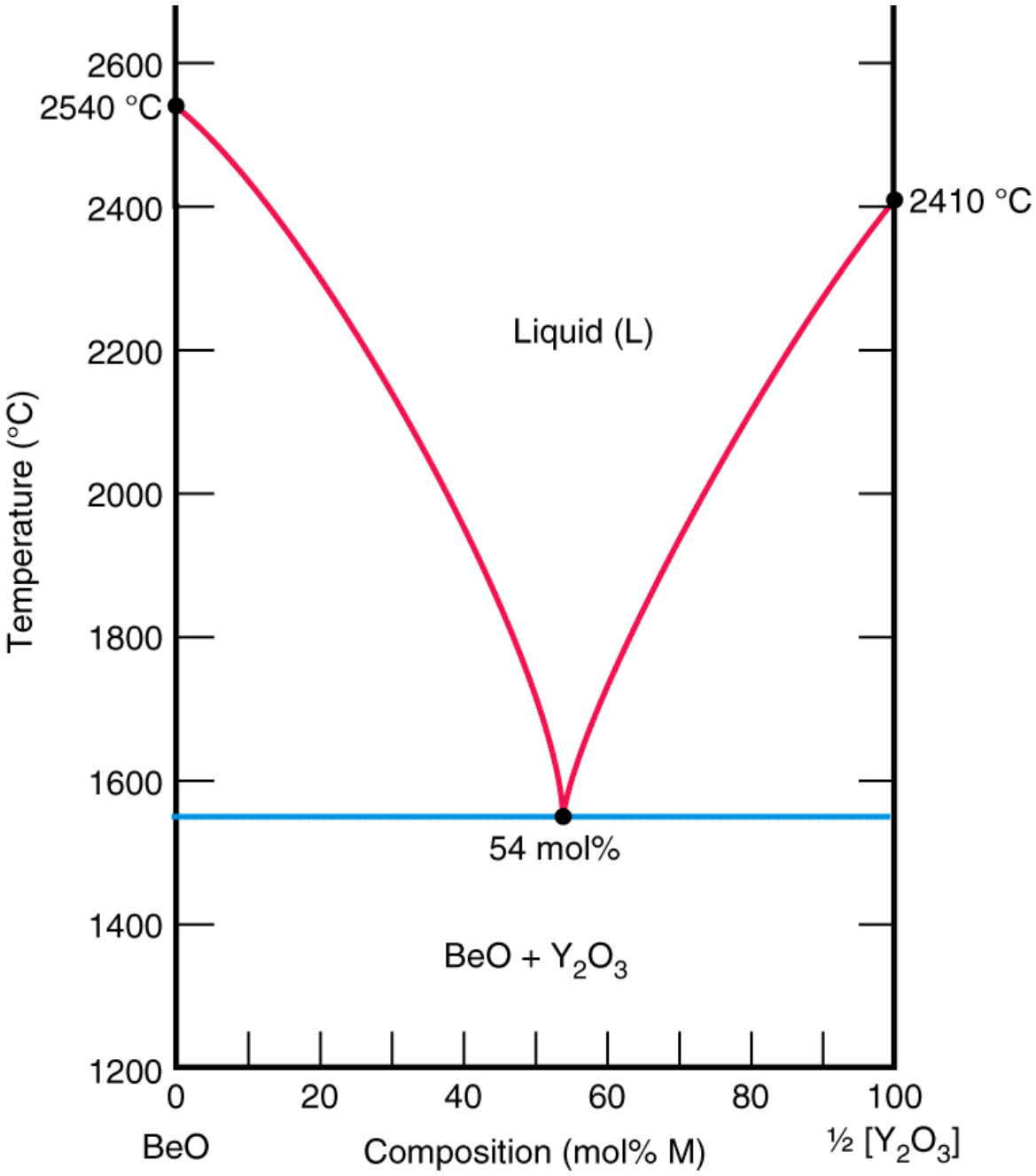
2.14 With reference to Question 2.13, a sample of composition 60 at.% Ru is made up and held at 2700 °C.

a. What is the composition of any solid phase present?

b. What is the composition of any liquid phase present?

c. How much of each phase is present?

2.15 The phase diagram of the BeO–Y_2O_3 system is given in the following text. Explain why this figure differs from that of the lead–tin system, Figure 2.14. Why does the composition axis use ½ [Y_2O_3] instead of Y_2O_3?



2.16 With respect to the phase diagram of the BeO–Y_2O_3 system, a composition is made up with 80 mol% BeO and held at 2000 °C.

a. What weights of the components are needed to make 100 g of sample?

b. How much solid is present?

c. What is the composition of the solid?

d. How much liquid is present?

e. What is the composition of the liquid?

2.17 With respect to the sample in the previous question, the material is cooled to 1400 °C.

a. What phases are present?

b. What are the compositions of the phases?

c. What are the proportions of each phase present?

2.18 With respect to the Pb–Sn phase diagram (Figure 2.14), a sample is made up with 40 at.% Sn.

a. What weights of lead and tin are needed to make 100 g solid?

b. What phase(s) is(are) present when the sample is held at 300 °C?

c. What are the compositions of the phase(s)?

d. How much of each phase is present at 300 °C?

2.19 With respect to the Pb–Sn phase diagram (Figure 2.14), a sample made up with 40 at.% Sn is cooled slowly to 250 °C.

a. What phases are present at 250 °C?

b. What is the composition of each phase?

c. How much of each phase is present?

2.20 With respect to the Pb–Sn phase diagram (Figure 2.14), the sample made up with 40 at.% Sn is cooled further to 100 °C.

a. What phases are present at 100 °C?

b. What is the composition of each phase?

c. How much of each phase is present?

2.21 With respect to the Fe–C phase diagram (Figure 2.18), an alloy with a composition of 1.5 wt% C is homogenised by heating for a long period at 1000 °C.

a. What phases are present?

b. How much of each phase is present?

c. What are the compositions of the phases?

2.22 With respect to the Fe–C phase diagram (Figure 2.18), an alloy with a composition of 1.5 wt% C is homogenised by heating for a long period at 1000 °C, is subsequently cooled slowly to 800 °C.

a. What phase(s) is(are) present?

b. How much of each phase is present?

c. What are the compositions of the phase(s)?

2.23 With respect to the Fe–C phase diagram (Figure 2.18), an alloy with a composition of 5 at.% C is homogenised by heating for a long period at 1350 °C.

a. What is the composition of the liquid phase present?

b. How much of the liquid phase is present?

c. What is the composition of the solid phase present?

d. How much solid phase is present?

2.24 With respect to the phase diagram of the WO_3–WO_2–ZrO_2 system (Figure 2.20), a sample is made up of an equimolar mixture of WO_3, WO_2 and ZrO_2 (1 : 1 : 1) and heated at 1100 °C to equilibrium.

a. What phases are present?

b. How much of each phase is present?

2.25 With respect to the phase diagram of the WO_3–WO_2–ZrO_2 system (Figure 2.20), a sample is made up of a mixture of 80 mol% WO_3, 10 mol% WO_2, and 10 mol% ZrO_2 (8 : 1 : 1) and heated at 1100 °C to equilibrium.

a. What phases are present?

b. How much of each phase is present?

2.26 According to the SiO_2–MgO–Al_2O_3 phase diagram (Figure 2.22), what phases may be present in steatite ceramics and cordierite ceramics?

3
Crystal Structures and Defects

- How does a lattice differ from a crystal structure?
- What is a unit cell?
- What are point defects?

Crystallography describes the ways in which the component atoms of a solid are arranged in crystals. Crystallography has developed via three routes. The first of these was observational, especially the early studies of mineral crystals. A second route, the mathematical description of the arrangement of objects in space, including the concept of symmetry, was developed in the latter years of the nineteenth century. The two approaches were unified with the exploitation of X-ray and other diffraction methods, which are now used to determine crystal structures, even of complex biological materials, on a routine basis. These structural characterisation techniques also lead to the characterisation of the structural defects present in crystalline solids and an appreciation of their importance in controlling both the physical and chemical properties of the bulk.

3.1 Crystal Geometry

3.1.1 Crystal Systems

It was long supposed that the beautiful shapes of mineral crystals were an expression of internal order and this regularity was described by the classification of these external shapes, the morphology or *habit* of crystals. The observation that many crystals could be cleaved into smaller and smaller units gave rise to the idea that all crystals were built up from elementary volumes, which came to be called (morphological) *unit cells*, with a shape defined by the crystal habit, and allocated to six *crystal families*, called anorthic, monoclinic, orthorhombic, tetragonal, hexagonal, and isometric. Later studies, especially the determination of the atomic positions in a crystal, demonstrated that all crystals could be classified into *seven crystal systems*, triclinic, monoclinic, orthorhombic, tetragonal, hexagonal, rhombohedral, and cubic (isometric). Each member is described by a set of reference axes, **a**, **b**, and **c**, which have a direction, and hence are vectors.[1] The parallelepiped formed by the three vectors **a**, **b**, and **c**, defines the *unit cell* of the lattice, with edges of length *a*, *b*, and *c*. The numerical values of the unit cell sides and the angles between them are collectively called the *lattice parameters* or (*unit*) *cell parameters*.

Clearly there are a number of ways of choosing axes or the corresponding unit cell for any crystal. In order to avoid ambiguity, the axes are chosen to form a conventional right-handed set, and are drawn so that the **a**-axis points out from the page, the **b**-axis points to the right and the **c**-axis is vertical (Figure 3.1). The angles between the axes are chosen to be equal to or greater than 90° whenever possible. These are labelled α, β, and γ, where α lies between *b* and *c*, β lies between *a* and *c*, and γ lies between *a* and *b*. Together with the unit cell dimensions, these uniquely characterise all crystals (Table 3.1).

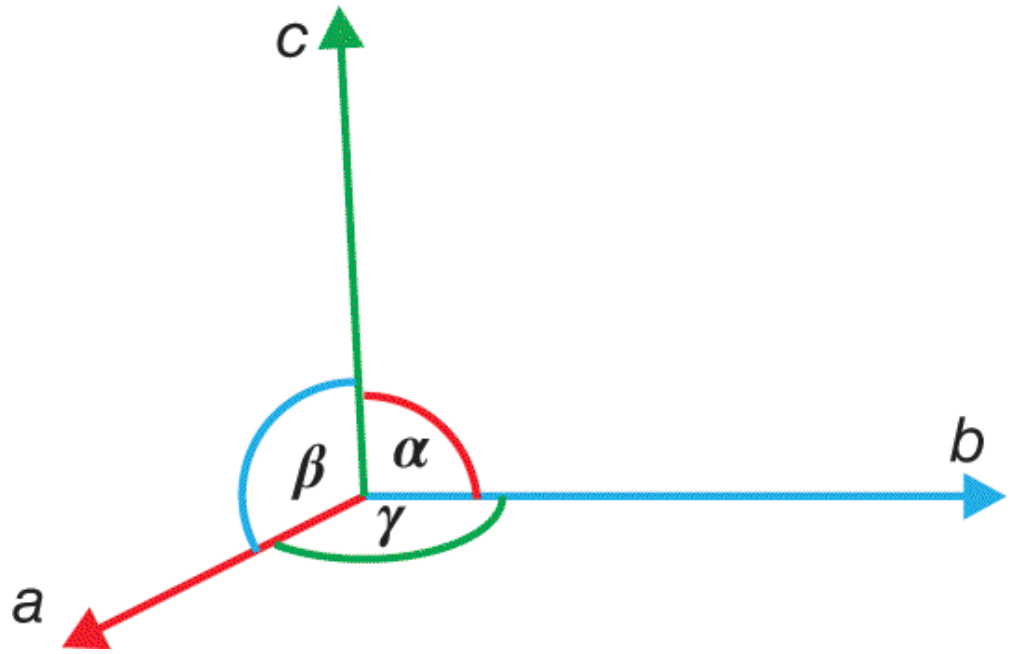


Figure 3.1 The reference axes used to characterise the crystal systems.

Table 3.1 The crystal systems.

System	Unit cell parameters
Cubic (isometric)	$a = b = c$; $\alpha = 90°$, $\beta = 90°$, $\gamma = 90°$
Tetragonal	$a = b \neq c$; $\alpha = 90°$, $\beta = 90°$, $\gamma = 90°$
Orthorhombic	$a \neq b \neq c$; $\alpha = 90°$, $\beta = 90°$, $\gamma = 90°$
Monoclinic	$a \neq b \neq c$; $\alpha = 90°$, $\beta \neq 90°$, $\gamma = 90°$
Triclinic	$a \neq b \neq c$; $\alpha \neq 90°$, $\beta \neq 90°$, $\gamma \neq 90°$
Hexagonal	$a = b \neq c$; $\alpha = 90°$, $\beta = 90°$, $\gamma = 120°$
Rhombohedral[a]	$a = b = c$; $\alpha = \beta = \gamma \neq 90°$ $a' = b' \neq c'$; $\alpha' = 90°$, $\beta' = 90°$, $\gamma' = 120°$

[a] Rhombohedral unit cells are often specified in terms of a bigger hexagonal unit cell.

3.1.2 Crystal Lattices

Crystal structures and crystal lattices are different, although these terms are frequently used as synonyms. A crystal *structure* is built of atoms. A crystal *lattice* is an infinite pattern of points, each of which must have the same surroundings in the same orientation. A lattice is a mathematical concept. If any lattice point is chosen as the origin, the position of any other lattice point is defined by:

$$\mathbf{P}(uvw) = u\mathbf{a} + v\mathbf{b} + w\mathbf{c}$$

where **a**, **b**, and **c** are vectors, called *basis vectors*, and u, v, and w are positive or negative integers. The parallelepiped formed by the three basis vectors **a**, **b**, and **c** defines the unit cell of the lattice, with edges of length a, b, and c. The numerical values of the unit cell side and the angles between them are the lattice parameters.

The seven crystal systems give rise to 14 three-dimensional lattices, called the *Bravais* lattices (Figure 3.2). Bravais lattices are sometimes called *direct lattices*. The smallest unit cell possible for any of the lattices, the one that contains just one lattice point, is called the *primitive* unit cell. A primitive unit cell, usually drawn with a lattice point at each corner, is labelled P. (Note that a lattice point at a unit cell corner is shared between eight neighbouring unit cells, and as each unit cell has eight corners, the cell contains one lattice point.) All other lattice unit cells contain more than one lattice point. A unit cell with a lattice point at each corner and one at the centre of the unit cell (thus containing two lattice points in total), is called a *body-centred* unit cell, and labelled I. A unit cell with a lattice point at each corner and one in the middle of each face, thus containing four lattice points, is called a *face-centred* unit cell, and labelled F. A unit cell with a lattice point at each corner and just one of the faces of the unit cell centred, thus containing two lattice points, is labelled *A-face-centred*, if the faces cut the **a**-axis, *B-face-centred* if the faces cut the **b**-axis and *C-face-centred* if the faces cut the **c**-axis.

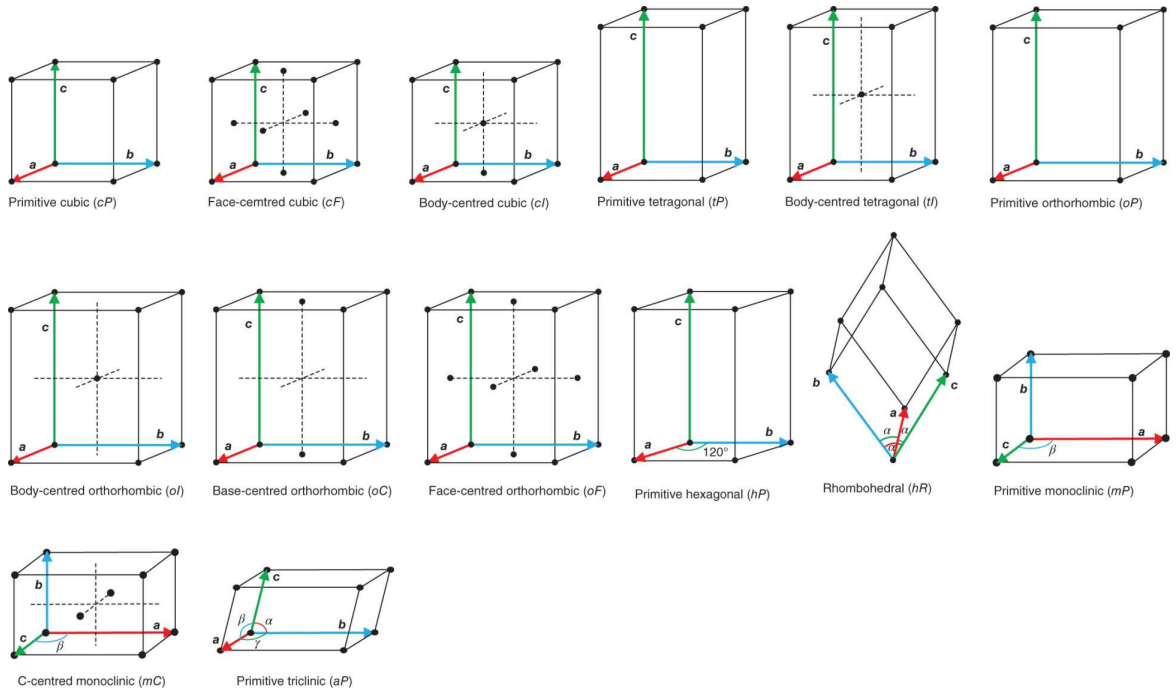


Figure 3.2 The 14 Bravais lattices. (Note that the lattice points are exaggerated in size and are not atoms. The monoclinic lattices have been drawn with the **b**-axis vertical, to emphasise that it is normal to the plane containing the *a*- and *c*-axes.)

A cubic lattice has three identical axes and a triclinic unit cell has no equivalent axes. Tetragonal, hexagonal, orthorhombic, and monoclinic unit cells all have one unique axis. The unique axis in the first three is designated the **c**-axis. The unique axis in a monoclinic unit cell is the **b**-axis.

3.1.3 Symmetry and Crystal Classes

The external shape, or habit, of a crystal is described as *isometric* (like a cube), *prismatic* (like a prism, often with six sides), *tabular* (like a rectangular tablet or thick plate), *lathy* (like a thin blade), or *acicular* (needle-like). An examination of the disposition of crystal faces leads to an appreciation that all crystals can be allocated to one of 32 *crystal classes*.

The crystal class mirrors the internal symmetry of the crystal. The internal symmetry of any isolated object, including a crystal, can be described by a combination of axes of rotation and mirror planes, all of which will be found to intersect in a point within the object. There are just 32 combinations of these symmetry elements that are found in crystals, each of which forms a crystallographic *point group*. The point group is

equivalent to the crystal class of a crystal, and the terms are often used interchangeably.

Point groups are used extensively in crystal physics to relate external and internal symmetry with observable physical properties. A centre of symmetry at a position (0, 0, 0) is such that any atom at a position (x, y, z) is accompanied by a similar atom at ($-x$, $-y$, $-z$), which, in crystallographic notation, is written with the negative signs above the symbols they apply to, thus $(\bar{x}, \bar{y}, \bar{z})$. Crystals that do not possess a centre of symmetry have one or more *polar directions* and *polar axes*. A polar axis is one that is not related by symmetry to any other direction in the crystal. That is, if an atom occurs at $+z$ on a polar **c**-axis, there is no similar atom at $-z$. This can be illustrated with reference to the commonly occurring polyhedral coordination shown by metal atoms. A regular octahedral group such as a TiO_6 octahedron is centrosymmetric, meaning that an O atom at $+z$ on the **c**-axis is paired with an O atom at $-z$. A regular SiO_4 tetrahedron is a group that lacks a centre of symmetry. In this latter case, the oxygen atom at $+z$ on the **c**-axis is not paired with a similar oxygen atom at $-z$ and the z-axis is a polar axis (Figure 3.3).

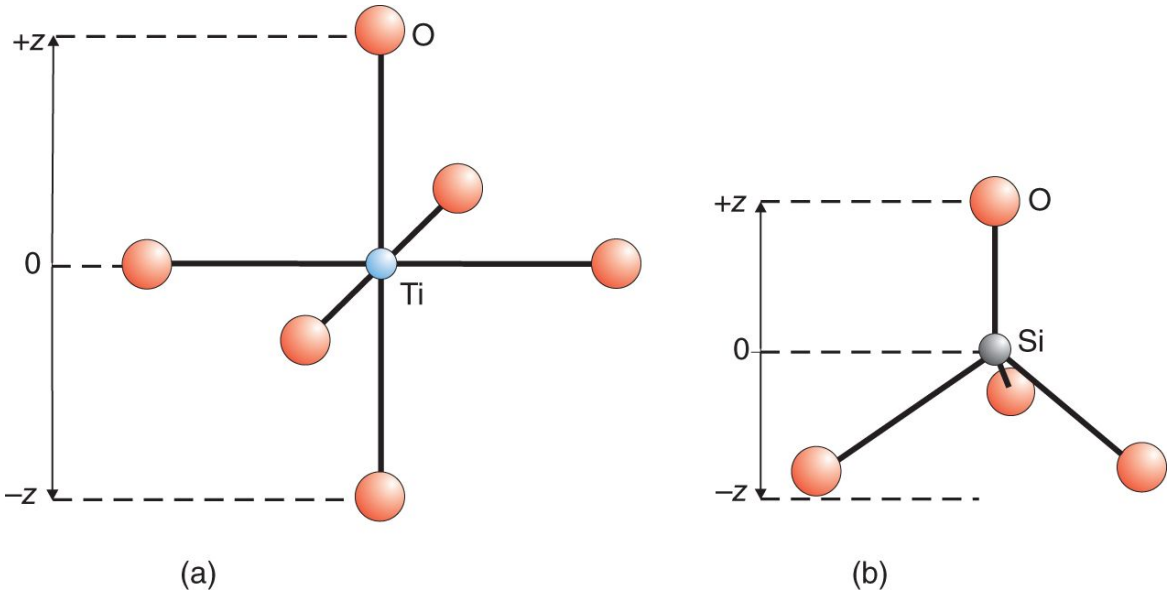


Figure 3.3 (a) A regular TiO_6 octahedron. (b) A regular SiO_4 tetrahedron.

The symmetry of the internal structure of a crystal is obtained by combining the point group symmetry with the symmetry of the lattice. It is found that 230 different patterns arise. These are called *space groups*. Every crystal structure can be assigned to a space group. The space group,

because it is concerned with the symmetry of the crystal structure, places severe restrictions on the location of atoms within the unit cell. The determination of a crystal structure generally starts with the determination of the correct space group for the crystal.

3.2 Crystal Structures

3.2.1 Unit Cells and Atomic Coordinates

All crystal structures can be built up from the Bravais lattices by placing an atom or a group of atoms at each lattice point. The crystal structure of a metal and that of a complex protein may both be described in terms of the same lattice, but whereas the number of atoms allocated to each lattice point is often just one for a metallic crystal, it may easily be thousands for a protein crystal. The number of atoms associated with each lattice point is called the *motif*. The motif is a fragment of structure that is sufficient, when repeated at each of the lattice points, to construct the whole of the structure. The crystal structure of a compound is thus specified by the unit cell type, the lattice parameters,[2] and the positions of the atoms in the cell.

The position of an atom within the unit cell is given as fractions x, y, and z, of the a, b, and c lattice parameters, respectively, and they are always measured parallel to the unit cell edge. An atom at the centre of a unit cell would have a position specified as (½, ½, ½), irrespective of the type of unit cell. Similarly, an atom at each corner of a unit cell is specified by (0, 0, 0). For an atom to occupy the centre of a face of the unit cell, the coordinates will be (½, ½, 0), (0, ½, ½), or (½, 0, ½), for C-face-centred, A-face-centred, and B-face-centred cells, respectively. Atoms at the mid-point of the unit cell edges are at positions (½, 0, 0), (0, ½, 0) or (0, 0, ½). The normal procedure of stacking the unit cells together means that these atoms will be duplicated in the same position in all unit cells (Figure 3.4). Because repetition of the unit cell must reproduce the crystal, the atomic contents of the unit cell must also be representative of the overall composition of the material.

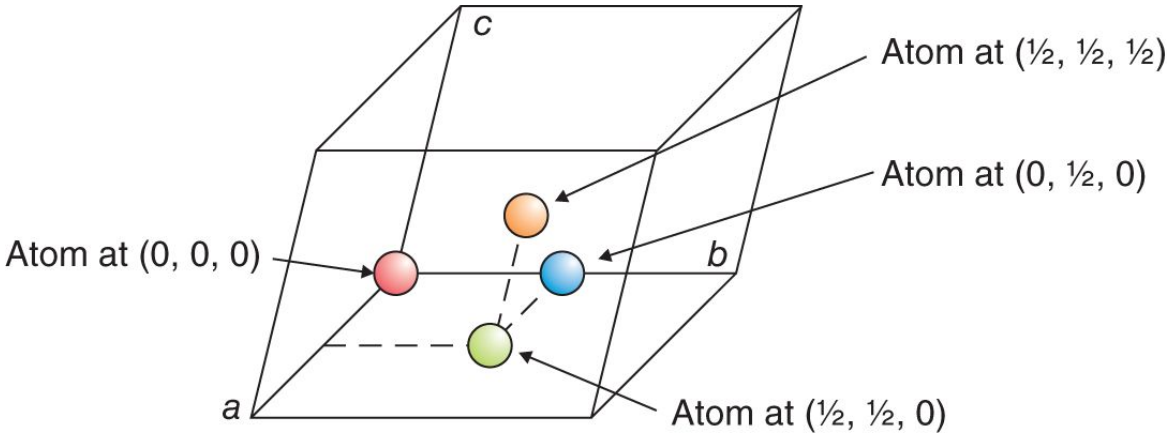


Figure 3.4 The positions of an atom in a unit cell, specified as fractions of the cell edges.

To avoid long repetitive lists of atom positions in complex structures, crystallographic descriptions usually list only the minimum number of atomic positions, which, when combined with the symmetry of the structure, given as the space group, generate all the atom positions in the unit cell. Additionally, the Bravais lattice type and the motif are often specified, as well as the number of formula units of the compound in the unit cell, written as *Z*.

A vast number of structures have been determined, and it is very convenient to group those with topologically identical structures together. Any one member of the group differs from any other in terms of the atoms in the unit cell, reflecting a change in chemical compound, but the atomic coordinates and unit cell dimensions change only slightly. Frequently, the group name is taken from the name of a mineral, as mineral crystals were the first solids used for structure determination. Thus, all solids with the halite (sodium chloride) structure have a unit cell similar to that of NaCl; examples being the oxides NiO, MgO, and CaO. Different compounds that crystallise with the same crystal structure, for example, the two alums, $NaAl(SO_4)_2{\cdot}12H_2O$ and $NaFe(SO_4)_2{\cdot}12H_2O$, are said to be *isomorphous*[3] or *isostructural*. Sometimes the crystal structure of a compound will change with temperature and with applied pressure. This is called *polymorphism*. Polymorphs of elements are known as *allotropes*. Graphite and diamond are two allotropes of carbon, formed at different temperatures and pressures.

3.2.2 Crystal Structures

3.2.2.1 The Face-Centred Cubic (fcc, A1) Structure

General formula M; example, copper Cu;

Lattice: cubic face-centred, a = 0.3615 nm.

Unit cell contents: Z = 4 Cu

Atom positions: 0, 0, 0; ½, ½, 0; 0, ½, ½; ½, 0, ½;

There are four lattice points in the face-centred cubic (fcc) unit cell and the motif is one atom at 0, 0, 0. Each atom has 12 nearest neighbours, and if the atoms are supposed to be hard touching spheres, the fraction of the volume occupied is 0.7405 (Figure 3.5).

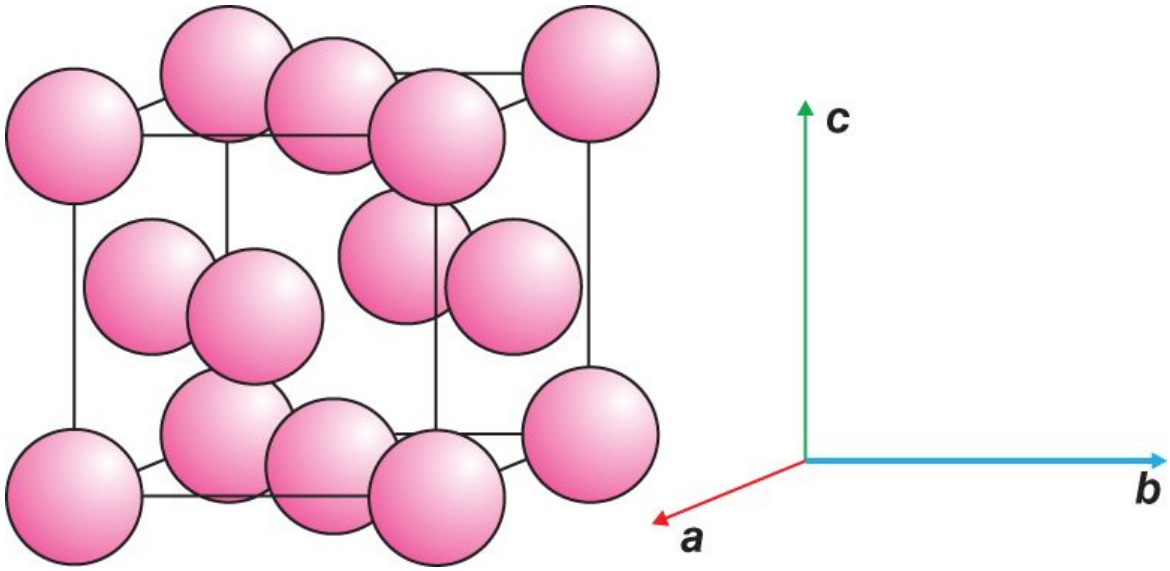


Figure 3.5 The face-centred cubic (fcc, A1) structure.

3.2.2.2 The Body-Centred Cubic (bcc, A2) Structure

General formula: M; example, tungsten W;

Lattice: cubic body-centred, a = 0.316 nm.

Z = 2 W

Atom positions: 0, 0, 0; ½, ½, ½;

There are two lattice points in the body-centred cubic (bcc) unit cell and the motif is one atom at 0, 0, 0. Each atom has eight nearest neighbours and six next nearest neighbours at only 15% greater distance. If the atoms are supposed to be hard touching spheres, the fraction of the volume occupied is 0.6802. The bcc structure is often the high temperature

structure of a metal that has a fcc structure at lower temperatures (Figure 3.6).

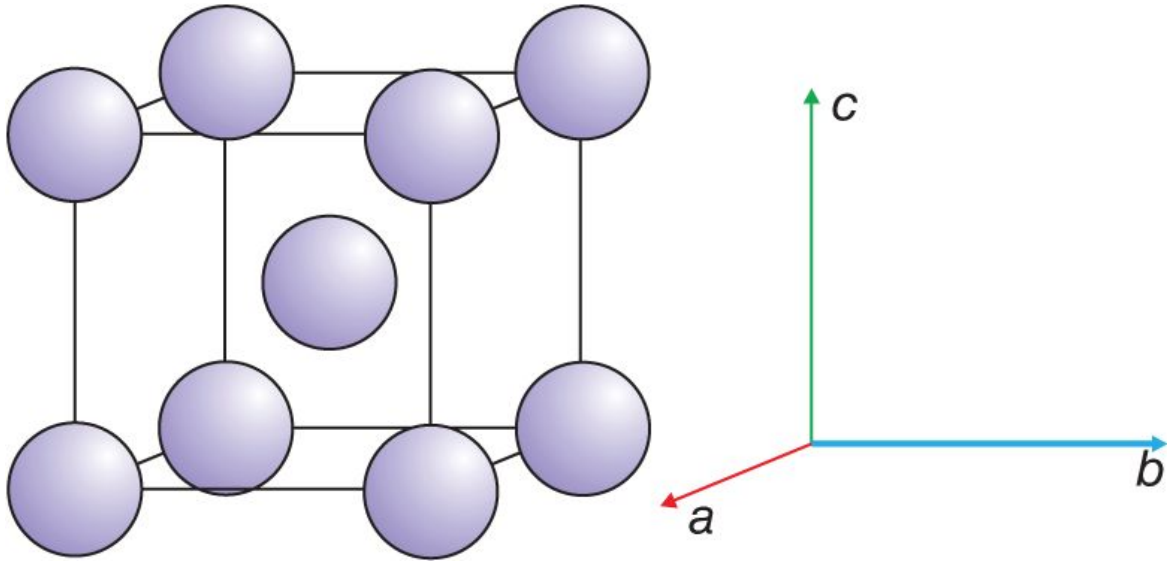


Figure 3.6 The body-centred cubic (bcc, A2) structure.

3.2.2.3 The Hexagonal Close-Packed (hcp, A3) Structure

General formula: M; example, magnesium Mg;

Lattice: primitive hexagonal, a = 0.321 nm, c = 0.521 nm.

Z = 2 Mg

Atom positions: 0, 0, 0; ⅓, ⅔, ½;

The lattice is primitive and hence there is only one lattice point in a unit cell. The motif is two atoms, one at (0, 0, 0) and the other at (⅓, ⅔, ½). If the atoms are supposed to be hard touching spheres, the fraction of the volume occupied is 0.7405 and the ratio c/a is equal to $\sqrt{8}/\sqrt{3} \approx 1.633$ (Figure 3.7).

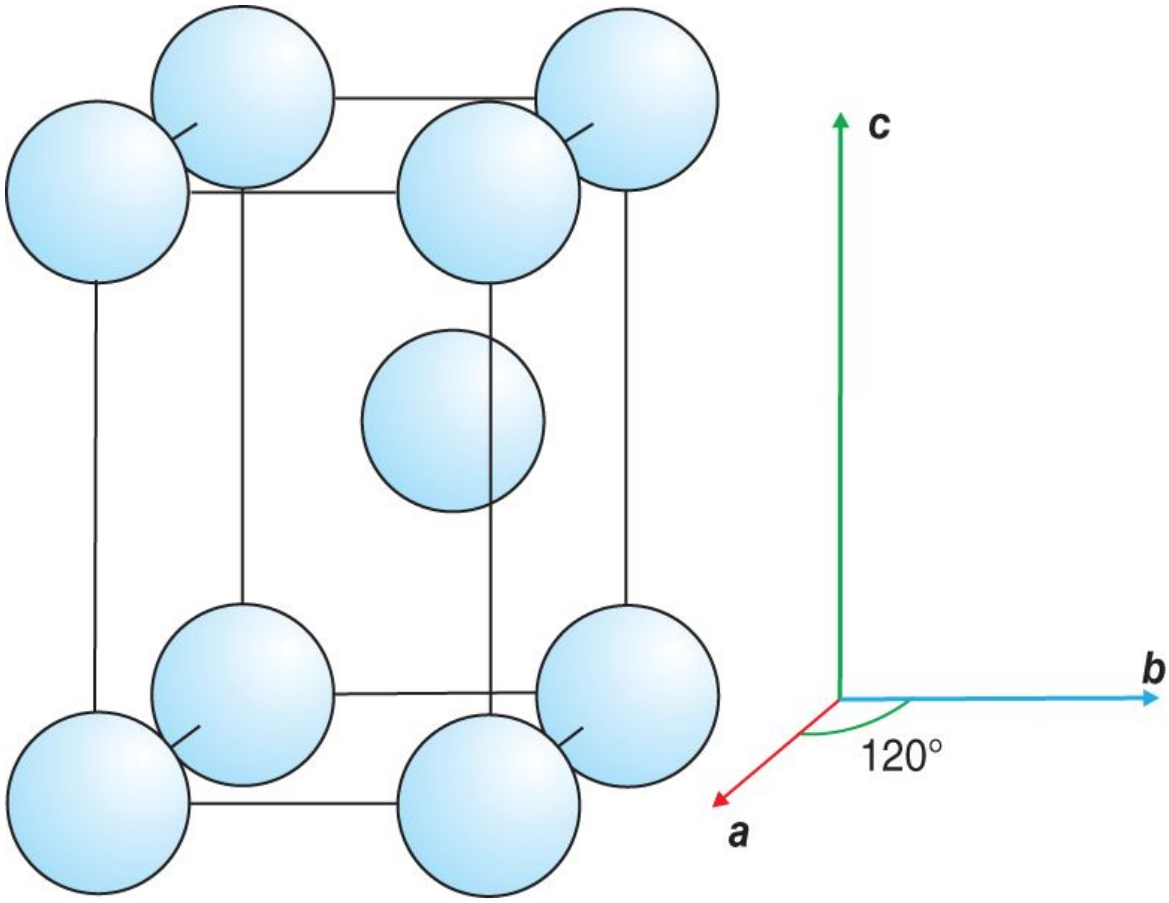


Figure 3.7 The hexagonal close-packed (hcp, A3) structure.

3.2.2.4 The Diamond Structure

General formula: M; example, diamond C;

Lattice: cubic face-centred, a = 0.356 nm.

Z = 8 C

Atom positions: 0, 0, 0; ¼, ¼, ¼; repeated in the face-centred pattern.

There are four lattice points in the face-centred unit cell, and the motif is two atoms, one atom at (0, 0, 0) and one at (¼, ¼, ¼). The structure is adopted by diamond, and in it, each carbon atom is bonded to four other carbon atoms that are arranged at the vertices of a tetrahedron (Figure 3.8). The bonds, of length 0.154 nm, are extremely strong sp^3-hybrids.

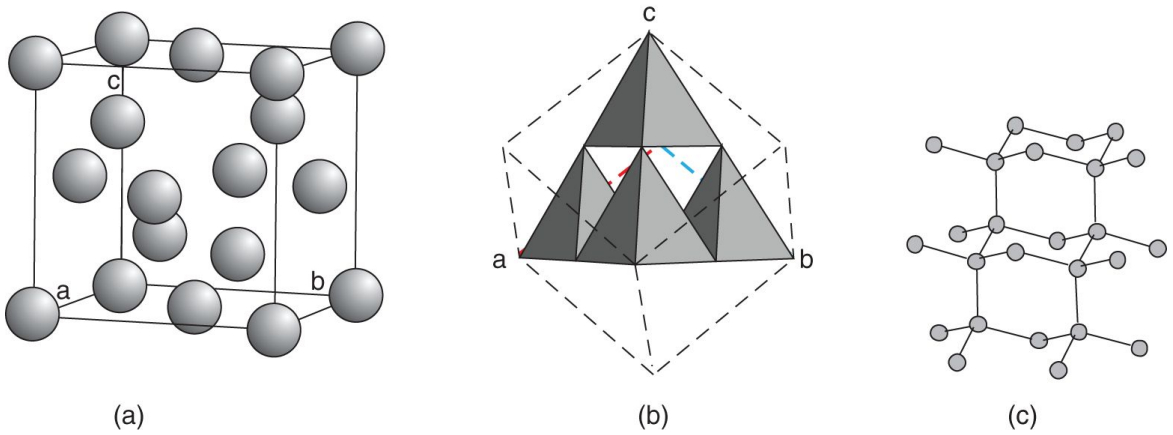


Figure 3.8 The structure of diamond: (a) unit cell; (b) cell viewed approximately along a cell diagonal; (c) tetrahedral bond geometry.

The elements silicon (a = 0.542 nm) and germanium (a = 0.564 nm) also have the same structure as diamond, as does grey tin (a = 0.649 nm) stable below a temperature of 13.2 °C.

3.2.2.5 The Graphite Structure

General formula: C; example, graphite C;

Lattice: primitive hexagonal, a = 0.246 nm, c = 0.671 nm.

Z = 4 C

Atom positions: 0, 0, 0; 0, 0, ½; ⅓, ⅔, 0; ⅔, ⅓, ½;

The lattice is primitive, giving one lattice point per unit cell. The motif is four atoms, at the positions specified earlier. The bonding in graphite, a form of elemental carbon, is closely related to that in benzene. The structure is made up of planar layers of carbon atoms bonded via sp^2 hybrid orbitals to give a strong sheet with a hexagonal geometry (Figure 3.9). Above and below the layers, π-bonds form a cloud of delocalised electrons that, because the layers are stacked directly on top of each other, repel each other strongly. This results in a large interlayer distance of 0.335 nm, compared with a C–C distance in the plane of 0.141 nm. The bonding between layers is very weak, made up of a van der Waals interaction between the delocalised electrons. Hence, although each layer of graphite is strong, the layers slide over one another easily. Graphite is easily cleaved in this direction and is a good dry lubricant. The delocalised electrons between the layers are similar to electrons in a metal and these make graphite into an electronic conductor parallel to the layers.

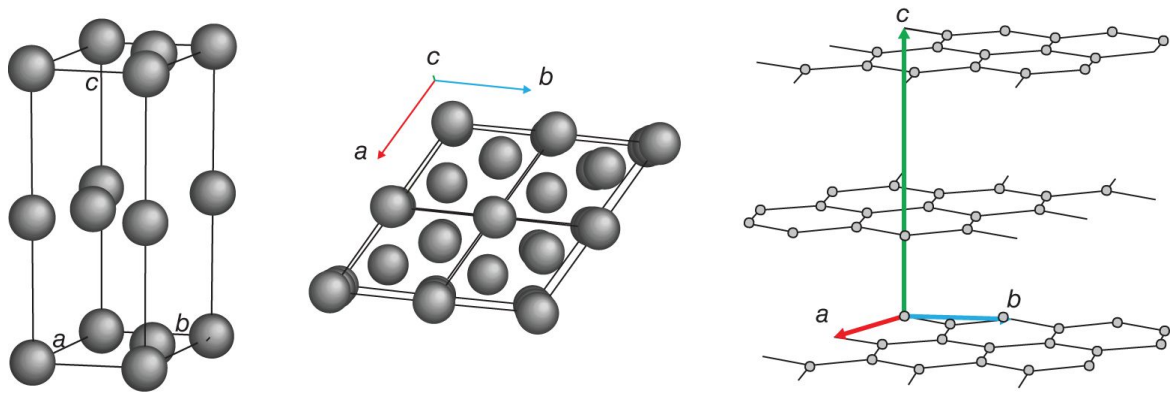


Figure 3.9 The structure of graphite: (a) unit cell; (b) projection down the **c**-axis; (c) the structure as hexagonal sheets.

Graphene (Figure 2.4) is a single graphite layer and carbon nanotubes (Figure 2.5) can be thought of as coiled up sheets of a single graphite layer.

3.2.2.6 The Halite (Rock Salt, Sodium Chloride) Structure

General formula: MX; example, NaCl;

Lattice: cubic face-centred, a = 0.563 nm.

Z = 4 NaCl

Atom positions: Na at 0, 0, 0; Cl at ½, ½, ½ repeated in the face-centred pattern.

There are four lattice points in the face-centred unit cell, and the motif is one Na atom at 0, 0, 0 and one Cl atom at ½, ½, ½. In this structure each ion is surrounded by six ions of the opposite type at the corners of a regular octahedron (Figure 3.10).

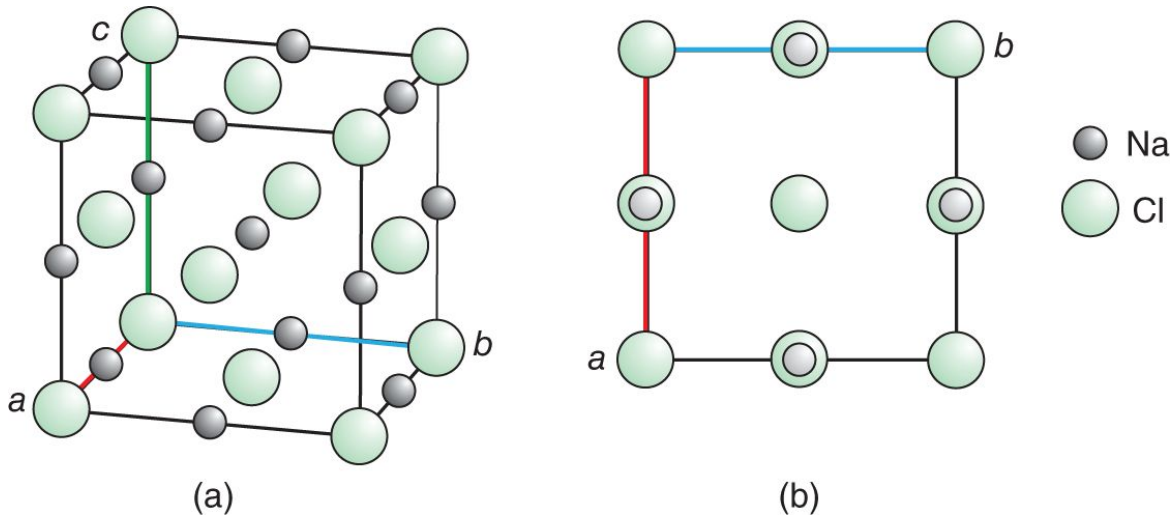


Figure 3.10 The halite (NaCl) structure: (a) perspective view; (b) projection down the **c**-axis.

This structure is adopted by many oxides, sulfides, halides, and nitrides with a formula MX.

3.2.2.7 The Perovskite Structure

General formula: ABX_3; example, $SrTiO_3$;

Lattice: cubic primitive, a = 0.3905 nm.

Z = 1 $SrTiO_3$

Atom positions: Sr: 0, 0, 0;

Ti: ½, ½, ½

O: ½, ½, 0; ½, 0, ½; 0, ½, ½;

The ideal perovskite structure is adopted by strontium titanate, $SrTiO_3$. There is one Ti atom at the cell centre and one Sr atom at the cell corners. The O atoms lie at the centres of the cell faces. Each Ti is surrounded by six O atoms arranged to give a regular octahedral coordination polyhedron (Figure 3.11).

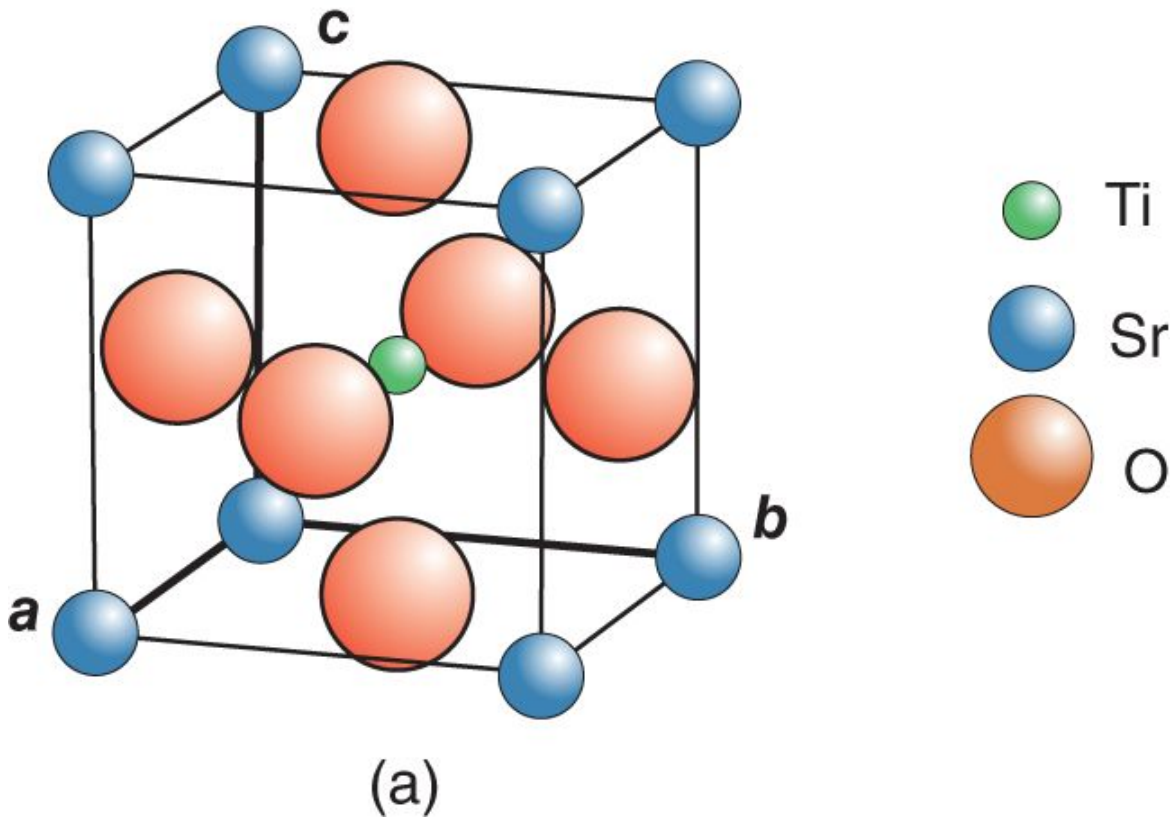


Figure 3.11 The cubic perovskite structure $SrTiO_3$.

The general formula of crystals with the perovskite structure is ABX_3, and all are structurally related to the mineral Perovskite, $CaTiO_3$. Most have a non-cubic unit cell and the structures, although topologically identical to $SrTiO_3$, usually have a lower symmetry, typically tetragonal or orthorhombic. The perovskite ABX_3 structure and slightly distorted variants of it are adopted by many compounds containing a large (A) and a medium sized (B) cation when combined with anions (X) including oxygen, halogens, and nitrogen.

3.2.2.8 The Spinel Structure

General formula: AB_2X_4; example, $MgAl_2O_4$;

Lattice: face-centred cubic, a = 0.809 nm.

Z = 8 $MgAl_2O_4$;

There are four lattice points in the face-centred unit cell, and the motif is two $MgAl_2O_4$ units. The oxygen atoms in the crystal structure are in the

same relative positions as the chlorine atoms in eight unit cells of halite, stacked together to form a 2 × 2 × 2 cube (Figure 3.12). In the cubic unit cell of spinel there are 32 oxygen atoms, to give a unit cell contents of $Mg_8Al_{16}O_{32}$. The magnesium atoms are surrounded by four oxygen atoms in the form of a tetrahedron and are said to occupy tetrahedral sites and the aluminium atoms are surrounded by six oxygen atoms in the form of an octahedron and are said to occupy octahedral positions or sites.

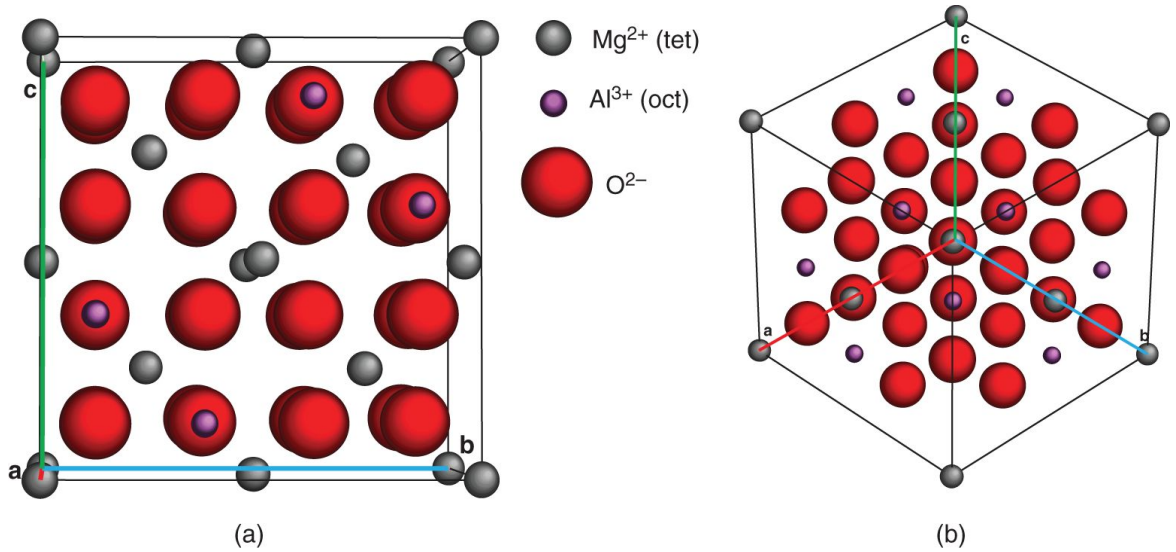


Figure 3.12 The $MgAl_2O_4$ normal spinel structure: (a) projected down the **c**-axis; (b) projected down a cell diagonal.

In each unit cell there are the same number of octahedral sites as there are oxygen ions, that is, 32, and twice as many tetrahedral sites as oxygen ions, that is 64. However, not all of these can be occupied. The A^{2+} and B^{3+} cations are inserted into this array in an ordered fashion, filling ½ of the available octahedral positions and ⅛ of the available tetrahedral positions. This means that there are 8 occupied tetrahedral sites and 16 occupied octahedral sites in a unit cell.

There are many compounds with this structure, of general formula AB_2X_4, collectively known as spinels, including halides, sulfides, and nitrides as well as oxides. Across this large group, there are two principle arrangements of cations found. If the 8 A^{2+} ions per unit cell are confined to the available tetrahedral sites, these are filled completely. The 16 B^{3+} ions are then confined to the octahedral sites. This cation distribution is often depicted as $(A)[B_2]X_4$, with the tetrahedral cations enclosed in ( ) and the octahedral cations enclosed in [ ]. This is called the *normal* spinel structure, and spinels with this arrangement of cations are said to be

normal spinels. If the 8 A^{2+} ions are placed in half of the available 16 octahedral sites, half of the B^{3+} ions must be placed in the remaining octahedral sites and the other half in the tetrahedral sites. This can be written as $(B)[AB]X_4$. This arrangement is called the *inverse* spinel structure and compounds with this cation arrangement are said to be inverse spinels.

In reality, very few spinels have exactly the normal or inverse structure, and these are sometimes called *mixed* spinels. The cation distribution between the two sites is a function of a number of parameters, including temperature. This variability is described by an *occupation factor* λ, which gives the fraction of B^{3+} cations in tetrahedral positions:

$$\lambda = B^{3+}_{tet}/B^{3+}_{total}$$

A normal spinel is characterised by a λ value of 0, and an inverse spinel by a value of 0.5. The spinel $MgAl_2O_4$ has a λ value of 0.05, and so is quite a good approximation to a normal spinel.

3.2.2.9 Lattice Parameters and Vegard's Law

The unit cell lattice parameters of a crystalline phase with a fixed composition, at a given temperature, are constant. If the solid has a composition range, as in a solid solution or an alloy, the cell parameters will vary. *Vegard's law*, first propounded in 1921, states that the lattice parameter of a solid solution of two phases with similar structures will be a *linear function* of the lattice parameters of the two end members of the composition range (Figure 3.13):

$$x = (a_{ss} - a_1)/(a_2 - a_1)$$

where a_1 and a_2 are the lattice parameters of the parent phases, a_{ss} is the lattice parameter of the solid solution or alloy, and x is the mole fraction of the parent phase with lattice parameter a_2. This 'law' is simply an expression of the idea that the cell parameters are a direct consequence of the sizes of the component atoms in the solid solution. Vegard's law, in its ideal form, is almost never obeyed exactly. A plot of cell parameters that lies below the ideal line is said to show a *negative* deviation from Vegard's law, and a plot that lies above the ideal line is said to show a *positive* deviation. In all cases, a plot of composition vs. cell parameters can be used to determine the composition of intermediate compositions in a solid solution.

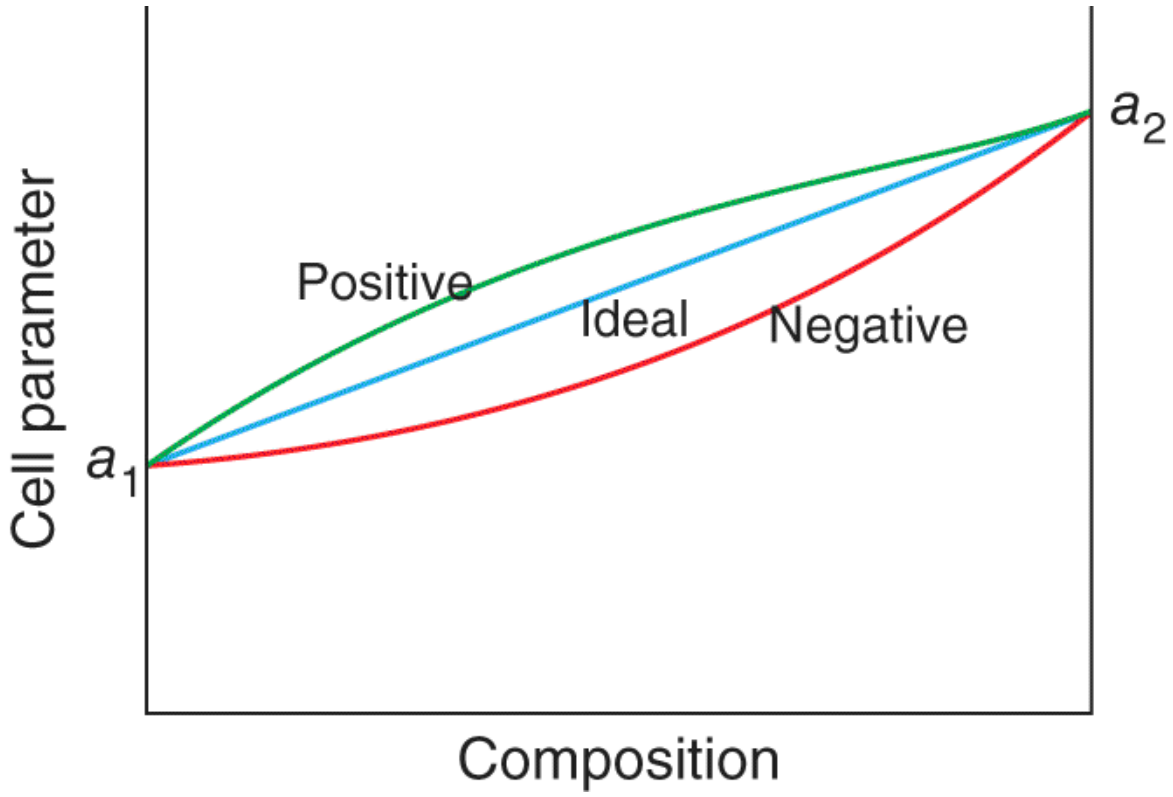


Figure 3.13 Vegard's law, relating unit cell lattice parameters to composition.

3.3 Crystal Planes and Directions

3.3.1 Miller Indices

The facets of a well-formed crystal or internal planes through a crystal structure or a lattice are specified in terms of *Miller Indices*. These indices, h, k, and l, written in round brackets (hkl), represent not just one plane, but the *set of all parallel planes* (hkl). The values of h, k, and l are related to the fractions of a unit cell edge, a, b, and c, respectively, intersected by this set of planes.

A plane that lies parallel to a cell edge, and so never cuts it, is given the index 0 (zero). A plane that passes across the end of the unit cell cutting the **a**-axis and parallel to the **b**- and **c**-axes of the unit cell has Miller indices (100) (Figure 3.14a). The indices indicate that the plane cuts the cell edge running along the **a**-axis at a position 1a and does not cut the cell edges parallel to the **b**- or **c**-axes at all. Remember that (100) represents a set of identical planes all separated by a, not just one plane. A plane parallel to this that cuts the **a**-cell edge in half, at $a/2$, has indices (200) (Figure 3.14b). Similarly, parallel planes cutting the **a**-edge at $a/3$ would have Miller indices of (300) (Figure 3.14c). Any *general plane* parallel to (100) is written as (h00). A general plane parallel to the **a**- and **c**-axes, perpendicular to the **b**-axis and so only cutting b has indices (0k0), and a general plane parallel to the **a**- and **b**-axes, and perpendicular to the **c**-axis, and so cutting c cell-edge has indices (00l) (Figure 3.14d,e).

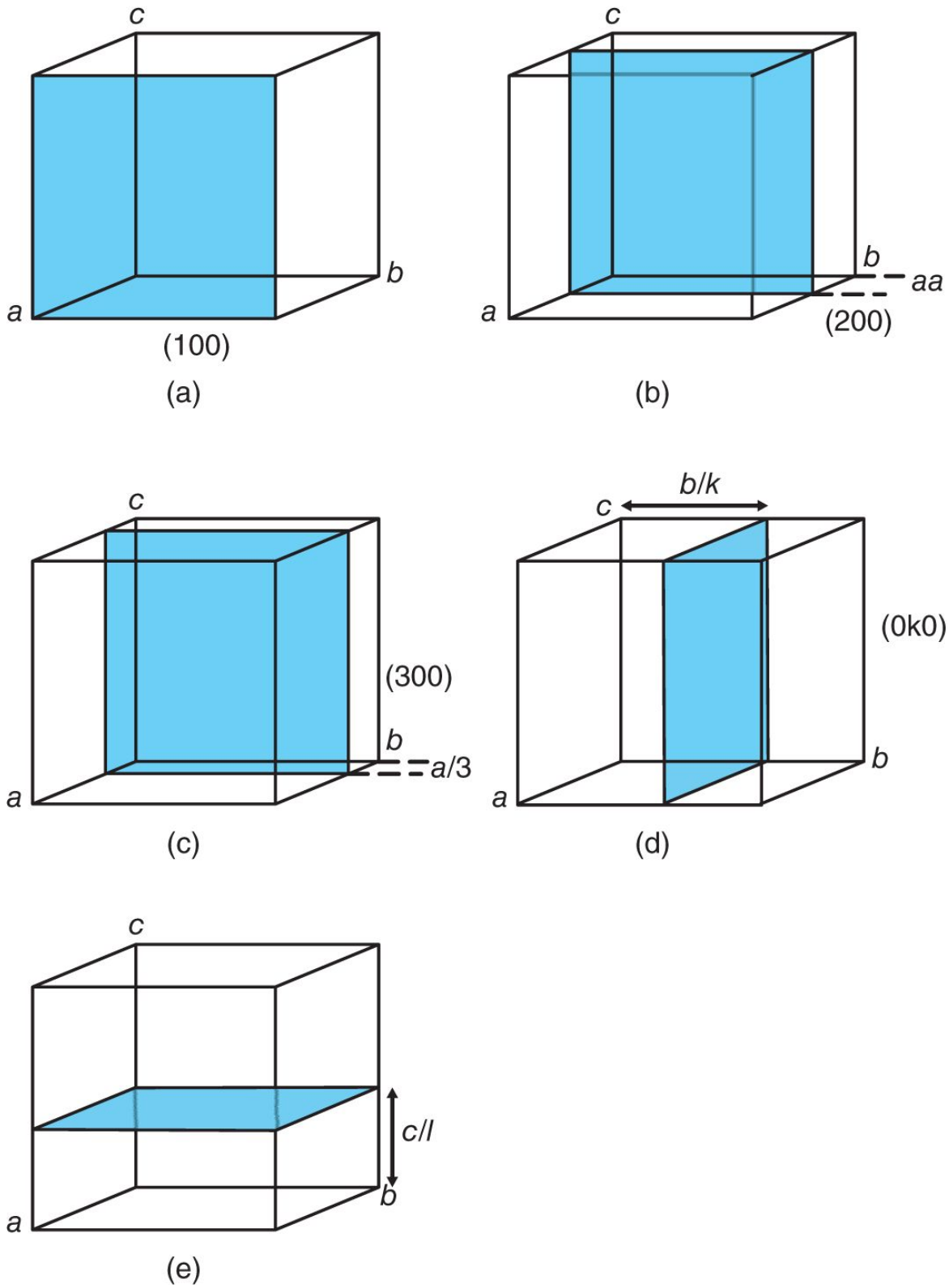


Figure 3.14 Miller indices of crystal planes: (a) (100); (b) (200); (c) (300); (d) (0*k*0); (e) (00*l*).

Planes that cut two edges and lie parallel to a third are described by indices (*hk*0), (0*kl*), or (*h*0*l*) (Figure 3.15). Negative intersections are written with a negative sign over the index, and pronounced *h* bar, *k* bar, and *l* bar. For example, there are four planes related to (110); a similar plane cuts the **b**-axis in 1*b* but the **a**-axis is cut in a negative direction, at −*a* and so has Miller indices ($\bar{1}$10), pronounced (one bar, one, zero) (Figure 3.16a). Two other related planes, one of which cuts the **b**-axis at −*b*, and so has Miller indices (1$\bar{1}$0), pronounced (one, one bar, zero) and the other, with Miller indices ($\bar{1}\bar{1}$0), can be drawn (Figure 3.16b). Because the Miller indices (*hkl*) refer to a set of planes, ($\bar{1}\bar{1}$0) is equivalent to (110) and (1$\bar{1}$0) is equivalent to ($\bar{1}$10). This notation is readily extended to cases where a plane cuts all three unit cell edges (Figure 3.17).

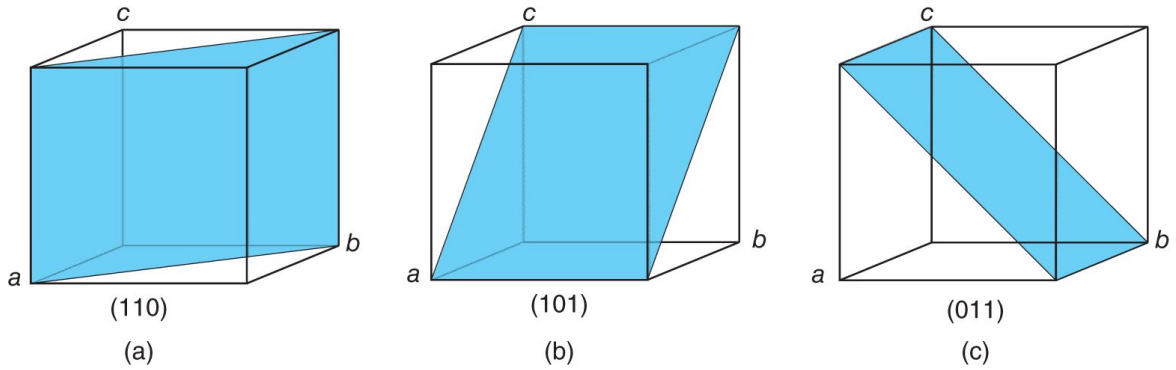


Figure 3.15 Miller indices of crystals planes: (a) (110); (b) (101); (c) (011).

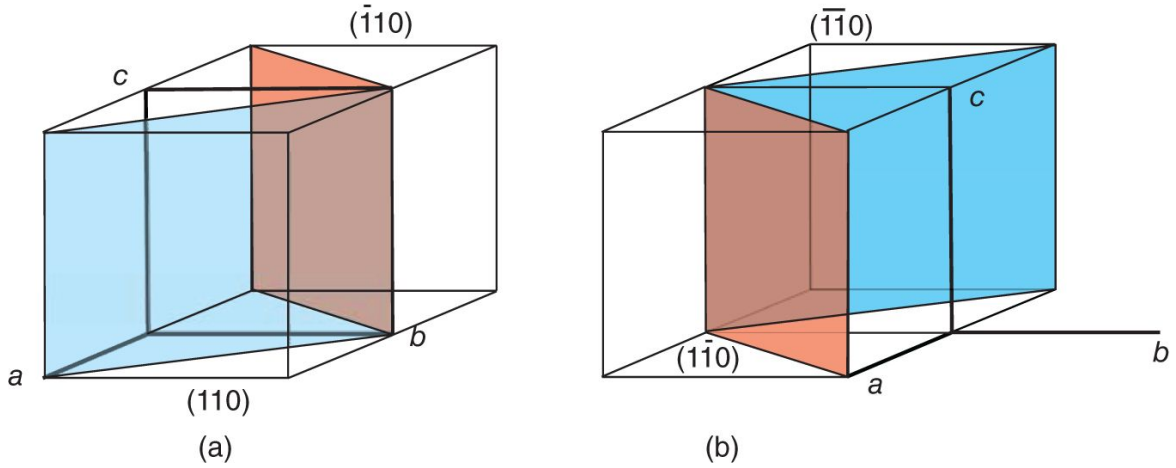


Figure 3.16 Miller indices of crystal planes: (a) (110) and ($\bar{1}$10); (b) (1$\bar{1}$0) and ($\bar{1}\bar{1}$0).

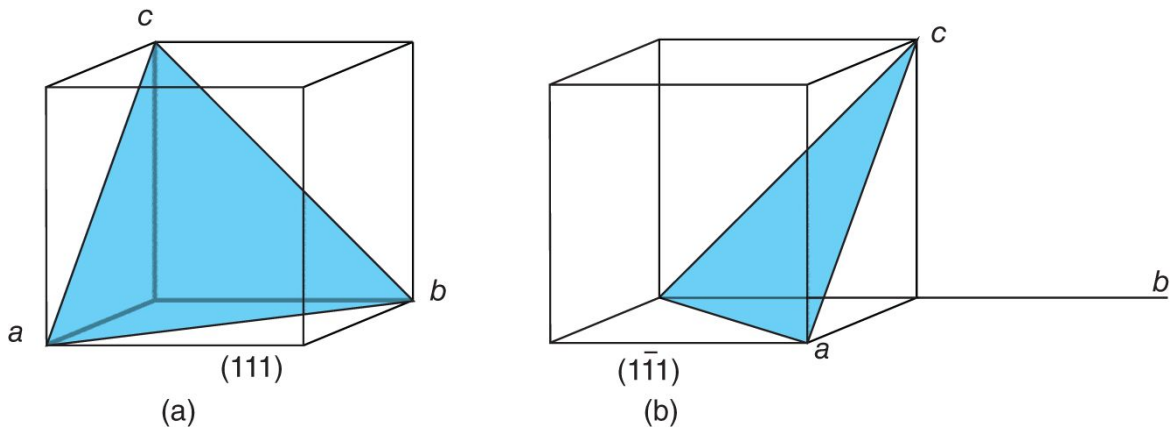


Figure 3.17 Miller indices of crystal planes: (a) (111); (b) (1 $\bar{1}$ 1).

In order to find the Miller indices of a plane, first draw the other members of the set, if not already present (Figure 3.18a,c). This is done by drawing parallel planes through each lattice point (Figure 3.18b,d). Move along the unit cell and count the number of spaces between planes crossed in moving a distance of one cell edge, *a*. If the spaces lie in the +**a** direction, this gives the index *h*. If the spaces lie in the −**a** direction, this gives index $\bar{h}$. Repeat this along **b** to get *k* and along **c** to get *l*.

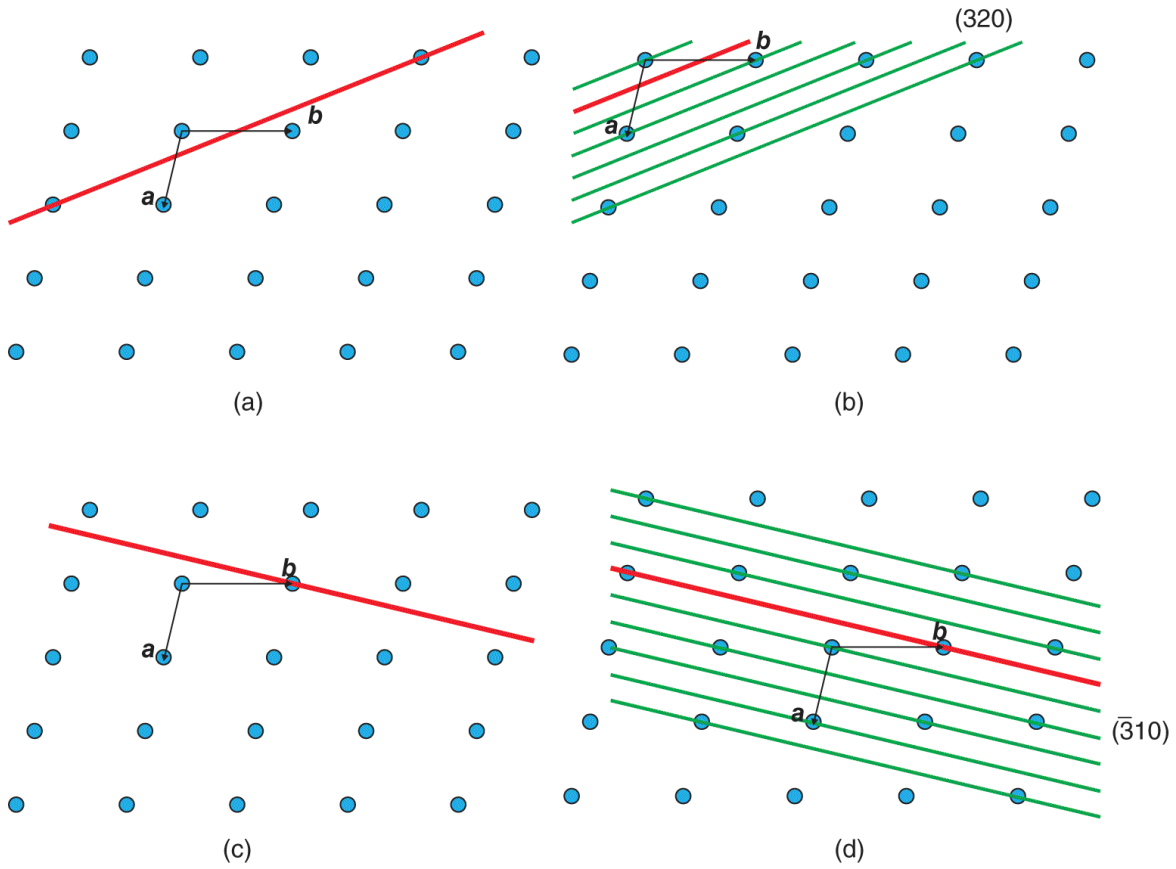


Figure 3.18 Determination of Miller indices: (a), (b), (320); (c), (d), (3 $\bar{1}$ 0).

In crystals of high symmetry, there are often several sets of (*hkl*) planes that are identical. For example, in a cubic crystal, the (100), (010), and (001) planes are identical in every way. Similarly, in a tetragonal crystal, (110) and ($\bar{1}$10) planes are identical. Sets of identical planes are designated {*hkl*}. Thus, in the cubic system, the symbol {100} represents the three sets of planes (100), (010), and (001), {110} represents the six sets of planes (110), (101), (011), ($\bar{1}$10), ($\bar{1}$01), and (0 $\bar{1}$ 1), and {111} represents the four sets (111), (11 $\bar{1}$), (1 $\bar{1}$ 1), and ($\bar{1}$11).

3.3.2 Hexagonal Crystals and Miller–Bravais Indices

The Miller indices of planes parallel to the **c**-axis in crystals with a hexagonal unit cell, such as magnesium, can be ambiguous. Figure 3.19 shows three such sets of planes, imagined to be perpendicular to the plane of the figure and parallel to **c**. Following the procedure just outlined, the sets have Miller indices A (110), B (1 $\bar{2}$ 0), and C ($\bar{2}$10). Although these seem to refer to different types of plane, clearly they are identical from the

point of view of atomic constitution. In order to eliminate this confusion, four indices (*hkil*), are often used to specify planes in a hexagonal crystal. These are called *Miller–Bravais indices* and are only used in the hexagonal system. The index *i* is given by:

$$h + k + i = 0, \text{ that is } i = -(h + k)$$

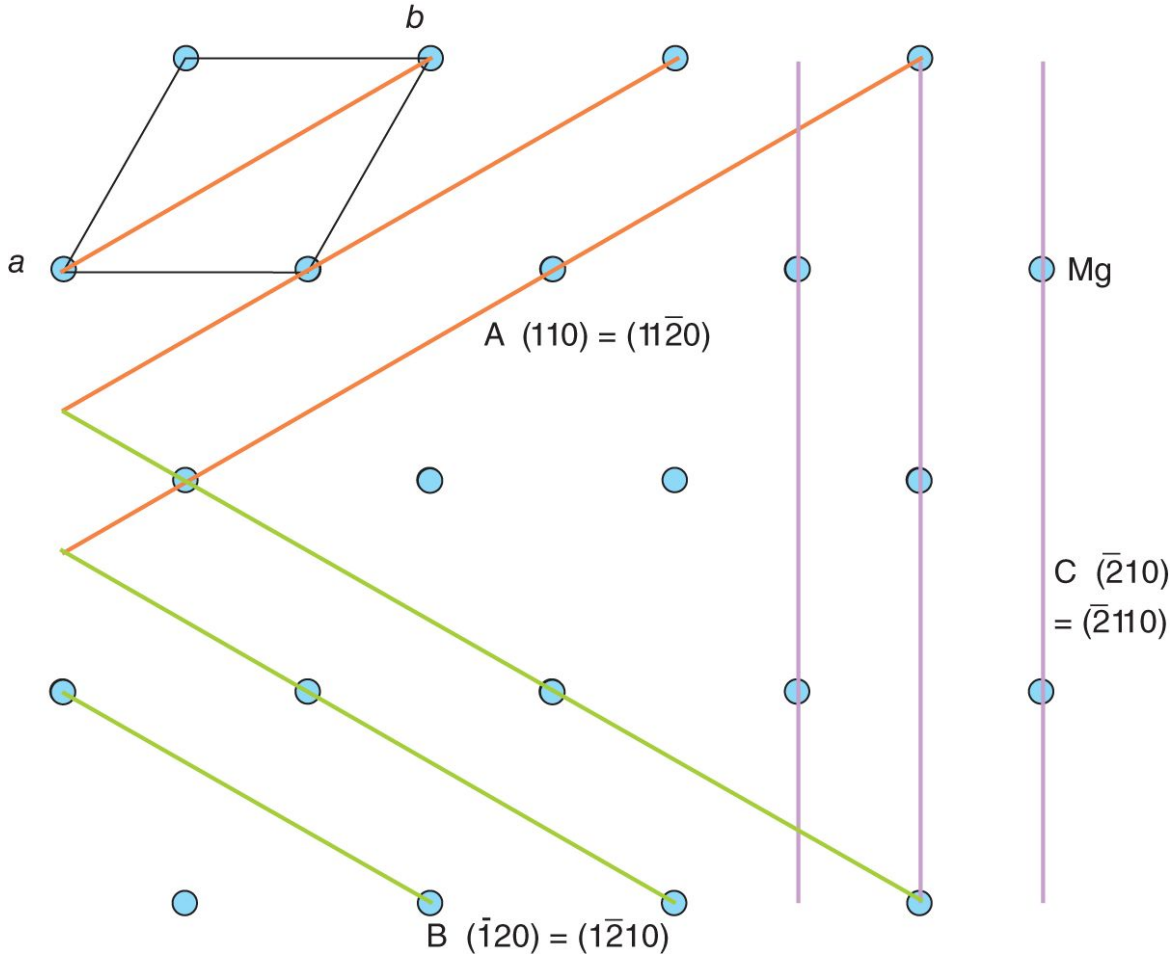
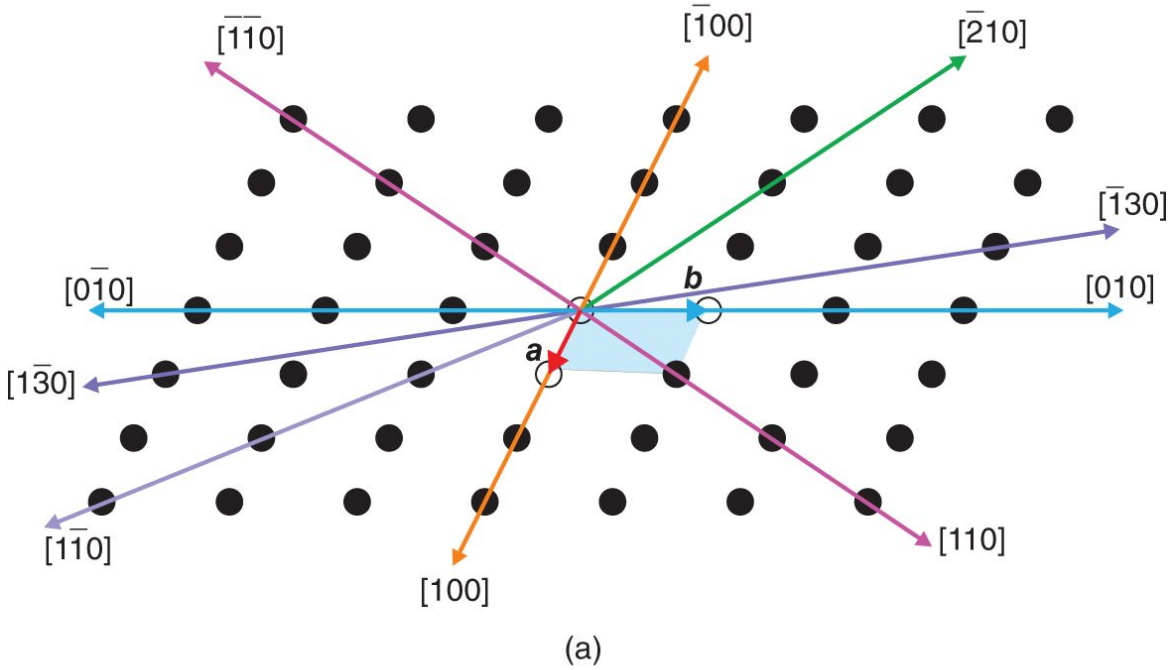


Figure 3.19 Miller indices in hexagonal crystals.

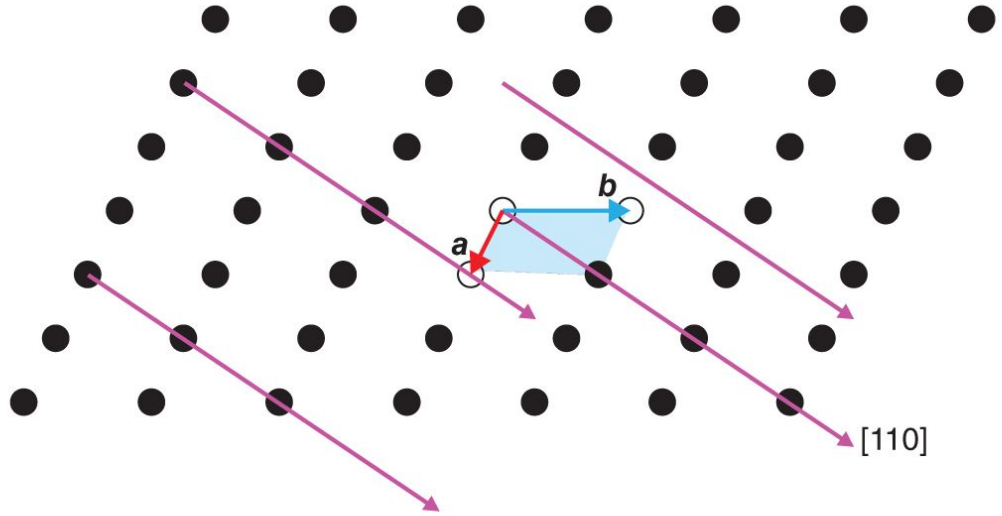
In reality this third index is not needed. However, it does help to bring out the relationship between the planes. Using four indices, the planes are A (11$\bar{2}$0), B (1$\bar{2}$10), and C ($\bar{2}$110). Because it is a redundant index, the value of *i* is sometimes replaced by a dot, to give indices (*h k . l*). This nomenclature emphasises that the hexagonal system is under discussion without actually including a value for *i*.

3.3.3 Directions

It is important to be able to specify directions in crystals in an unambiguous fashion. Directions are written generally as [*uvw*] where the three indices *u*, *v*, and *w* define the coordinates of a point with respect to the crystallographic **a**-, **b**-, and **c**-axes. The index *u* gives the coordinates in terms of *a* along the **a**-axis, the index *v* gives the coordinates in terms of *b* along the **b**-axis, and the index *w* gives the coordinates in terms of *c* along the **c**-axis. The direction [*uvw*] is simply the vector pointing from the origin to the point with coordinates *u*, *v*, *w* (Figure 3.20a). For example, the direction [100] is parallel to **a**, [010] is parallel to **b**, and [001] is parallel to **c**. Because directions are vectors, [*uvw*] is not identical to [$\overline{u}\overline{v}\overline{w}$], in the same way that the direction 'North' is not the same as the direction 'South'. Remember, though, that [*uvw*] means the set of all parallel directions or vectors because the origin of the coordinate system is not fixed and can always be moved to the starting point of the vector (Figure 3.20b). A North wind is always a North wind, regardless of where you stand.



(a)



(b)

Figure 3.20 (a) Directions in a lattice; (b) parallel directions have the same indices.

As with the Miller indices, it is sometimes convenient to group together all directions that are identical by virtue of the symmetry of the structure. These are represented by the notation ⟨*uvw*⟩. The symbol ⟨100⟩ represents the six directions [100], [$\bar{1}$00], [010], [0$\bar{1}$0], [001], [00$\bar{1}$] for a cubic crystal

A *zone* is a set of planes, all of which are parallel to a single direction, called the *zone axis*. The zone axis [*uvw*] is perpendicular to the plane (*uvw*) in cubic crystals but *not* in crystals of other symmetry.

It is sometimes important to specify a vector with a definite length; perhaps to indicate the displacement of one part of a crystal with respect to another part, such as the relative displacement of one side of a boundary with respect to the other. In such a case, the direction of the vector is written as earlier, and a prefix is added to give the length. The prefix is usually expressed in terms of the unit cell dimensions. For example, in a cubic crystal, a displacement of two unit cell lengths parallel to the **a**-axis would be written 2*a*[100].

As with Miller indices, directions in hexagonal crystals are sometimes specified by a four-index system [*u′v′tw′*], called *Weber indices*. The conversion of a three-index set to a four-index set is given by:

$$[uvw] \rightarrow [u'v'tw']$$

$$u' = n/3(2u - v)$$

$$v' = n/3(2v - u)$$

$$t = -(u' + v')$$

$$w' = nw$$

In these equations, *n* is a factor *sometimes* needed to make the new indices into smallest integers. Thus directions [00*l*] always transform to [000*l*]. The three equivalent directions in the basal (0001) plane of a hexagonal crystal structure such as magnesium (Figure 3.21) are obtained by using the above transformations. The correspondence is:

$$[100] = [2\bar{1}\bar{1}0]$$

$$[010] = [\bar{1}2\bar{1}0]$$

$$[\bar{1}\bar{1}0] = [\bar{1}\bar{1}20]$$

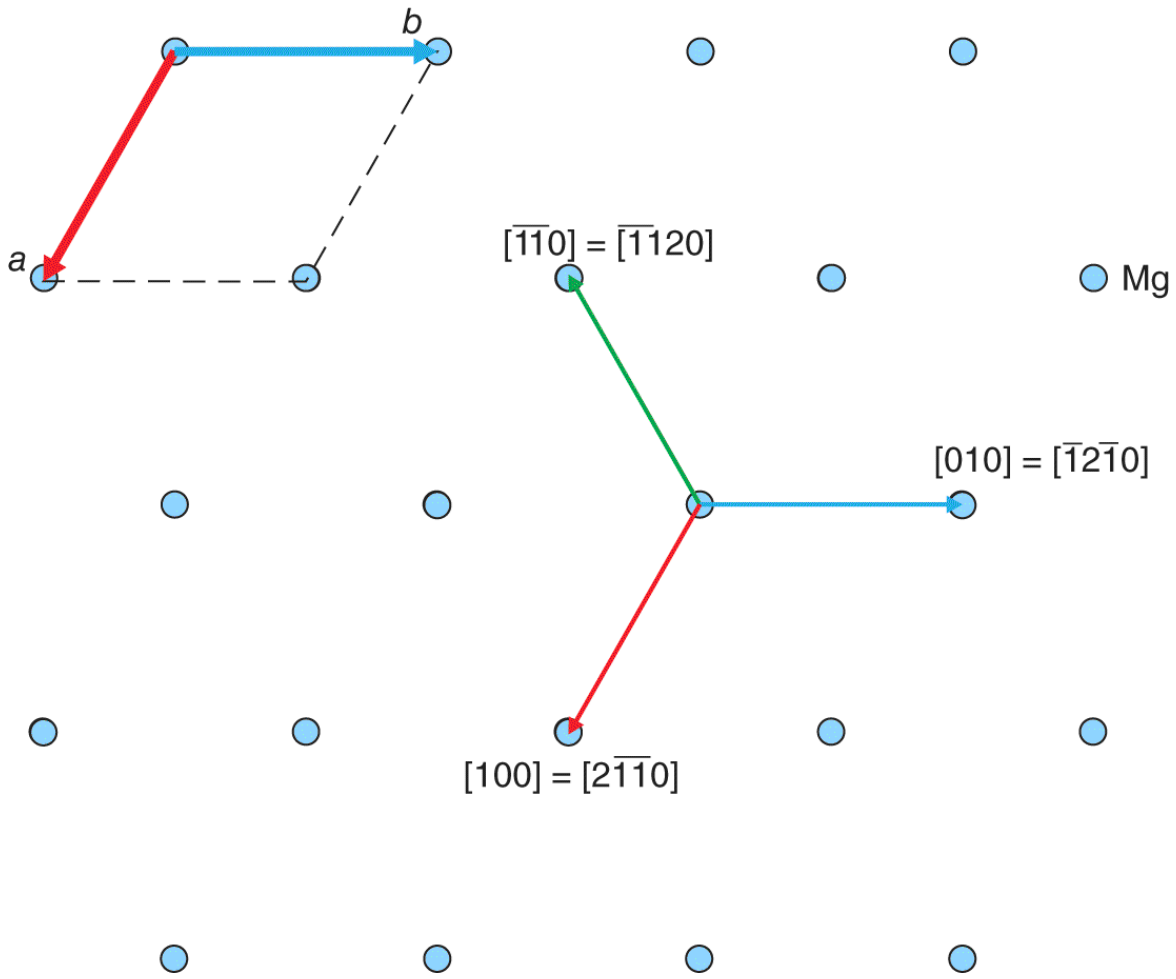


Figure 3.21 Directions in the (001) plane of a hexagonal crystal.

The relationship between directions and planes depends upon the symmetry of the crystal. In cubic crystals (and *only* cubic crystals), the direction [*hkl*] is normal to the plane (*hkl*).

3.3.4 Interplanar Spacings

The *interplanar spacing*, d_{hkl}, the spacing between (*hkl*) planes in the crystal or lattice and the unit cell volume in terms of the lattice parameters, is given in Table 3.2.

Table 3.2 Interplanar spacing and unit cell volume.

System	Interplanar spacing, d_{hkl}	Unit cell volume
Cubic	$1/d^2_{hkl} = [h^2 + k^2 + l^2]/a^2$	a^3
Tetragonal	$1/d^2_{hkl} = [(h^2 + k^2)/a^2] + [l^2/c^2]$	a^2c
Orthorhombic	$1/d^2_{hkl} = [h^2/a^2] + [k^2/b^2] + [l^2/c^2]$	abc
Monoclinic	$1/d^2_{hkl} = [h^2/a^2\sin^2\beta] + [k^2/b^2] + [l^2/c^2\sin^2\beta] - [(2hl\cos\beta)/(ac\sin^2\beta)]$	$abc \sin\beta$
Triclinic[a]	$1/d^2_{hkl} = [1/V^2]\{[S_{11}h^2] + [S_{22}k^2] + [S_{33}l^2] + [2S_{12}hk] + [2S_{23}kl] + [2S_{13}hl]\}$	$abc\sqrt{(1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma)}$
Hexagonal	$1/d^2_{hkl} = [4/3][(h^2 + hk + k^2)/a^2] + [k^2/b^2] + [l^2/c^2)]$	$[\sqrt{(3)}/2][a^2c] \approx 0.866\, a^2c$
Rhombohedral	$1/d^2_{hkl} = \{[(h^2 + k^2 + l^2)\sin^2\alpha + 2(hk + kl + hl)(\cos^2\alpha - \cos\alpha]/[a^2(1 - 3\cos^2\alpha + 2\cos^3\alpha)]\}$	$a^3\sqrt{(1 - 3\cos^2\alpha + 2\cos^3\alpha)}$

[a] $S_{11} = b^2c^2\sin^2\alpha$; $S_{22} = a^2c^2\sin^2\beta$; $S_{33} = a^2b^2\sin^2\gamma$;

$$S_{12} = abc^2(\cos\alpha\cos\beta - \cos\gamma);\ S_{23} = a^2bc(\cos\beta\cos\gamma - \cos\alpha);$$

$$S_{13} = ab^2c(\cos\gamma\cos\alpha - \cos\beta);$$

V = unit cell volume.

3.4 Crystal Density

3.4.1 Density Estimation

The theoretical density of a crystal can be found by calculating the mass of all the atoms in the unit cell. The mass of an atom, m_A, is its molar mass divided by the Avogadro constant, N_A:

$$m_A = \text{molar mass}/N_A$$

The total mass of all of the atoms in the unit cell is then:

$$n_1\, m_1 + n_2\, m_2 + n_3\, m_3 \ldots /\mathrm{N_A}$$

where n_1 is the number of atoms of type 1, with a molar mass of m_1, and so on. This is written in a more compact form as:

$$\sum_{i=1}^{q} n_i m_i / N_A$$

where there are q different atom types in the unit cell. The density, ρ, is simply the total mass divided by the unit cell volume, V:

$$\rho = \left(\sum_{i=1}^{q} n_i m_i / N_A \right) / V$$

To count the number of atoms in a unit cell, use the information:

- An atom within the cell counts as 1
- An atom in a face counts as 1/2
- An atom on an edge counts as 1/4
- An atom on a corner counts as 1/8

A quick method to count the number of atoms in a unit cell is to displace the unit cell outline to remove all atoms from corners, edges, and faces. The atoms remaining, which represent the unit cell contents, are all within the boundary of the unit cell and count as 1.

3.4.2 The Density of NaCl

As an example, the density of NaCl is determined in the following way. The unit cell of the NaCl structure contains four sodium (Na) and four chlorine (Cl) atoms and has a cubic unit cell with a = 0.5500 nm. The mass of the unit cell, m, is then given by:

$$\begin{aligned} m &= [(4 \times 22.99) \\ &\quad + (4 \times 35.453)]/1000 \times N_A \\ &= 3.882 \times 10^{-25}\ \text{kg} \end{aligned}$$

where 22.99 g mol^{-1} is the molar mass of sodium, 35.453 g mol^{-1} is the molar mass of chlorine, and N_A is the Avogadro constant, 6.022 14 × 10^{23} mol^{-1} and the factor 1000 is to convert grams to kilograms.

The volume, V, of the cubic unit cell is given by a^3, thus:

$$\begin{aligned} V &= (0.5500 \times 10^{-9})^3\ \text{m}^3 \\ &= 1.663\,75 \times 10^{-28}\ \text{m}^3. \end{aligned}$$

The density, ρ, is given by the mass m divided by the volume, V:

$$\begin{aligned} \rho &= 3.882 \times 10^{-25}\ \text{kg}/1.66375 \times 10^{-28}\ \text{m}^3 \\ &= 2333\ \text{kg m}^{-3} \end{aligned}$$

The measured density is 2165 kg m^{-3}. The theoretical density is almost always slightly greater than the measured density because real crystals contain defects that act so as to reduce the total mass per unit volume.

3.4.3 The Density of Crystals with a Variable Composition

For a solid that has a variable composition, such as an alloy or a phase that exists over a composition range, the density will vary with composition. In addition, it is found that the unit cell dimensions change in a regular way across the phase range. These two techniques can be used in conjunction with each other to determine the defects that may be responsible for the composition change. (Note that as both these techniques are averaging

techniques: they say nothing about the real structure or organisation of the defects, but they do suggest first approximations.)

The general procedure is to determine the unit cell dimensions, the crystal structure type and the real composition of the material. The ideal composition of the unit cell will be known from the structure type. The ideal composition is adjusted by the addition of extra atoms (interstitials or substituted atoms) or removal of atoms (vacancies) to agree with the real composition. A calculation of the density of the sample assuming either that interstitials or vacancies are present is then made. This is compared with the measured density to discriminate between the two alternatives.

The method can be illustrated by reference to iron monoxide. Iron monoxide, often known by its mineral name of wüstite, has the halite (NaCl) structure. In the normal halite structure, there are four metal and four non-metal atoms in the unit cell, and compounds with this structure have an ideal composition $MX_{1.0}$. Wüstite has a composition that is always oxygen rich compared with the ideal formula of $FeO_{1.0}$. Data[4] for one sample found an oxygen:iron ratio of 1.058, a density of 5728 kg m^{-3}, and a cubic lattice parameter of 0.4301 nm. The real composition can be obtained by assuming either that there are extra oxygen atoms in the unit cell or that there are iron vacancies present.

Model A: Assume that the iron atoms in the crystal are in a perfect array, identical to the metal atoms in halite and an excess of oxygen is due to extra oxygen atoms over and above those on the normal anion positions. Thus, in this model, the unit cell must contain 4 atoms of Fe and $4(1+x)$ atoms of oxygen. The unit cell contents are Fe_4O_{4+4x} and the (measured) composition is $FeO_{1.058}$.

The mass of one unit cell is m_A,

$$\begin{aligned} m_A &= [(4 \times 55.85) + (4 \times 16 \times (1+x))]/N_A \\ &= [(4 \times 55.85) + (4 \times 16 \times 1.058)]/N_A \text{ g} \\ &= 4.834 \times 10^{-25} \text{ kg} \end{aligned}$$

The volume, V, of the cubic unit cell is given by a^3, thus:

$$V = (0.4301 \times 10^{-9})^3 \text{ m}^3 = 7.9562 \times 10^{-29} \text{ m}^3.$$

The density, ρ, is given by the mass m_A divided by the volume, V:

$$\begin{aligned}\rho &= 4.834 \times 10^{-25}/7.9562 \times 10^{-29} \\ &= 6076\ \text{kg m}^{-3}\end{aligned}$$

Model B: Assume that the oxygen array is perfect and identical to the non-metal atom array in the halite structure. As there are more oxygen atoms than iron atoms, the unit cell must contain some vacancies on the iron positions. In this case, one unit cell will contain 4 atoms of oxygen and $(4 - 4x)$ atoms of iron. The unit cell contents are $Fe_{4-4x}O_4$ and the (measured) composition is $Fe_{1/1.058}O_{1.0}$ or $Fe_{0.945}O$.

The mass of one unit cell is m_B:

$$\begin{aligned}m_B &= [(4 \times (1 - x) \times 55.85) + (4 \times 16)]/N_A \\ &= [(4 \times 0.945 \times 55.85) + (4 \times 16)]/N_A\ \text{g} \\ &= 4.568 \times 10^{-25}\ \text{kg}\end{aligned}$$

The density, ρ, is given by m_B divided by the volume, V, to yield

$$\begin{aligned}\rho &= 4.568 \times 10^{-25}/(7.9562 \times 10^{-29}) \\ &= 5741\ \text{kg m}^{-3}\end{aligned}$$

The difference in the two values is surprisingly large. The experimental value of the density, 5728 kg m^{-3}, is in good accord with Model B, which assumes vacancies on the iron positions. This indicates that the formula should be written $Fe_{0.945}O$.

3.5 Structural Relationships

A list of atomic positions is often not very helpful when a variety of structures have to be compared. In this section, two ways of looking at structures that facilitate comparisons are described. In the first of these, structures are described as built up by packing together spheres, and in the other, in terms of polyhedra linked by corners and edges.

3.5.1 Sphere Packing

The structure of many crystals can conveniently be described in terms of an ordered packing of spheres, representing spherical atoms or ions. Although there are many ways of doing this, just two main arrangements are sufficient to describe a large number of structures. Both are built of layers, each of which consists of a hexagonal arrangement of spheres just touching each other to fill the space as much as possible (Figure 3.22a). In the first arrangement, called *hexagonal close-packing* (*hcp*), a second layer fits into the dimples in the first layer, and the third layer is stacked in dimples on top of the second layer to lie over the first layer (Figure 3.22b). This sequence is repeated indefinitely. If the position of the spheres in the first layer is labelled A, and the positions of the spheres in the second, B, the complete stacking is described by the sequence ...ABABAB...

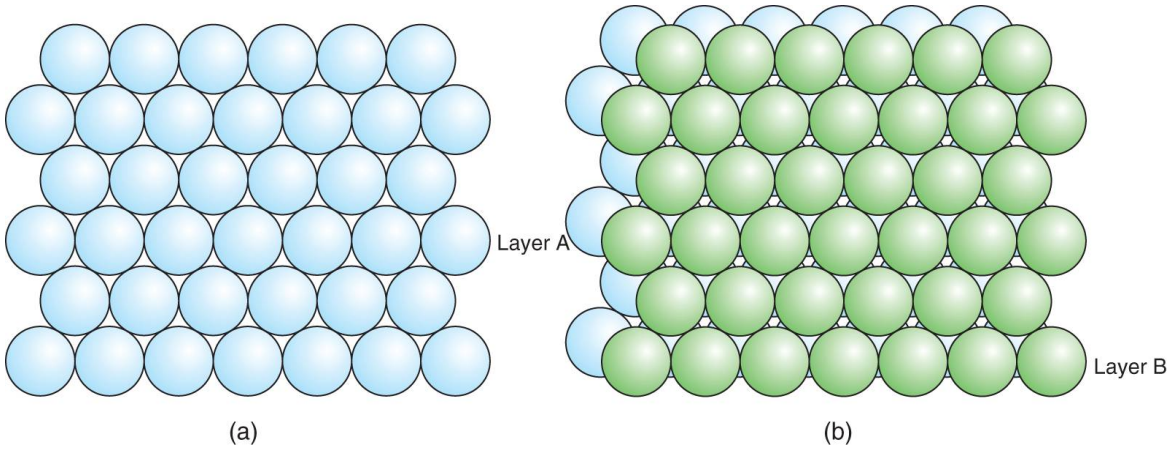


Figure 3.22 Hexagonal close-packing (hcp) of spheres: (a) single layer; (b) two layers making up the repeat ...AB....

The structure has a *hexagonal* symmetry and unit cell. The **a**- and **b**-axes lay in the close-packed A sheet, and the hexagonal **c**-axis is perpendicular to the stacking and runs from one A sheet to the next. There are two spheres (two atoms) in a unit cell, at positions 0, 0, 0 and ⅓, ⅔, ½ (Figure 3.23).

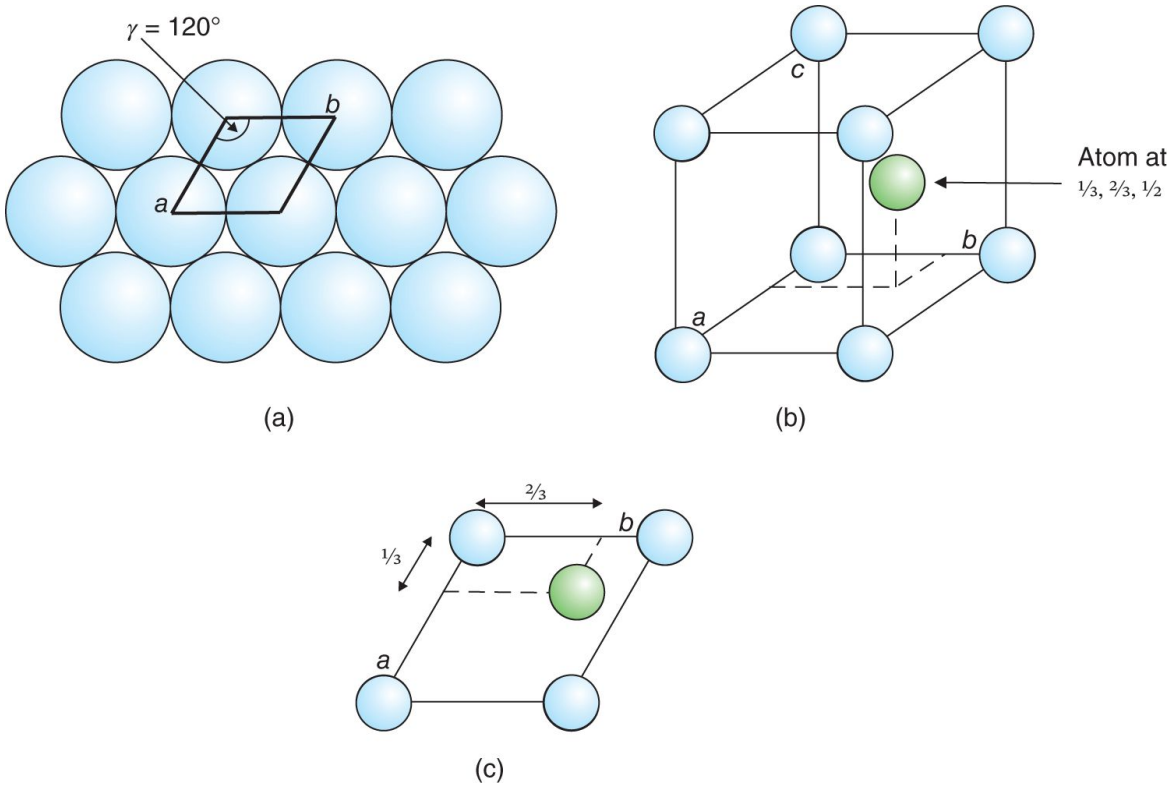


Figure 3.23 Hexagonal close-packing unit cell: (a) basal layer; (b) perspective view; (c) projection down (001).

If the spheres just touch, the lattice parameter, a, is given by:

$$a = b = 2r$$

where r is the sphere radius. The spacing of the layers, d, is given by:

$$d = \sqrt{8}r/\sqrt{3} \approx 1.633\,r$$

The lattice parameter in this direction, c, is equal to $2d$. The ratio of c/a in this ideal sphere packing is $\sqrt{8}/\sqrt{3} \approx 1.633$. The structure is equivalent to the hexagonal Mg structure, although the c/a ratio departs from the ideal value of 1.633 in the majority of crystals.

The second structure of importance, *cubic close-packing* (*ccp*), has a three-layer repeat. Two layers of spheres, A and B, pack as before. The position of the third layer, which occupies dimples in the B-layer directly below it, is not above either A or B, and is given the position label C (Figure 3.24). This three-layer stacking is repeated indefinitely, thus ... ABCABC...

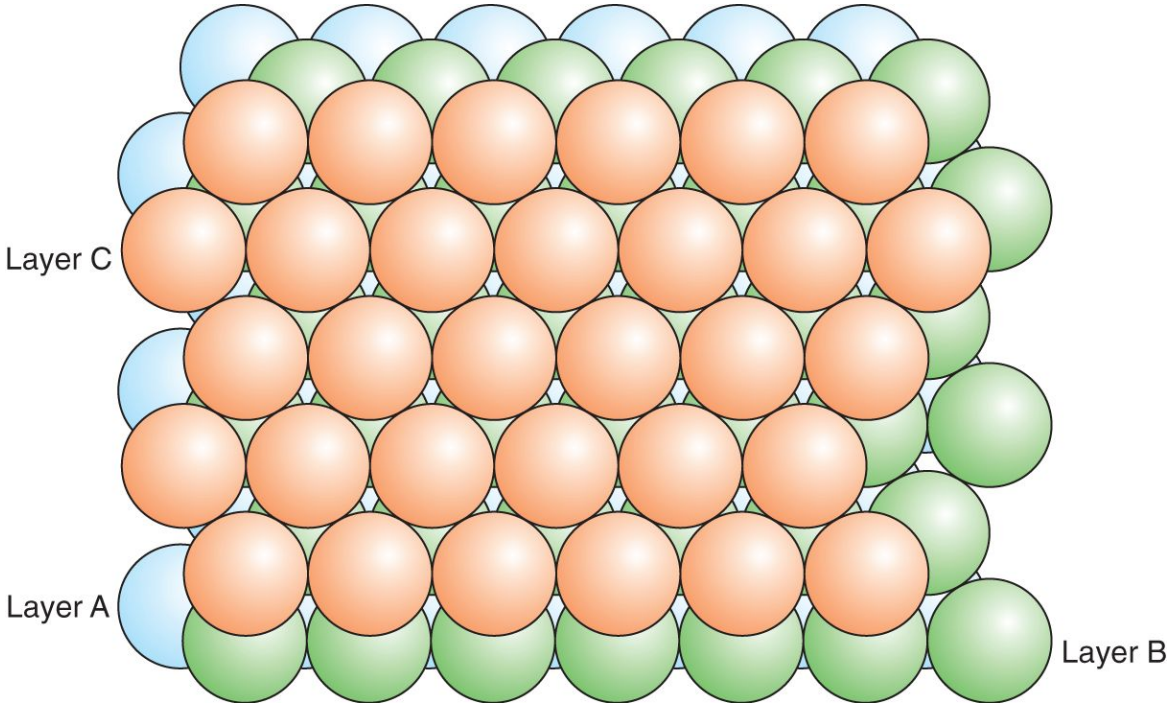


Figure 3.24 Cubic close-packing (ccp) of spheres. The layers follow the sequence ...ABCABC....

Although this arrangement can be described in terms of a hexagonal unit cell, the structure has cubic symmetry, and this description is always chosen. The close-packed layers of spheres lie perpendicular to the [111] direction and form (111) planes (Figure 3.25). The spacing of the close packed planes for an ideal packing, d, is ⅓ body diagonal of the cubic unit cell, i.e. $a/\sqrt{3}$. If the spheres just touch, the relationship between the sphere radius, r, and the lattice parameter a is:

$$r = a/\sqrt{8}$$

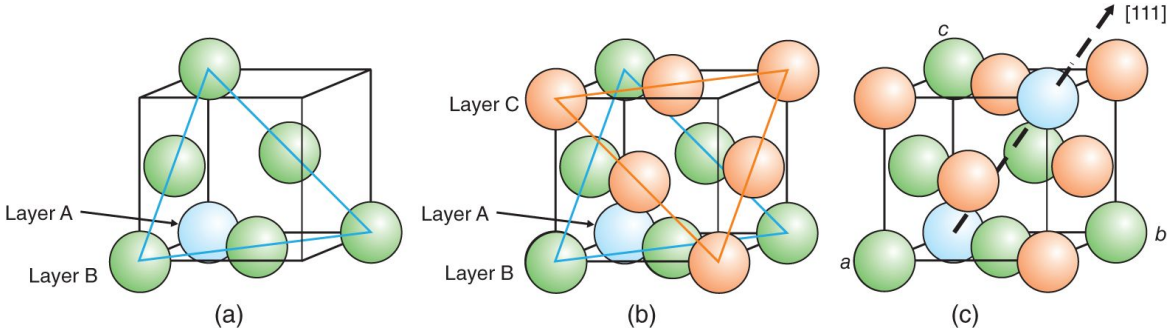


Figure 3.25 The cubic close-packing (ccp) unit cell: (a) layers A and B; (b) layers A, B and C; (c) perspective view of unit cell.

The relationship between the spacing of the close packed planes of spheres, d, the cell parameter a and r, is therefore:

$$d = a/\sqrt{3} = (\sqrt{8}\, r)/\sqrt{3} \approx 1.633\, r$$

which is identical to that in hexagonal closest packing. The ccp structure is identical to the fcc structure of Cu (Figure 3.5).

Hexagonal, hcp and cubic, ccp represent the (equally) densest packing of the spheres. The fraction of the total volume occupied by the spheres, when they touch, is 0.7405. Other sphere arrangements are found less commonly. The structure of the metal lanthanum is given by the sequence ...ABAC... In addition, a number of more complex packing sequences have been found, especially in silicon carbide and zinc sulfide.

3.5.2 Ionic Structures in Terms of Anion Packing

Relationships between structures can often be revealed by assuming that they are built of large spherical anions that form a close packed array (Table 3.3). Between these are small gaps, which are called *interstices*, *interstitial sites*, or *interstitial positions*, which are of two types (Figure 3.26). In one, three spheres in the lower layer are surmounted by one sphere in the layer above, or vice versa. The geometry of this site is that of a *tetrahedron*. In the other, a lower layer of three spheres and an upper layer of three spheres form a site with an *octahedral* geometry.

Table 3.3 Structures in terms of anion packing.

		Sequence of anion layers	
Fraction of tetrahedral sites occupied	**Fraction of octahedral sites occupied**	**...ABAB... (hex)**	**...ABCABC... (cubic)**
0	1	NiAs (nicolite)	NaCl (halite)
½	0	ZnO, ZnS (wurtzite)	ZnS (sphalerite or zinc blende)
0	⅔	Al_2O_3 (corundum)	—
0	½	TiO_2 (rutile), α-PbO_2	TiO_2 (anatase)
⅛	½	Mg_2SiO_4 (olivine)	$MgAl_2O_4$ (spinel)

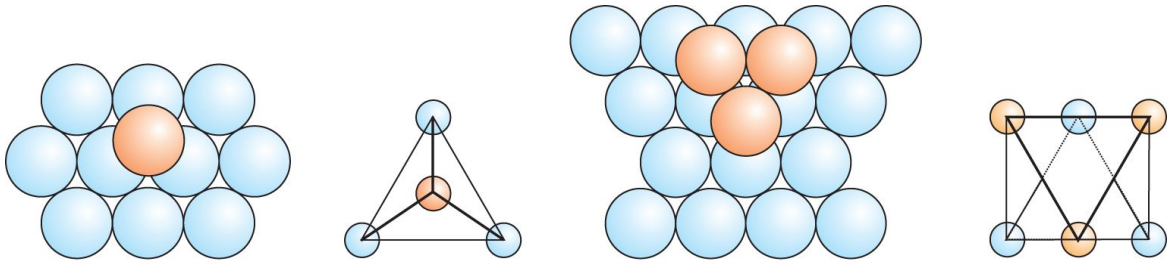


Figure 3.26 Sites between close-packed layers of spheres: (a), (b) tetrahedral; (c), (d) octahedral.

In the two close-packed sequences ...ABCABC... and ...ABAB..., there are 2*N* tetrahedral interstices and *N* octahedral interstices for every *N* anions. Structures are derived by placing cations into the interstices, making sure that the total positive charge on the cations is equal to the total negative charge on the anions. The formula of the structure can be found by counting up the numbers of type of ion present.

Compounds with halite (NaCl) structure can be thought of as a cubic close-packed array of X anions in which each octahedral site contains an

M cation. As there are equal numbers of octahedral sites and anions in the structure the formula of the phases is MX. In the case of halide anions, X^-, to maintain charge balance, each cation must have a charge of +1, and the alkali halide, MX, structure adopted by LiF, NaF, and KF and the other alkali halides, results. Should oxygen anions, O^{2-}, form the anion array, the cations must necessarily have a charge of 2+, to ensure that the charges balance and the oxides will have a formula MO, typified by a number of oxides, including MgO, CaO, SrO, BaO, MnO, FeO, CoO, and NiO.

Should the anions adopt hexagonal close-packing and all of the octahedral sites contain a cation, the hexagonal analogue of the halite structure is produced. The formula of the crystal is again MX and the structure is that of nicolite (NiAs), which is adopted by a number of alloys and metallic sulfides, including NiAs, CoS, VS, FeS, and TiS.

Filling only a fraction of the octahedral positions in the hexagonal packed array of anions in an ordered way replicates many well-known structures. Filling ⅔ of the octahedral sites in an ordered way produces the *corundum* structure, adopted by the oxides α-Al_2O_3, V_2O_3, Ti_2O_3, and Fe_2O_3. Of the structures that form when only half of the octahedral sites are occupied, those of rutile (TiO_2) and α-PbO_2 are best known. The difference between the two structures lies in the way in which the cations are ordered. In the rutile form of TiO_2, the cations occupy straight rows of sites, while in α-PbO_2 the rows are staggered.

A large number of structures can be generated by the various patterns of filling either the octahedral or tetrahedral interstices. The number can be extended if both types of position are occupied, as in spinel, the structure of which derives from a cubic close-packed array of oxygen ions with an ordered filling of ½ of the available octahedral sites and ⅛ of the available tetrahedral sites.

Structures containing cations in tetrahedral sites can be described in exactly the same way. In this case, there are twice as many tetrahedral sites as anions, and so if all sites are filled the formula of the solid will be M_2X. When half are filled this becomes MX, and so on.

3.5.3 Polyhedral Representations

It is often necessary to focus upon the surroundings of a particular atom or ion in a solid, and for this purpose, structures drawn in terms of

polyhedra are helpful. The polyhedra selected are generally metal–non-metal coordination polyhedra, composed of a central metal surrounded by non-metal atoms. Reducing the non-metal atoms to points and then joining the points by lines constructs the polyhedral shape. The two commonest polyhedra utilised for this are the (often idealised) tetrahedron (Figure 3.27) and octahedron (Figure 3.28). These polyhedra are then linked together to build up the complete structure. The advantage of using polyhedral representations of solids is that family relationships can be clearly illustrated. The disadvantage is that important structural details are often ignored, especially when polyhedra are idealised.

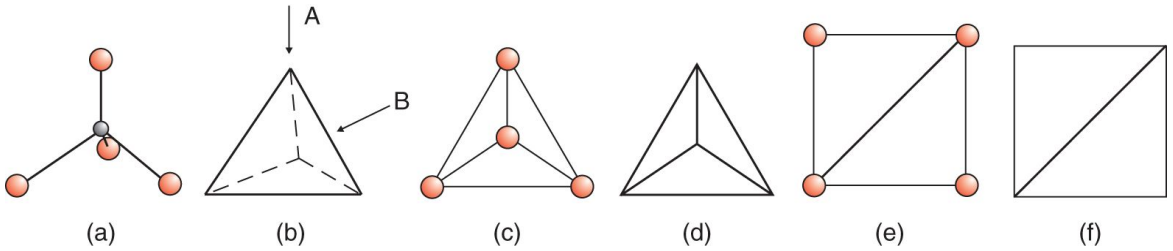


Figure 3.27 Representations of tetrahedra: (a), (b) perspective view; (c), (d) viewed down direction A; (e), (f) viewed down direction B.

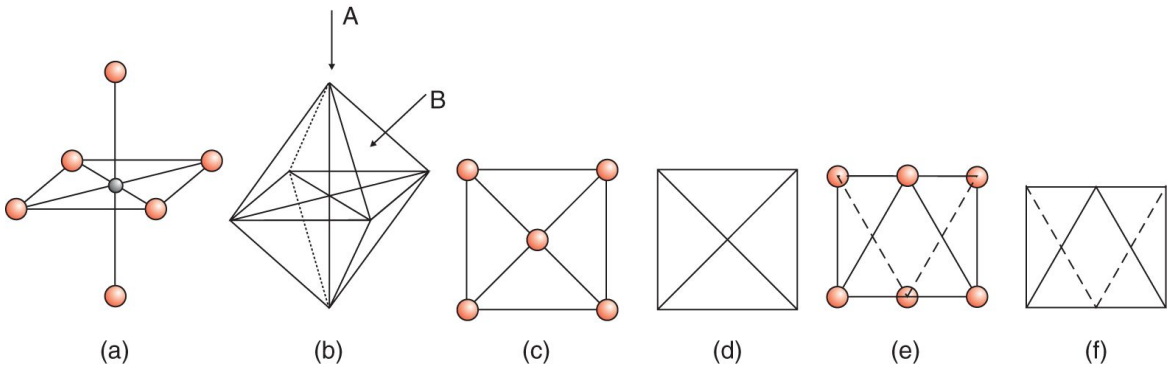


Figure 3.28 Representations of octahedra: (a), (b) perspective view; (c), (d) viewed down direction A; (e), (f) viewed down direction B.

The complex families of silicates are most easily compared if the structures are described in terms of linked tetrahedra. The tetrahedral shape used is the idealised coordination polyhedron of the SiO_4 unit (Figure 3.27). Each silicon atom is linked to four oxygen atoms by sp^3-hybrid bonds. The SiO_4 units are very strong and persist during physical and chemical reactions, so that structural transformations of silicates are often most easily visualised in terms of the rearrangement of the SiO_4 tetrahedra. Figure 3.29 shows the way in which the SiO_4 tetrahedra are linked in the commonest form of silica, α-(SiO_2) or α-quartz. The unit cell

contains three tetrahedra, and the heights of these, along the **c**-axis, show that they form a spiral. The direction of this spiral gives the material the ability to rotate the plane of polarisation of a light beam directed along the **c**-axis and differentiates it from β-quartz, which has an opposite spiral and rotates the beam in the opposite direction.

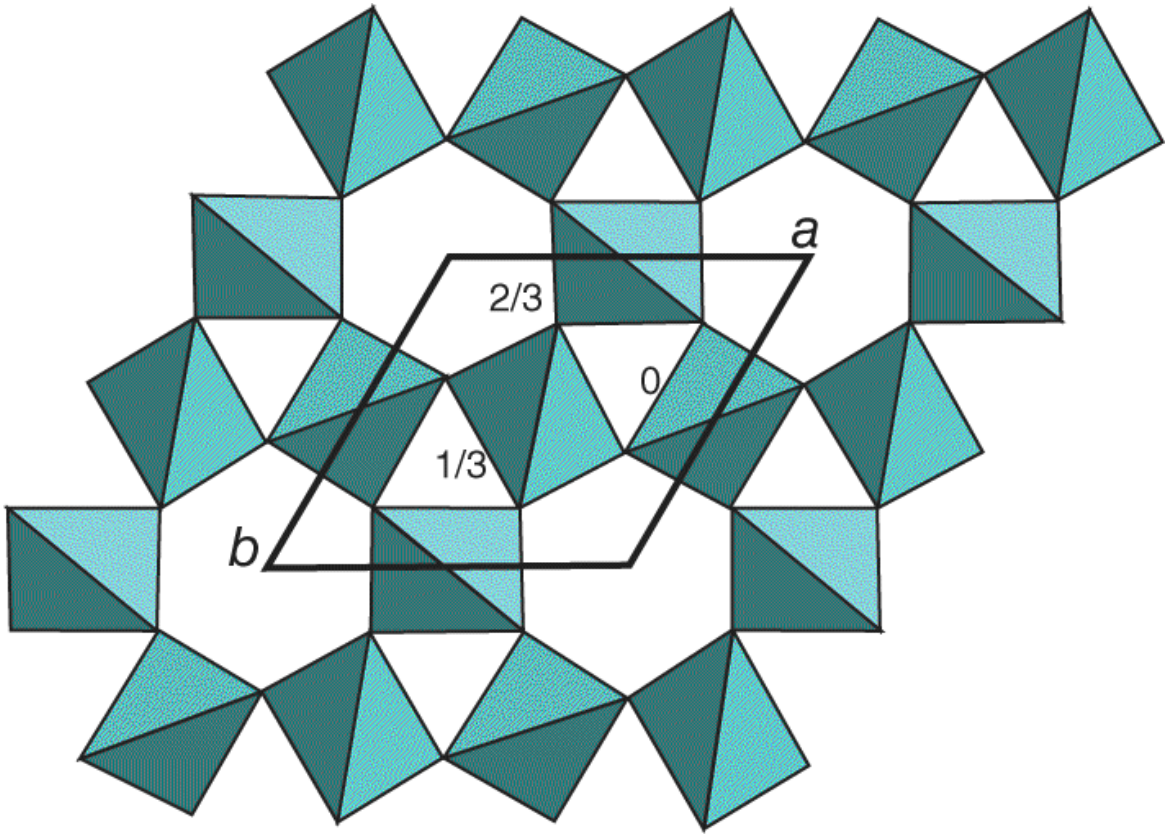


Figure 3.29 The structure of α-quartz (SiO_2) projected down the **c**-axis. The heights of the three tetrahedra within each unit cell are 0, ⅓, and ⅔ *c*.

Octahedral coordination is frequently adopted by the important 3d transition-metal ions. The structure of rhenium trioxide, ReO_3, in terms of linked ReO_6 octahedra, has the appearance of a three-dimensional chessboard (Figure 3.30a). This structure is similar to that of tungsten trioxide, WO_3, but in the latter compound, the octahedra are distorted slightly, so that the symmetry is reduced from cubic in ReO_3, to monoclinic in WO_3. The idealised cubic ABO_3 perovskite structure (see Figure 3.11) is similar, but has the large A cation in the centre of the cage of BO_6 octahedra (Figure 3.30b), although the structure can also be depicted with the large A cations at the cell corners surrounding a central BO_6 octahedron (Figure 3.30c). Most real perovskites are built of slightly

distorted BO_6 octahedra, which reduce the symmetry from cubic to orthorhombic or monoclinic.

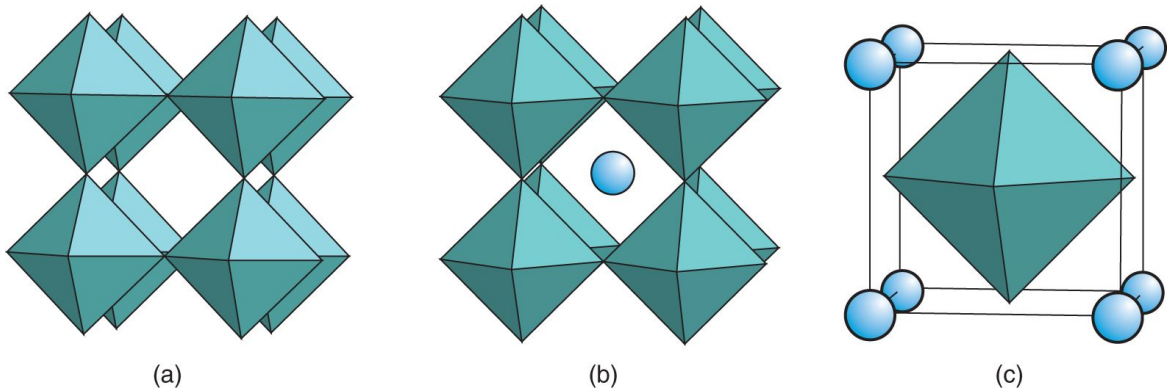


Figure 3.30 (a) The cubic ReO_3 structure represented as corner-shared ReO_6 octahedra; (b, c) the idealised cubic perovskite, ABO_3 structure.

3.6 Point Defects

Defects in crystalline solids are important because they modify properties. For example, just a trace of chromium impurity changes colourless corundum (aluminium oxide) crystals into ruby. Metals are ductile when linear defects (dislocations) are free to move. Crystals dissolve and react at increased rates at points where dislocations intersect external surfaces. It is necessary to have an idea of the types of defect that form and the role that they play in the control of properties in order to understand the behaviour of solids. This section describes defects at a single atom site in a crystal – *point defects*, while the Section 3.7 describes linear, planar, and volume defects.

3.6.1 Point Defects in Crystals of the Elements

Crystals of the solid elements such as silicon are monatomic, containing only one atom type. The simplest *localised* defect that we can imagine in a crystal is a mistake at a *single* atom site. These defects are called *point defects*. Two types of point defect can occur in a *pure* crystal, an atom can be absent from a normally occupied position, to create a *vacancy*, or an atom can occupy a position normally empty to form an *interstitial*, sometimes called a *self-interstitial* (Figure 3.31). Such vacancies and interstitials, which occur in even the purest of materials, are called *intrinsic defects*.

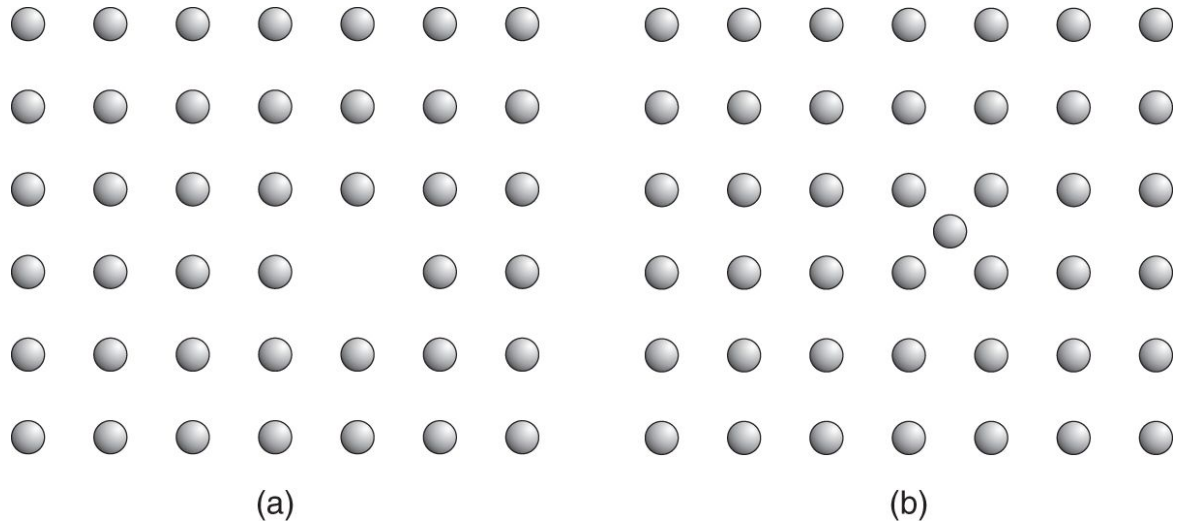


Figure 3.31 Idealised point defects in a pure monatomic crystal: (a) vacancy; (b) interstitial.

For these defects to be stable, the Gibbs energy of a crystal containing defects must be less than the Gibbs energy of a crystal without defects.

Initially, a population of point defects lowers the Gibbs energy, but ultimately large numbers of point defects result in an increase in the Gibbs energy. Thermodynamics allows the position of the minimum and the approximate number of point defects present in a crystal to be calculated. The number of defects per unit volume, n_d, is expressed by the formula:

$$n_d \approx N \exp(-\Delta H/k_B T)$$

where N is the number of sites affected by defects per unit volume, ΔH is the enthalpy (loosely the heat energy), needed to form a single defect, k_B is the Boltzmann constant, and T is the absolute temperature. The minimum in the curve represents the equilibrium situation that will exist at a given temperature and shows that even a pure crystal will contain an equilibrium population of point defects, n_e (Figure 3.32). The *fraction* of atom sites, which contain a defect, n_d/N, at any temperature, can be calculated if the enthalpy of defect formation is known.

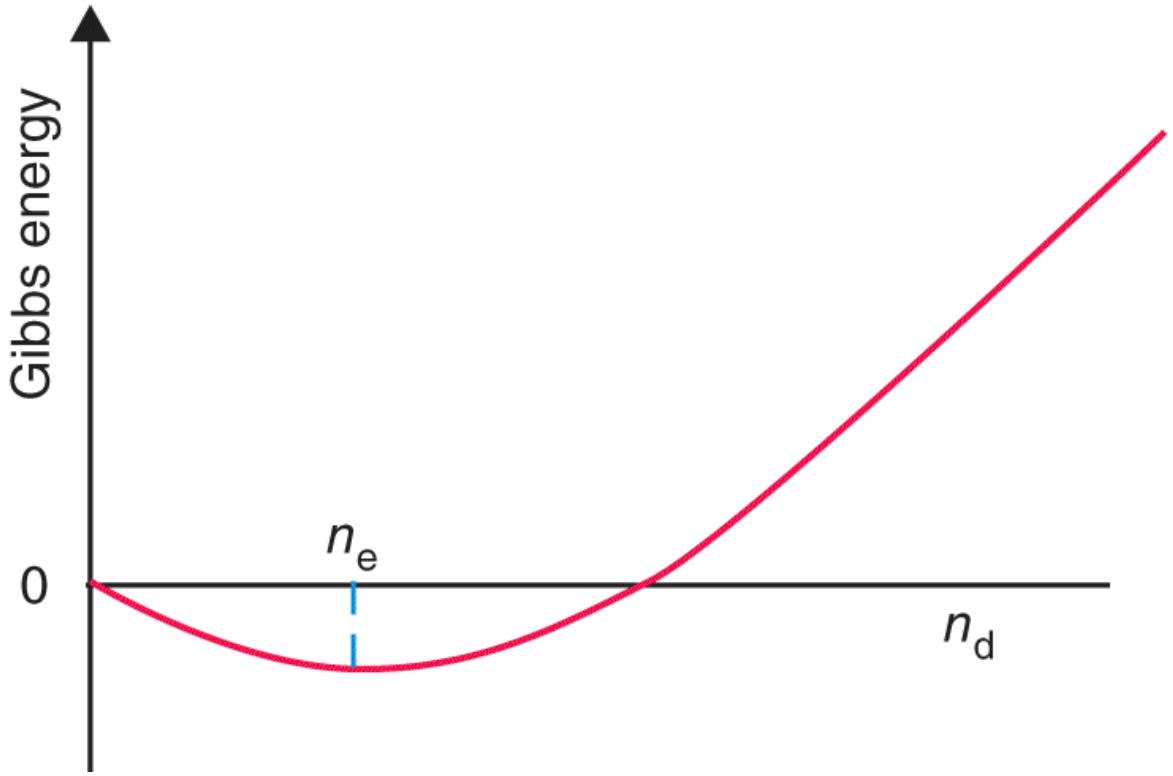


Figure 3.32 The Gibbs energy of a pure crystal as a function of the number of point defects present.

To obtain the absolute number of defects in the solid, it is necessary to know the number of atoms in a unit volume of the crystal. This value is

obtained from the crystal structure of the compound as the number of atoms in the unit cell divided by the unit cell volume. For example, the unit cell of silicon is cubic, with a side of 0.5431 nm, and contains eight atoms of silicon, which gives a value of 5×10^{28} atoms m^{-3}. Alternatively, it is possible to obtain the same information from the density of the material. The relative molar mass of an element of density ρ contains N_A atoms and the value of N is given by:

$$N = \rho \times N_A / \text{molar mass}$$

where N_A is Avogadro's constant.

No material is completely pure, and *foreign atoms* will be present. If these are undesirable or accidental, they are known as *impurities*, but if they have been added deliberately, to change the properties of the material on purpose, they are called *dopants*. Foreign atoms can rest on sites normally occupied by the parent atom type to form *substitutional defects*. Foreign atoms may also occupy normally empty positions to create *interstitial impurities* or *interstitial dopants* (Figure 3.33). There is no simple thermodynamic formula for the number of impurities present in a crystal.

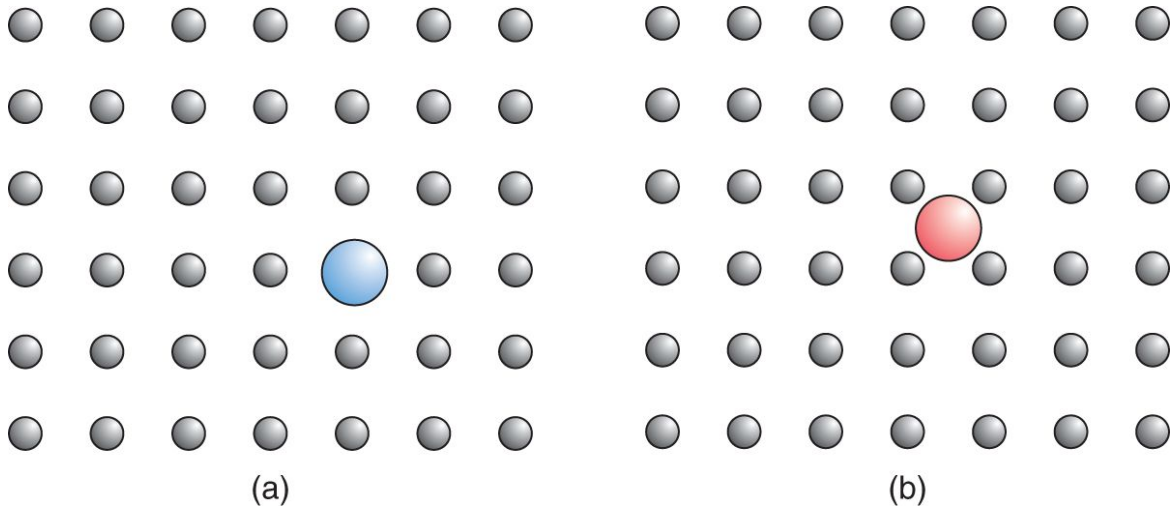


Figure 3.33 Impurity or dopant point defects in a crystal, schematic: (a) substitutional; (b) interstitial.

3.6.2 Solid Solutions

When quite large numbers of impurity atoms enter a crystal, without changing the crystal structure, the resultant phase is referred to as a *solid solution*. A *substitutional* solid solution forms with two similar elements when both share the same sites in the crystal. The copper–nickel system

provides an example. Both parent phases adopt the same fcc crystal structure and in alloys the two metals are distributed at random. Instances also abound in mineral crystals, almost all of which contain varying populations of substitutional impurities. Many precious gemstones owe their exquisite colours to such impurities. Ruby, for example, is a dilute solid solution of about 0.5% Cr_2O_3 in the isostructural Al_2O_3. The Fe_3O_4–Fe_2TiO_4 system forms a complete solid solution, with the Ti atoms partially substituting for Fe as the phase region is crossed. *Interstitial* solid solutions form when atoms enter spaces between the atoms in a crystal. The impurities must be small and are typically elements from the first row of the Periodic Table, such as C and N. The austenite phase found in steel is composed of iron containing up to 2.14 wt% of interstitial carbon. The incorporation of hydrogen into a number of metals, especially Li, Al, and Mg, is actively being studied as a method of hydrogen storage.

3.6.3 Schottky and Frenkel Defects

When point defects are introduced into nominally ionic compounds, the charges must remain balanced so as to maintain neutrality. This implies that stable populations of defects on anion, cation, or interstitial sites are necessary, although these defects need not be near to each other in the crystal. The defects arising from balanced populations of cation and anion vacancies in a crystal are known as the *Schottky defects* (Figure 3.34a). Any ionic crystal of formula MX must contain equal numbers of cation vacancies and anion vacancies; the ratio of two anion vacancies to one cation vacancy will hold in all compounds of formula MX_2, and in crystals with a formula M_2X_3, two vacancies on the cation sublattice will pair with three vacancies on the anion sublattice.

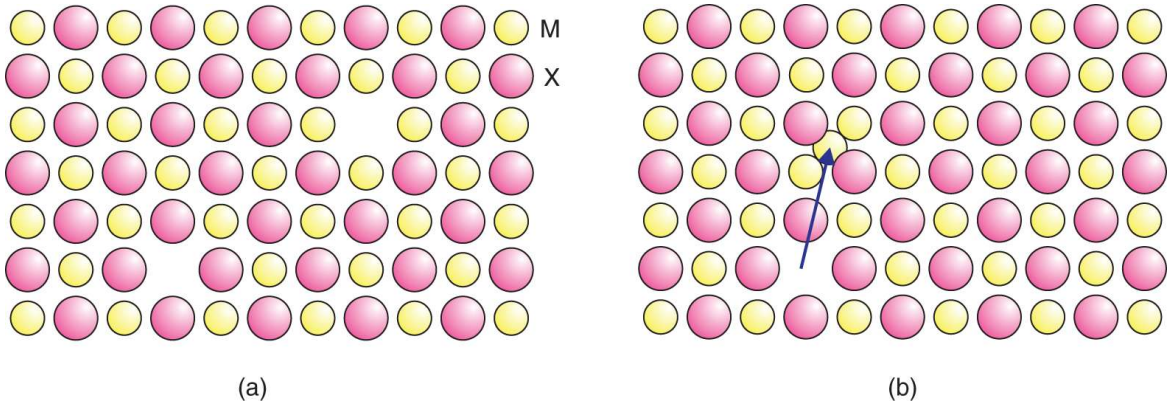


Figure 3.34 Point defects in an ionic crystal, formula MX, schematic: (a) Schottky defect; (b) Frenkel defect.

Under equilibrium conditions, the Gibbs energy of a crystal, G, is lower if it contains a small population of Schottky defects, similar to the situation shown in Figure 3.32. This means that Schottky defects will always be present in crystals at temperatures above 0 K, and hence Schottky defects are *intrinsic defects*. The approximate number of Schottky defects, n_S, in crystal with a formula MX, at equilibrium is given by:

$$n_S \approx Ne^{-\Delta H_S/2k_BT}$$

where N is the number of ions per unit volume, ΔH_S is the enthalpy (loosely the heat energy), required to form *one* defect, k_B is the Boltzmann constant, and T is the temperature (K), or:

$$n_S \approx Ne^{-\Delta H_S/2RT}$$

where ΔH_S is in J mol^{-1} and represents the enthalpy required to form 1 mol of Schottky defects, and R is the gas constant (Table 3.4).

Table 3.4 The formation enthalpy of Schottky defects in alkali halide compounds with the halite (NaCl) structure.

Compound	ΔH_S (kJ mol^{-1})	ΔH_S (eV[a])	Compound	ΔH_S (kJ mol^{-1})	ΔH_S (eV[a])
LiF	225.2	2.33	NaBr	203.0	2.10
LiCl	204.1	2.12	NaI	140.9	1.46
LiBr	173.4	1.80	KF	262.0	2.72
LiI	102.4	1.06	KCl	244.5	2.53
NaF	233.1	2.42	KBr	224.6	2.33
NaCl	225.7	2.34	KI	153.0	1.59

[a] The energy unit eV, electron volt, is widely used for defect energies.

As well as paired vacancies, interstitial defects can also form in ionic crystals. In the simplest case, an ion from one sub-lattice moves to a normally empty interstitial place in the crystal, leaving a vacancy behind. Such vacancy/interstitial pairs are known as *Frenkel defects* (Figure 3.34b). Because the total number of ions present does not change, there is no need for charge balance to be considered. For example, a Frenkel defect on the anion sublattice in fluorite, CaF_2, consists of one F^- ion displaced to an interstitial site together with the anion vacancy so formed.

Frenkel defects occur in silver bromide, AgBr. In this compound some of the Ag^+ ions move from the normal positions to sit at normally empty places to generate interstitial silver ions and leave behind vacancies on some of the normally occupied silver sites. Frenkel defects in AgBr make possible both black and white and colour film photography.

As with Schottky defects, the presence of a small number of Frenkel defects reduces the Gibbs energy of a crystal and so Frenkel defects are intrinsic defects. The formula for the equilibrium concentration of Frenkel defects in a crystal is similar to that for Schottky defects. Note, however, that the number of interstitial positions that are available to a displaced ion, N^*, need not be the same as the number of normally occupied positions, N, that the ion moves from. The number of Frenkel defects, n_F, present in a crystal of formula MX at equilibrium is given by:

$$n_F \approx (NN^*)^{1/2} e^{-\Delta H_F/2k_B T}$$

where ΔH_F is the enthalpy of formation of a single Frenkel defect, k_B is the Boltzmann constant, and T the absolute temperature (K). This is also often expressed in molar quantities:

$$n_F \approx (NN^*)^{1/2} e^{-\Delta H_F/2RT}$$

where ΔH_F is in J mol^{-1} and represents the enthalpy required to form 1 mol of Frenkel defects, and R is the gas constant.

The fraction of interstitials, $n_F/(NN^*)^{1/2}$, at any temperature can be calculated if the enthalpy of defect formation, ΔH_F is known. To obtain the absolute number of defects in the solid, it is necessary to know the number of ions affected by Frenkel disorder, N, and the number of sites that can accept an interstitial, N^*, in a unit volume of the crystal. These values can be assessed by a consideration of the crystal structure of the compound. The two numbers, N and N^*, are not usually identical, but in cases that they are, the formulae for Frenkel defects become identical to those for Schottky defects (Table 3.5).

Table 3.5 The formation enthalpy of Frenkel defects.

Compound	ΔH_F (kJ mol^{-1})	ΔH_F (eV)	Compound	ΔH_F (kJ mol^{-1})	ΔH_F (eV)
AgCl[a]	139.7	1.45	CaF_2[b]	261.4	2.71
AgBr[a]	109.0	1.13	SrF_2[b]	167.4	1.74
β-AgI[a]	57.8	0.60	BaF_2[b]	184.3	1.91

[a] Frenkel defects on the cation sublattice of a halite (NaCl) structure compound.

[b] Frenkel defects on the anion sublattice of a fluorite (CaF_2) structure compound.

3.6.4 Non-stoichiometric Compounds

Many non-molecular solids are found to exist over a composition range. This variation is considered normal in alloys but unusual in non-metallic compounds such as oxides. Non-metallic materials with a composition range are called *non-stoichiometric compounds*. Two ways in which this composition variation can occur are described.

Zirconia, ZrO_2, is an important oxide as it remains inert and stable at temperatures of up to 2500 °C and finds use in many high temperature applications. Unfortunately, pure zirconia fractures when cycled from high to low temperatures repeatedly, because the crystal structure changes at approximately 1100 °C. This shortcoming is overcome by reacting zirconia with calcia, CaO. The product, *calcia-stabilised zirconia*, exists over a wide composition range and for practical purposes calcia-stabilised zirconia can be cycled from high to low temperatures without problem.

This significant modification in the properties of zirconia is brought about by the introduction of defects into the crystal. As with Schottky defects, it is important that charge balance is maintained during the reaction of ZrO_2 and CaO. Cubic calcia-stabilised zirconia crystallises with the fluorite (CaF_2) structure. In the present case, the parent material is ZrO_2. The stabilised phase has Ca^{2+} cations in some of the positions that are normally filled by Zr^{4+} cations, that is, cation substitution has occurred (Figure 3.35). As the Ca^{2+} ions have a lower charge than the Zr^{4+} ions, the crystal will show an overall negative charge if the formula is written $Ca^{2+}{}_xZr^{4+}{}_{1-x}O_2$. The crystal compensates for the extra negative charge by

leaving some of the anion sites unoccupied. The number of vacancies on the anion sublattice needs to be identical to the number of calcium ions in the structure for exact neutrality. Thus, each Ca^{2+} added to the ZrO_2 produces an oxygen vacancy at the same time and the formula of the crystal is $Ca^{2+}_{x}Zr^{4+}_{1-x}O_{2-x}$. (Also see Section 5.14.3 for further information on ionic conductivity in this and related stabilised zirconia phases.)

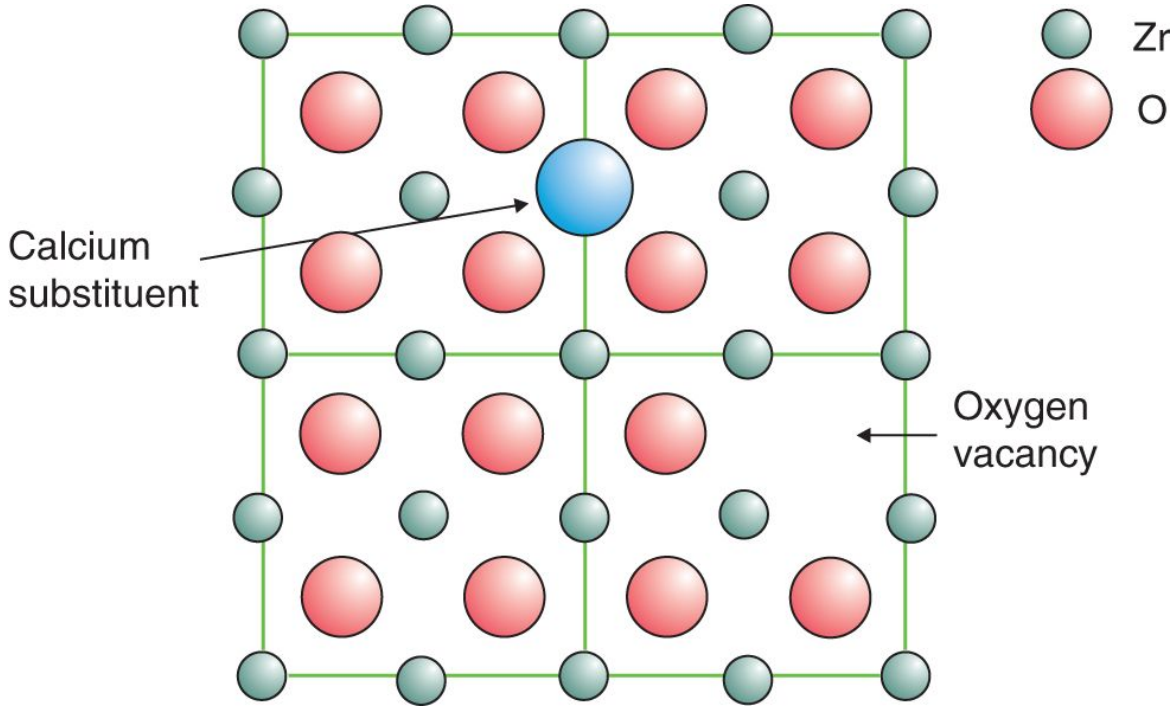


Figure 3.35 The structure of calcia-stabilised zirconia, schematic. The unit cell of the parent phase, ZrO_2, is outlined.

In cases where a cation of higher valence is substituted for the native cation in a crystal, charge balance will also be disturbed. This can occur, for example, if Mg^{2+} impurities are present in a crystal of NaCl. The Mg^{2+} cations will occupy Na^{+} sites forming substitutional impurity defects. In order to maintain charge balance, each Mg^{2+} impurity must be balanced by a vacancy on the cation sublattice. As a rule, in a crystal that contains substitutional impurity cations of a higher charge, cation vacancies tend to form in the sublattice of the lower charged ions.

It is also possible to vary the composition of a solid by introducing extra atoms into spaces within the crystal. This process is called *interpolation*. The likelihood of finding that a non-stoichiometric composition range is due to the presence of interpolated atoms in a crystal will depend upon the

openness of the structure and the size of the impurity. One of the most important groups of materials that use interpolation to modify properties is layered structures. In these materials atoms are taken in between the layers. The resulting compounds, often called *insertion compounds*, or *intercalation compounds*, are finding increasing use in batteries (Section 7.3). They are typified by disulfides such as TiS_2 (Figure 3.36), CoS_2, and NbS_2. The layers of these structures consist of strongly bonded cations in octahedral coordination; each layer being linked by weak van der Waals bonds and they form part of the group of van der Waals materials (Section 1.1.3). Small atoms such as Li can enter the structure between these weakly bound layers to form non-stoichiometric phases such as Li_xTiS_2. Because the bonding between the layers is weak, this process is easily reversible the compound can act as a convenient reservoir of Li atoms in lithium batteries.

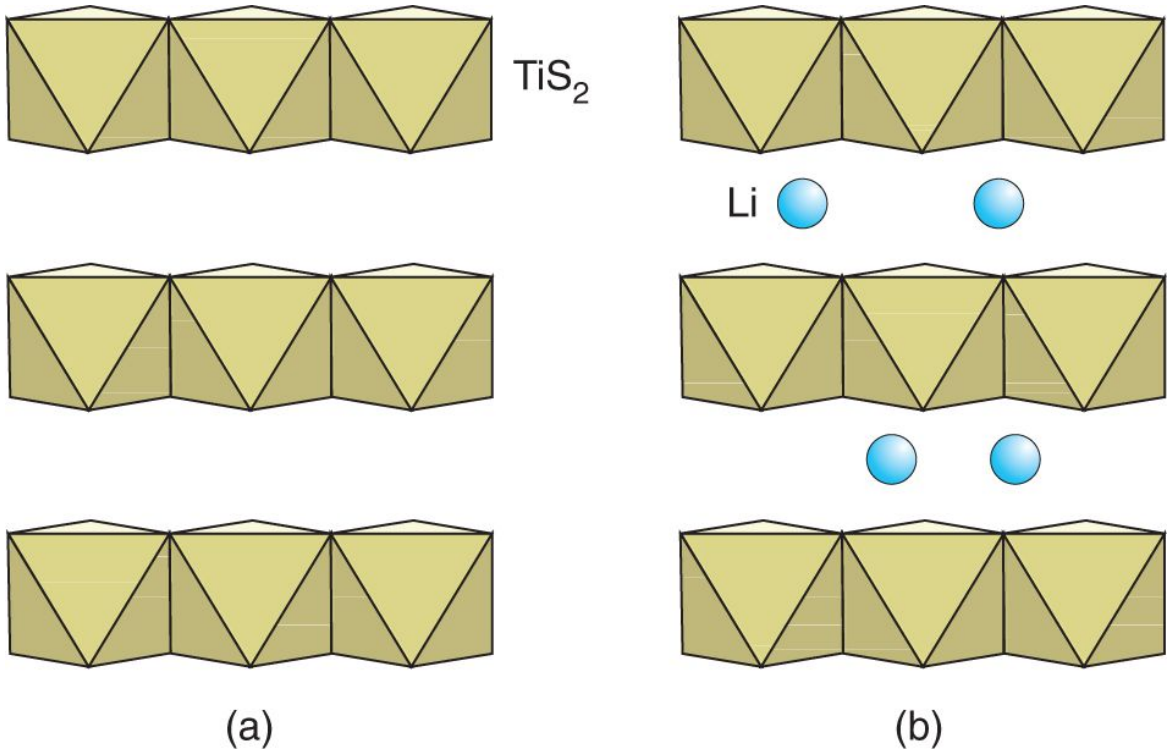


Figure 3.36 The layered structure of TiS_2, schematic: (a) without and (b) containing interpolated Li atoms.

Another group of similar van der Waals materials are derived from graphite. The layers of carbon atoms in graphite are only weakly linked by van der Waals bonding. Both atoms and molecules are able to enter the structure between the layers. As with Li_xTiS_2, lithium insertion into

graphite, Li_xC, is used as a battery material. The material that forms when fluorine is incorporated, CF_x, is used as a solid lubricant (also see Section 1.4)

3.6.5 Point Defect Notation

Point defect populations profoundly affect both the physical and chemical properties of materials. In order to describe these consequences, a simple and self-consistent set of symbols is required. The most widely employed system is the *Kröger–Vink notation* (Table 3.6). In this notation vacancies are indicated by the symbol V. (Because V is also the chemical symbol for the element vanadium, where confusion may occur the symbol for a vacancy is written Va.) The atom that is absent from a normally occupied site is specified by the normal chemical symbol for the element, written as a subscript. Thus, in NiO, for example, the symbol V_O would represent an oxygen atom vacancy and V_{Ni} a nickel atom vacancy.

Table 3.6 The Kröger–Vink notation for defects in crystals.[a]

Defect type	Notation	Defect type	Notation
Metal vacancy at metal (M) site	V_M	Non-metal vacancy at non-metal (Y) site	V_Y
Impurity metal (A) at metal (M) site	A_M	Impurity non-metal (Z) at non-metal site	Z_Y
Interstitial metal (M)	M_i	Interstitial non-metal (Y)	Y_i
Neutral metal (M) vacancy	V_M^x	Neutral non-metal (Y) vacancy	V_Y^x
Metal (M) vacancy with negative effective charge	V_M'	Non-metal (Y) vacancy with positive effective charge	$V_Y^\bullet$
Interstitial metal (M) with positive effective charge	$M_i^\bullet$	Interstitial non-metal (X) with negative effective charge	X_i'
Interstitial metal (M) with n positive effective charges	$M_i^{n\bullet}$	Interstitial non-metal (Y) with n negative effective charges	$Y_i^{n\prime}$
Free electron	e'	Free hole	$h^\bullet$
Associated defects	$(V_M V_Y)$		

[a] The definitive definitions of this nomenclature and further examples are to be found in the IUPAC Red Book on the Nomenclature of Inorganic Chemistry, Chapter I.6.

An impurity is given its normal chemical symbol and the site occupied is written as a subscript, using the chemical symbol for the atom that normally occupies the site. Thus, an Mg atom on an Ni site in NiO would be written as Mg_{Ni}. Interstitial positions, positions in a crystal not normally occupied by an atom, are denoted by the subscript i. For example, F_i would represent an interstitial fluorine atom in, say, a crystal of CaF_2.

It is possible for one or more lattice defects to associate with one another, that is, to cluster together. These are indicated by enclosing the components of such a cluster in parentheses. As an example, $(V_M V_X)$

would represent a defect in which a vacancy on a metal site and a vacancy on a non-metal site are associated as a vacancy pair.

One of the most difficult problems when working with defects, especially in ionic crystals, is to decide on the charge on the defects. The Kröger–Vink notation considers only *effective* charges. The effective charge on a defect is the charge that the defect has *with respect to the charge that would be present at the same site in a perfect crystal*. In order to distinguish effective charges from real charges, the superscript $'$ is used for each unit of effective negative charge and the superscript $^{\bullet}$ is used for each unit of effective positive charge. The real charges on a defect are still given the superscript symbols – and +.

For example, a sodium vacancy, V_{Na}, in the NaCl structure has a real charge of 0. However, the normal charge at the site in a perfect crystal, due to the presence of Na^+, is +1. Relative to the normal situation at the site, the vacancy appears to bear an *effective negative charge* equivalent to −1. Hence, a vacancy at a sodium ion (Na^+) site in NaCl would be written as V_{Na}'. In general, the absence of a positive ion will leave a vacancy with a negative effective charge relative to the normally occupied site. Multiple effective negative charges can exist, and are written using superscript n'. A Ca^{2+} ion vacancy in a crystal of CaO will bear an effective negative charge of $2'$, and the vacancy has the symbol $V_{Ca}^{2'}$.

The same reasoning indicates that the absence of a negative ion will leave a positive effective charge relative to a normal site occupied by a negative ion. A vacancy at a chloride ion (Cl^-) site is positively charged relative to the normal situation prevailing at an anion site in the crystal and would be written $V_{Cl}^{\bullet}$. In general, the absence of a negative ion will endow a site with a positive effective charge. Multiple effective positive charges can also exist. An oxide ion (O^{2-}) vacancy in a crystal of CaO has the symbol $V_O^{2\bullet}$.

Substitution of an ion with one valence by another with a different valence, *aliovalent substitution*, will create a charged defect. For example, a divalent ion such as Ca^{2+} substituted for a monovalent Na^+ on a sodium site in NaCl gives a defect with an effective charge of one, represented by the symbol $Ca_{Na}^{\bullet}$.

In a crystal containing defects, some fraction of the electrons may be free to move through the matrix. These are denoted by the symbol e'. The superscript $'$ represents the effective single negative charge on the electron

and it is written in this way to emphasise that it is considered relative to the surroundings rather than as an isolated real point charge. The counterparts to electrons in semiconducting solids are holes, represented by the symbol $h^{\bullet}$. Each hole will bear an effective positive charge of +1, which is represented by the superscript $^{\bullet}$ to emphasise that it is considered relative to the surrounding structure.

Interstitial sites, which are normally unoccupied in a crystal, will have no pre-existing charge. When an atom or an ion occupies an interstitial site, its effective charge is the same as the real charge. Thus, a Zn^{2+} ion at an interstitial site is given the symbol $Zn_i^{2\bullet}$.

Not all defects carry effective charges. Frequently this need not be noted. For instance, suppose that a sodium ion in NaCl, represented by Na_{Na}, is substituted by a potassium ion, represented by K_{Na}. Clearly, the defect will have no effective charge. The same could be said of a neutral lithium atom introduced into an interstitial site in titanium disulfide, TiS_2, which would be written Li_i. If it is necessary to emphasise that the defect is neutral in terms of effective charge a superscript *x* is used. Thus, a K^+ ion substituted for a Na^+ ion could be written K_{Na}^{x} when the effective charge situation needs to be specified. Similarly, an interstitial Li atom could be represented as Li_i^{x} to emphasise the lack of an effective charge on the defect when it is essential to do so.

3.7 Linear, Planar, and Volume defects

3.7.1 Dislocations

It has long been known that the strength of a metal crystal is far less than the theoretical strength. Moreover, metals can be deformed easily and retain the new shape, a process called *plastic deformation*, while ceramic solids fracture under the same conditions. These typical mechanical properties of metals are due to the presence of *linear* defects called *edge dislocations* that are free to move in the solid when stresses are applied. If dislocation movement is impeded or impossible, the material becomes hard and brittle. This is so in ceramics, in which dislocation movement is impeded, at least in part, by the charges on the ions.

Edge dislocations consist of an extra ½ plane of atoms inserted into the crystal (Figure 3.37a). These dislocations are instrumental in allowing metals to undergo plastic deformation, as a force applied to the crystal will allow both the top and bottom parts to slide past each other parallel to the plane where the dislocation tip is found, called the *slip plane* (see also Section 8.3). The disruption to the crystal introduced by a dislocation is characterised by the *Burgers vector*. The Burgers vector of a dislocation is determined by drawing a circuit in the crystal, a *Burgers circuit*, from atom to atom, in a region of crystal away from the defect (Figure 3.37b,c). The Burgers circuit starts and end on the same atom in a perfect crystal, but will not do so if the circuit contains a dislocation. The vector describing this failure, running from the start atom to the end atom, is the *Burgers vector*, **b**, of the dislocation. The Burgers vector of an edge dislocation is perpendicular to the dislocation line.

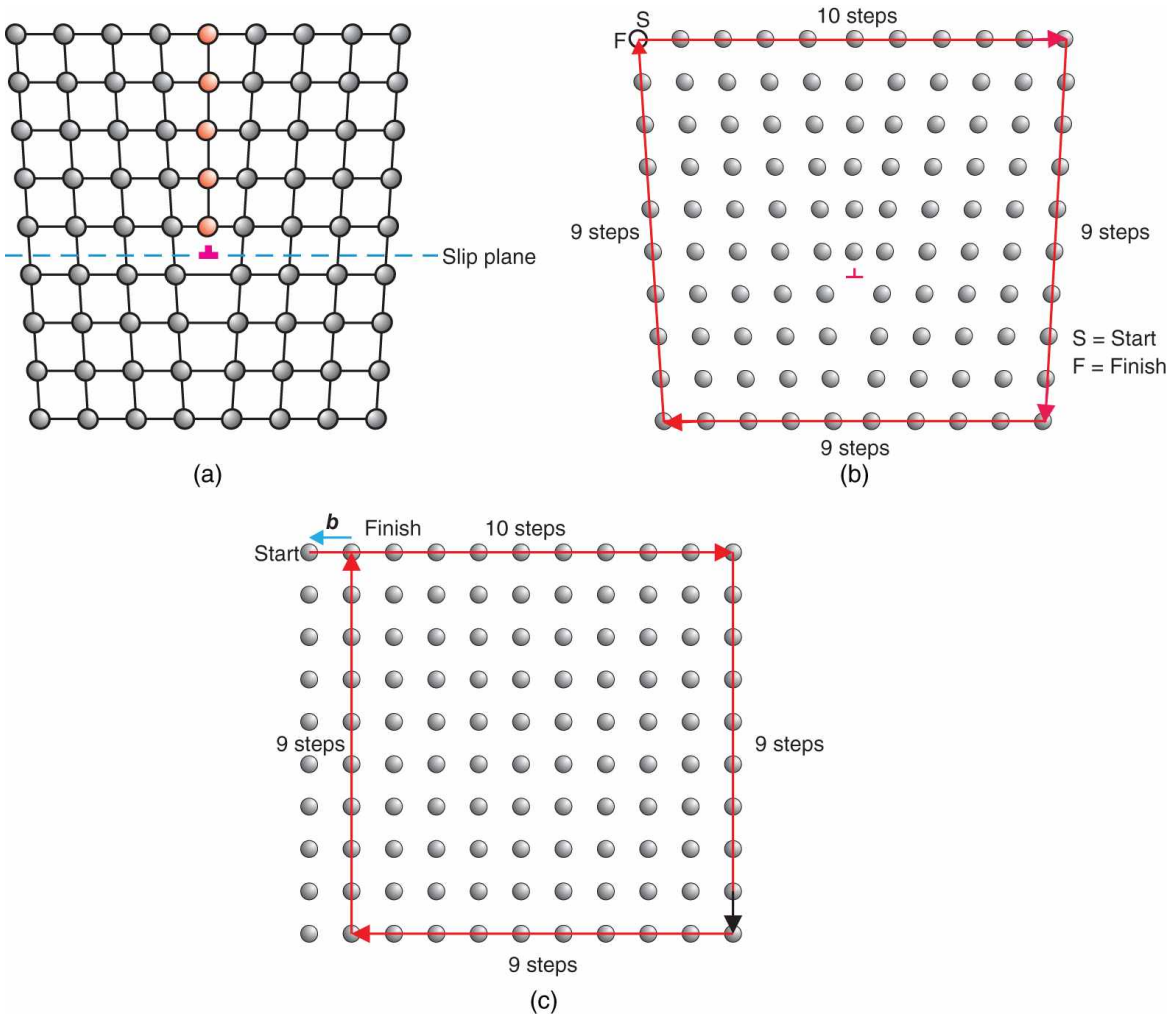


Figure 3.37 (a) an edge dislocation in a metal crystal: (b), (c) a circuit around an edge dislocation and a perfect crystal, defining the Burger's vector, **b**.

Screw dislocations, the second important category of dislocation, look rather like spiral staircases. The dislocation can be (hypothetically) formed by cutting halfway through a crystal and sliding the regions on each side of the cut parallel to the cut, to create spiralling atom planes (Figure 3.38a,b). The dislocation line is the central axis of the 'staircase'. The Burgers vector of a screw dislocation is parallel to the dislocation line and is determined in the same way as edge dislocations (Figure 3.38c,d). Screw dislocations play an important part in crystal growth.

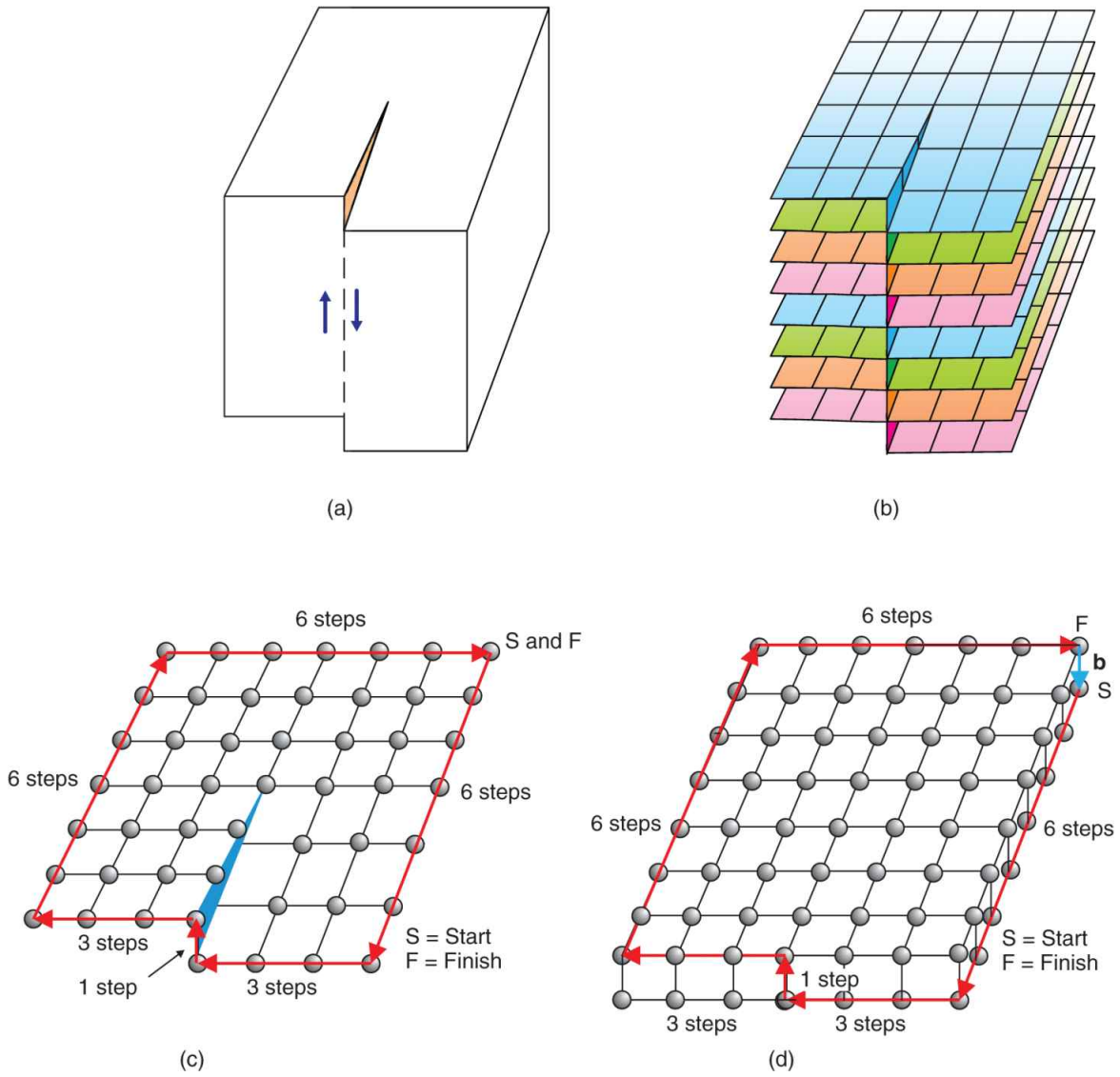


Figure 3.38 Screw dislocations: (a), (b) screw dislocation formed by displacement of two halves of the crystal; (c), (d) a circuit around a screw dislocation and a perfect crystal, defining the Burger's vector, **b**.

A dislocation line separates a region of crystal that has been displaced relative to an adjoining part. This disruption means that the dislocation must either end on the surface of a crystal or else form a closed loop. *Dislocation loops* have been found to occur frequently in crystals. They can form by the aggregation of vacancies or interstitials on a plane in a metal crystal. These point defects become increasingly mobile at high temperatures and if a crystal containing a population of either vacancies or interstitials is held at a high temperature, these defects can migrate from site to site. In some circumstances they can aggregate on a single atomic plane to form a dislocation loop (Figure 3.39). Although in such a

case the dislocation that delineates the loop is an edge dislocation, as dislocation loops grow they can change character, so that at one part of the loop the Burgers vector is parallel to the dislocation line and at another part it is parallel to it. The nature of the dislocation changes from pure edge to pure screw. Elsewhere on the loop, the dislocation has an intermediate character and is called a *mixed dislocation*. Finally, it can be noted that not all dislocations have unit step Burgers vectors and many have a Burgers vector that is less than the repeat distance of the structure. These are called *partial dislocations*.

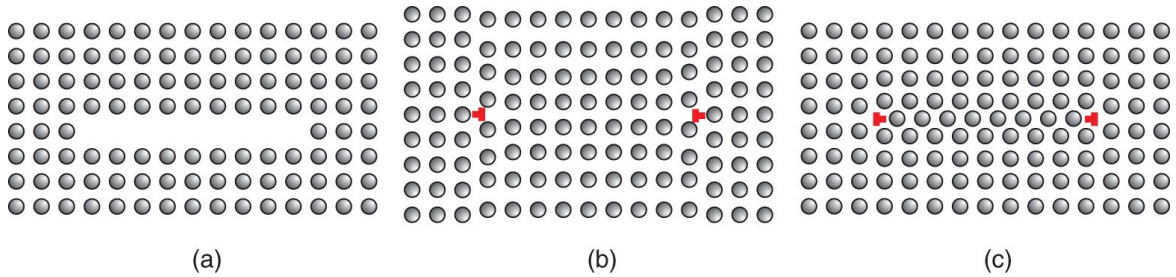


Figure 3.39 Formation of dislocation loops: (a) the aggregation of vacancies onto a single plane; (b) the collapse of the plane; (c) the aggregation of interstitials.

3.7.2 Planar Defects

In many cases, the most important planar defects in a solid are the *external surfaces*. The creation of solid surfaces involves an energy cost, called *surface energy*. This energy arises because of the unbalanced nature of the bonding forces between the atoms in the surface compared with those between the same atoms situated in the bulk. The external surfaces of solids may dominate the properties of the sample. For example, catalysts must have large *active* surfaces in order to function. Rates of reaction during corrosion are frequently determined by the amount of surface exposed to the corrosive agent. Porous solids, including zeolites, which are naturally occurring aluminosilicates consisting of open frameworks constructed of corner sharing $(Si,Al)O_4$ tetrahedra, and metal organic frameworks (MOFs), essentially display a large surface area that can be exploited in a variety of ways.

Grain boundaries are interfaces between crystallites in a polycrystalline array (Figure 3.40a). The energy of these boundaries, much of which is a form of surface energy, depends upon the crystallographic planes that make up the boundaries. Annealing, that is, heating for extended periods at temperatures high enough to allow for extensive atom diffusion, will

cause rearrangement to occur, leading to lower energy configurations. In these cases, there is often a relationship between the crystallography of the material and the boundary planes. Polymorphic transitions between metal phases, such as the transition between α-iron (ferrite) and γ-iron (austenite) (see Figure 2.18), are often initiated at grain boundaries. Although the atomic structure that characterises grain boundaries has been difficult to determine, high-resolution electron microscopy has shown that these regions are not, as was generally assumed, amorphous, but consist of small numbers of well-defined crystalline volumes, many of which only occur in the boundary region and are not found in bulk samples.

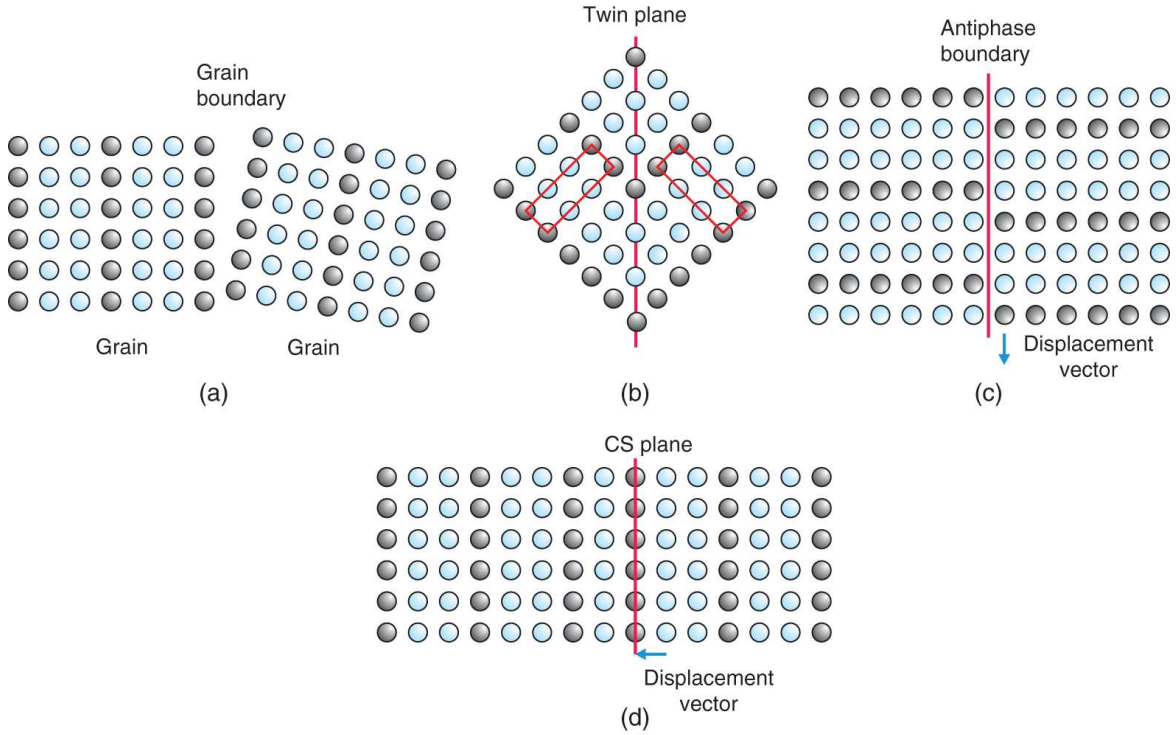


Figure 3.40 Planar defects in crystals: (a) grain boundary; (b) twin plane; (c) antiphase boundary; (d) crystallographic shear plane.

In terms of crystal strength, grain boundaries may show two opposite tendencies. They are frequently weaker than the crystal matrix, and the mechanical strength of many solids is limited by the presence of grain boundaries. In metals, grain boundaries prevent dislocation motion and reduce ductility and in this sense can enhance the mechanical properties of solids. Grain boundaries also increase the electrical resistance of a polycrystalline solid compared to a single crystal, and introduce scattering and opacity into otherwise transparent solids.

Twin boundaries are planes in which the crystal matrix on one side mirrors the crystal matrix on the other (Figure 3.40b). The mirror plane or *twin plane* may not be identical to the plane along which the two mirror-related parts of the crystal join, which is called the *composition plane*. Twin boundaries affect mechanical, optical, and electronic properties of materials in a similar way to grain boundaries.

Antiphase boundaries (APB's) are planes within a crystal across which one part of the crystal has been displaced with respect to the other side (Figure 3.40c). The vector describing the displacement of the two parts of the crystal is *parallel* to the boundary plane. When the displacement vector of the boundary is at an angle to the interface so that there is a collapse of the crystal, the boundary is called a *crystallographic shear (CS) plane* (Figure 3.40d). In these latter boundaries, collapse is equivalent to the removal of one or more planes of atoms and the composition of the crystal changes slightly. Regular arrays of crystallographic shear planes in a crystal lead to a series of new compounds, called a *homologous series*. For example, removal of oxygen from titanium dioxide (rutile), TiO_2, causes ordered arrays of CS planes to form giving rise to the homologous series Ti_nO_{2n-1}, running from Ti_4O_7 to Ti_9O_{17} and $Ti_{10}O_{19}$.

3.7.3 Volume Defects: Precipitates

Volume defects are regions of an impurity phase in the matrix of a material, typified by *precipitates* in a solid (Figure 3.41). Precipitates form in a variety of circumstances. Solid solutions are often not stable at low temperatures, and decreasing the temperature of a solid solution slowly will frequently lead to the formation of precipitates of a new crystal structure within the matrix of the solid solution. Glasses are inherently unstable, and a glass may slowly recrystallise so that precipitates of crystalline material appear. Precipitates have important effects on the mechanical, electronic, and optical properties of solids. Precipitation hardening is an important process used to strengthen metal alloys. In this technique, precipitates are induced to form in the alloy matrix by carefully controlled heat treatment. These precipitates interfere with dislocation movement and have the effect of hardening the alloy significantly. Precipitates are deliberately introduced into opal glass to produce the required opacity.

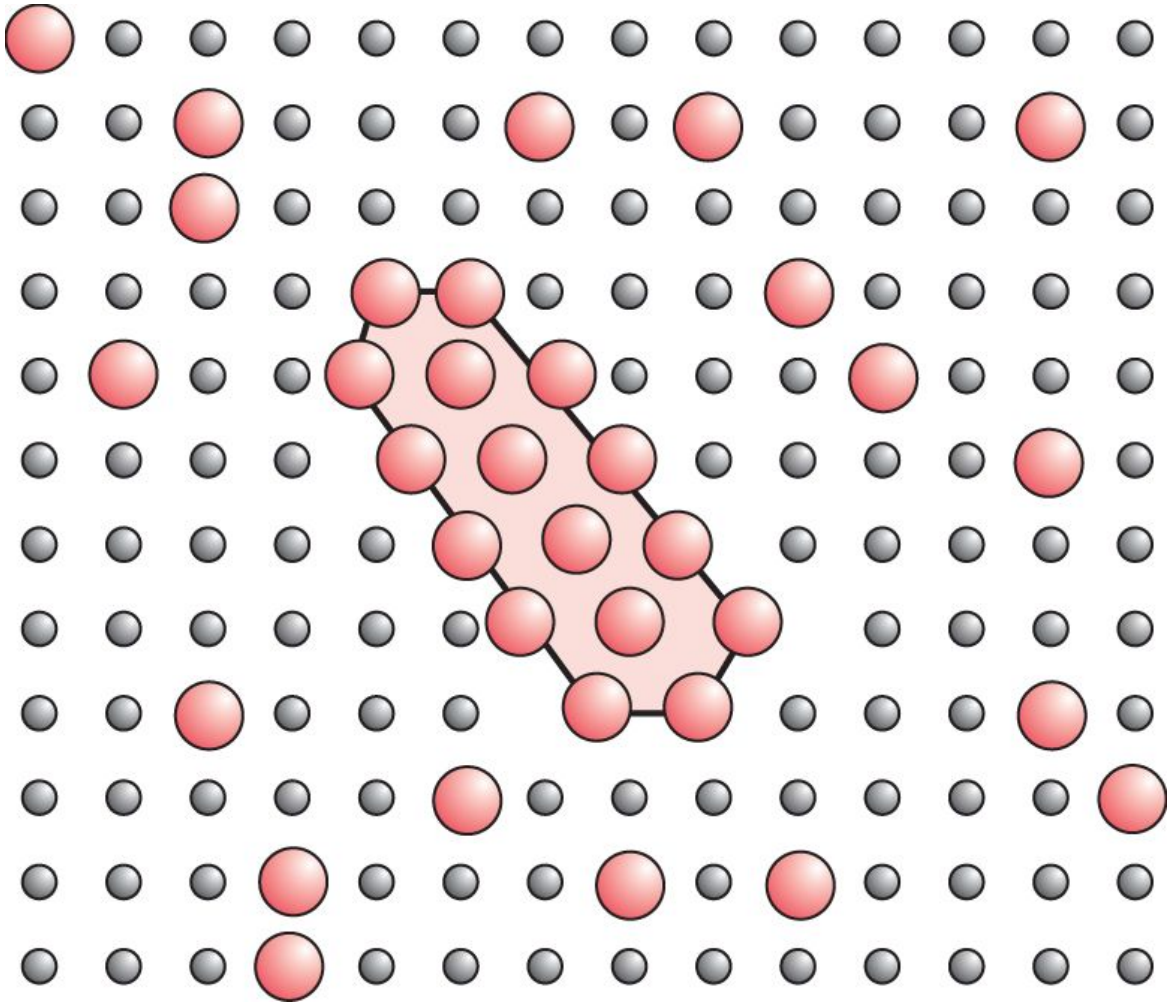


Figure 3.41 A precipitate formed by the clustering of impurity atoms in a crystal.

Reference

Jette, E.R. and Foote, F. (1933). An X-ray study of the Wüstite (FeO) solid solutions. *J. Chem. Phys.***1**: 29.

Further Reading

Crystal Structures:

Giacovazzo, C., Monaco, H.L., Artioli, G. et al. (2002). *Fundamentals of Crystallography*, 2e. Oxford: International Union of Crystallography, Oxford University Press.

O'Keeffe, M. (1977). On the arrangement of ions in crystals. *Acta Crystallogr.***A33**: 924.

O'Keefe, M. and Hyde, B.G. (1985). An alternative approach to non-molecular crystal structures. *Struct. Bond.***61**: 77–144.

Tilley, R.J.D. (2020). *Crystals and Crystal Structures*, 2e. Chichester: Wiley.

Defects:

Li, G., Blake, G.R., and Palestra, T.M. (2017). Vacancies in functional materials for clean energy storage and harvesting: the perfect imperfection. *Chem. Soc. Rev.***46**: 1693.

Meiners, T., Frolov, T., Rudd, R.E. et al. (2020). Observations of grain boundary phase transformations in an elemental crystal. *Nature***579**: 375.

Schmitt, R., Nenning, A., Kraynis, O. et al. (2020). A review of defect structure and chemistry in ceria and its solid solutions. *Chem. Soc. Rev.***49**: 554.

Tilley, R.J.D. (2020). *Defects in Solids*. Hoboken, NJ: Wiley.

Problems and Exercises

Quick Quiz

3.1 The basis vectors in a lattice define:

a. The unit cell.

b. The crystal structure.

c. The atom positions.

3.2 The number of Bravais lattices is:

a. 12.

b. 13.

c. 14.

3.3 A face centred (F) unit cell contains:

a. One lattice point.

b. Two lattice points.

c. Four lattice points.

3.4 A face-centred unit cell with a lattice point in the plane cutting the **b**-axis is:

a. A-face-centred.

b. B-face-centred.

c. C-face-centred.

3.5 A crystal system is:

a. A set of axes.

b. A lattice.

c. A unit cell.

3.6 A tetragonal unit cell has:

a. $a = b = c$.

b. $a = b \neq c$.

c. $a \neq b \neq c$.

3.7 A crystal class represents:

a. The internal symmetry of a crystal.

b. The unit cell of a crystal.

c. The crystal lattice.

3.8 A point group is identical to:

a. A crystal structure.

b. A crystal lattice.

c. A crystal class.

3.9 The Miller indices (*hkl*) represent:

a. A single plane in a crystal structure.

b. A set of parallel planes in a crystal structure.

c. A family of planes related by symmetry in a crystal structure.

3.10 An (*h*00) plane in a crystal structure is:

a. Parallel to the **a**- and **b**-axes.

b. Parallel to the **b**- and **c**-axes.

c. Parallel to the **a**- and **c**-axes.

3.11 A (110) plane in a crystal cuts:

a. The **a**- and **b**-axes.

b. The **b**- and **c**-axes.

c. The **a**- and **c**-axes.

3.12 {*hkl*} means:

a. A set of parallel planes.

b. A group of non-equivalent planes.

c. A family of symmetry related planes.

3.13 Miller–Bravais indices (*hkil*) are used with:

a. All non-cubic crystals.

b. Hexagonal crystals.

c. Primitive crystals.

3.14 A direction in a crystal structure is represented by:

a. {*uvw*}.

b. [*uvw*].

c. ⟨*uvw*⟩.

3.15 The atom coordinates ½, ½, ½ represent an atom at:

The centre of a unit cell.

The face centres of a unit cell.

The middle of the edges of a unit cell.

3.16 In a cubic close-packed (A1) unit cell there are:

a. One atom.

b. Two atoms.

c. Four atoms.

3.17 In a body-centred cubic (A2) unit cell there are:

a. One atom.

b. Two atoms.

c. Four atoms.

3.18 In a hexagonal (A3) unit cell there are:

a. One atom.

b. Two atoms.

c. Four atoms.

3.19 The symbol *Z* gives the number of:

a. Formula units in a unit cell.

b. Atoms in a unit cell.

c. Atom positions in a unit cell.

3.20 The sphere packing giving rise to a hexagonal structure is:

a. ...ABABAB....

b. ...ABCABC....

c. ...ABACABAC....

3.21 In a cubic close-packed array of *N* spheres there are:

a. *N* tetrahedral interstices.

b. 2*N* tetrahedral interstices.

c. $4N$ tetrahedral interstices.

3.22 In a hexagonal close-packed array of N spheres there are:

a. N octahedral interstices.

b. $2N$ octahedral interstices.

c. $4N$ octahedral interstices.

3.23 A tetrahedron is a polyhedron with:

a. Four triangular faces.

b. Six triangular faces.

c. Eight triangular faces.

3.24 An octahedron is a polyhedron with:

a. Four triangular faces.

b. Six triangular faces.

c. Eight triangular faces.

3.25 The structure of lithium oxide can be thought of as having anions in a cubic close-packed array with lithium ions in all of the tetrahedral positions. The formula of the oxide is:

a. Li_2O.

b. LiO.

c. LiO_2.

3.26 The alloy nickel arsenide has a structure in which all of the arsenic atoms are in a hexagonal close-packed array and the nickel atoms occupy all of the octahedral positions. The formula of nickel arsenide is:

a. Ni_3As.

b. Ni_2As.

c. NiAs.

3.27 The wurtzite structure of zinc sulfide has the sulfur atoms in a hexagonal close-packed array while the zinc atoms occupy half of the tetrahedral positions. The formula of the sulfide is:

a. Zn_2S.

b. ZnS.

c. ZnS_2.

3.28 The number of different intrinsic point defects possible in a single crystal of a pure element is:

a. One.

b. Two.

c. Three.

3.29 A Schottky defect in a crystal of potassium bromide, KBr, consists of:

a. A potassium vacancy and a bromide interstitial.

b. A potassium vacancy and a bromide vacancy.

c. A potassium interstitial and a potassium vacancy.

3.30 A Frenkel defect in a crystal of silver bromide, AgBr, consists of:

a. A silver vacancy and a bromide interstitial.

b. A silver vacancy and a bromide vacancy.

c. A silver interstitial and a silver vacancy.

3.31 The Burgers vector and dislocation line are normal to each other in:

a. A screw dislocation.

b. A partial dislocation.

c. An edge dislocation.

3.32 The Burgers vector and dislocation line are parallel to each other in:

a. A screw dislocation.

b. A partial dislocation.

c. An edge dislocation.

3.33 A dislocation loop has:

a. No Burgers vector.

b. A Burgers vector always normal to the loop.

c. A Burgers vector that changes along the periphery of the loop.

3.34 Dislocations are:

a. Planar defects.

b. Line defects.

c. Point defects.

3.35 A twin boundary in a solid is an example of a:

a. Line defect.

b. Planar defect.

c. Volume defect.

3.36 A precipitate is a:

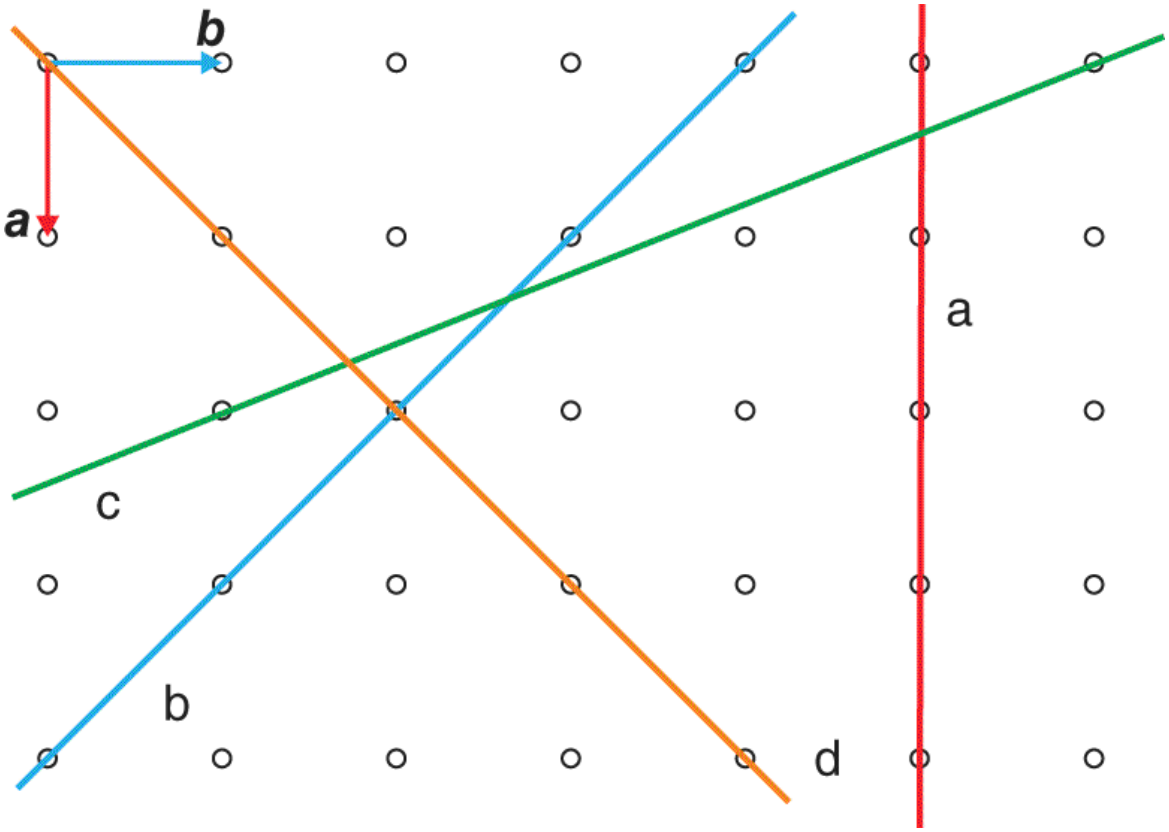
a. Point defect.

b. Planar defect.

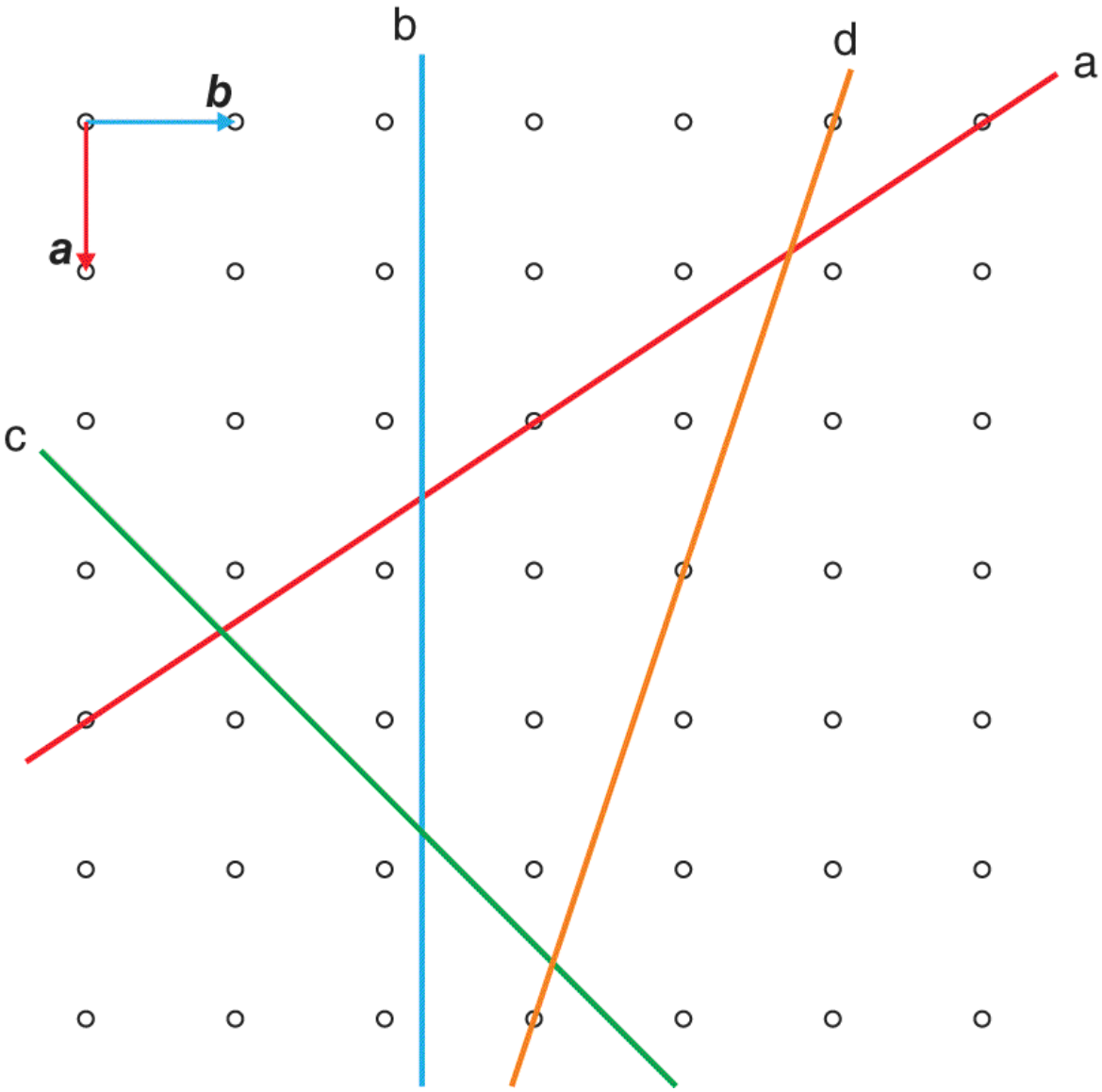
c. Volume defect.

Calculations and Questions

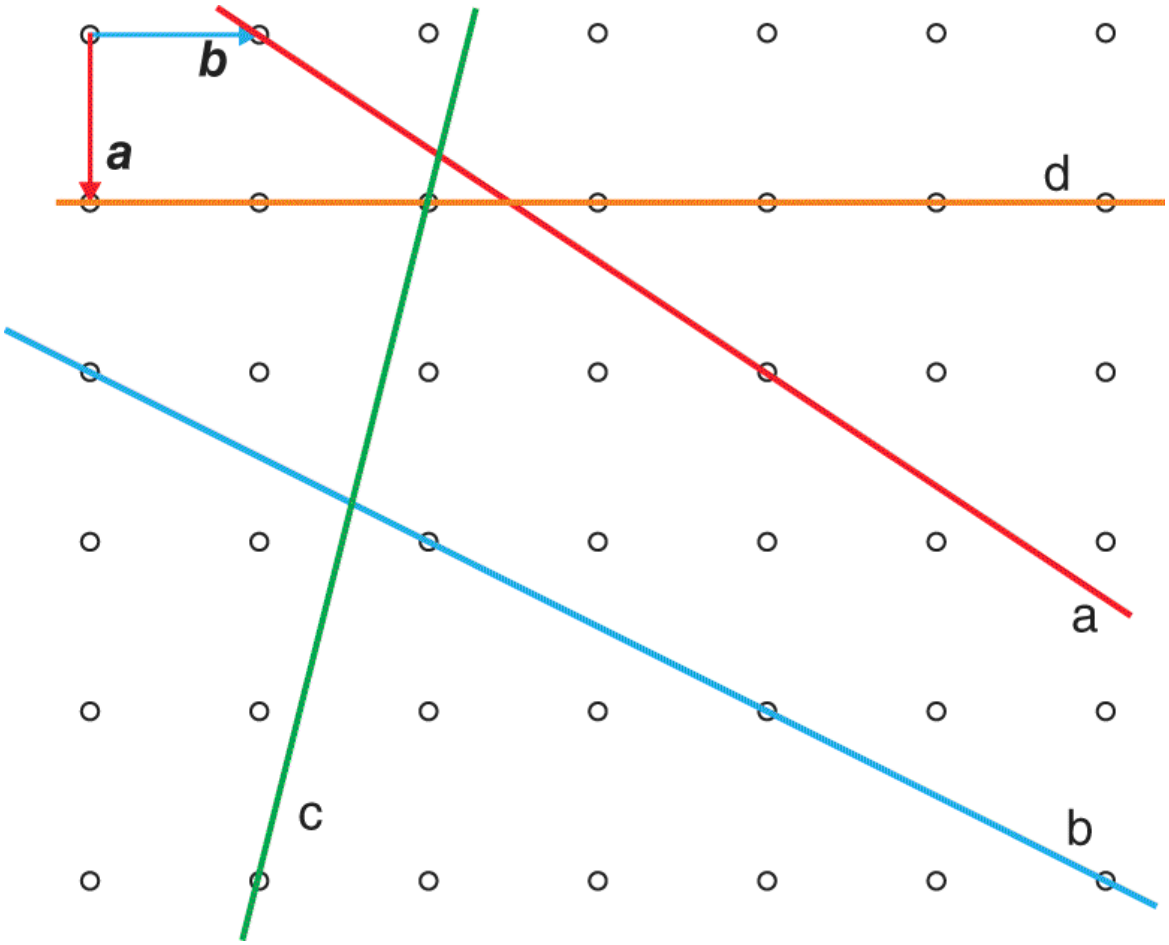
3.1 Index the planes (a)–(d) shown in the figure.



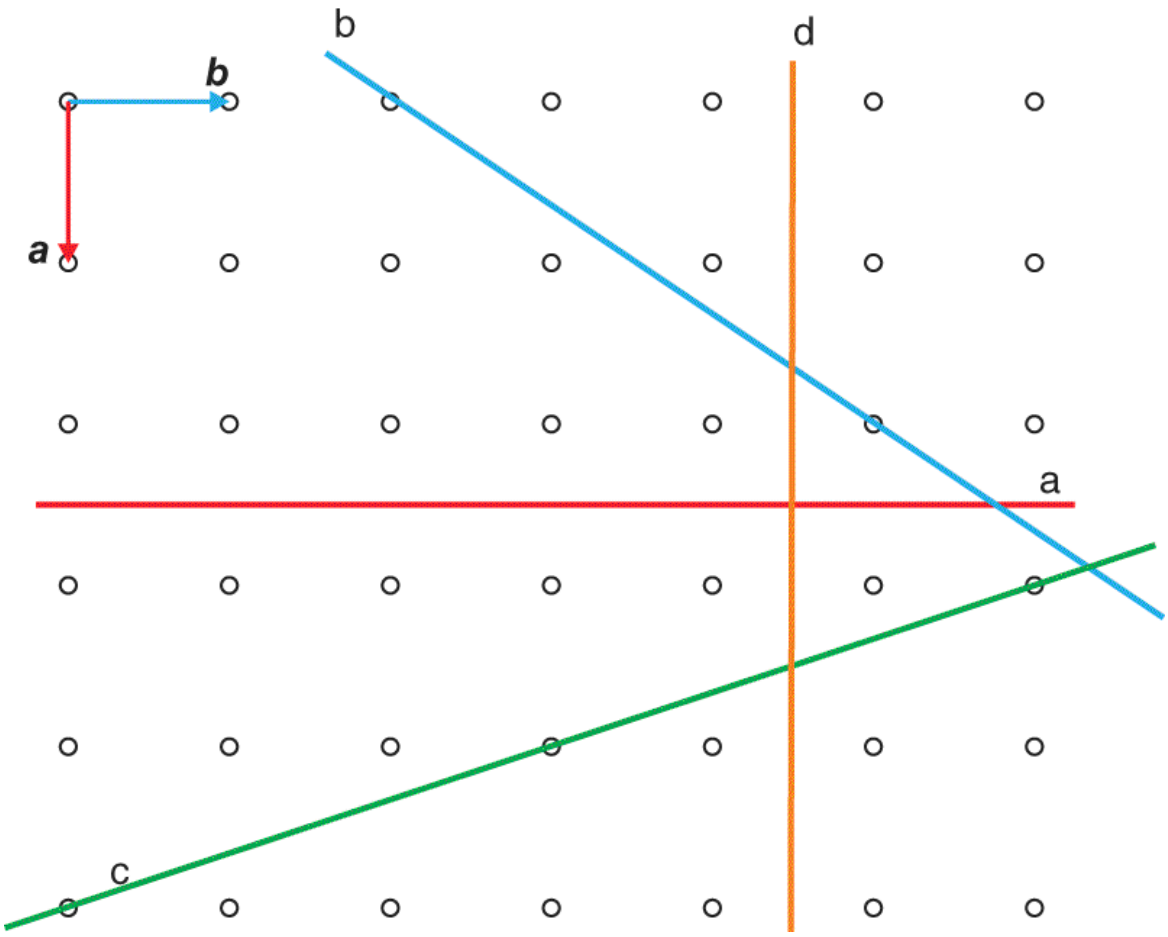
3.2 Index the planes (a)–(d) shown in the figure.



3.3 Index the planes (a)–(d) shown in the figure.



3.4 Index the planes (a)–(d) shown in the figure.



3.5 Sketch the (111) and $(1\bar{1}1)$ planes in a cubic crystal.

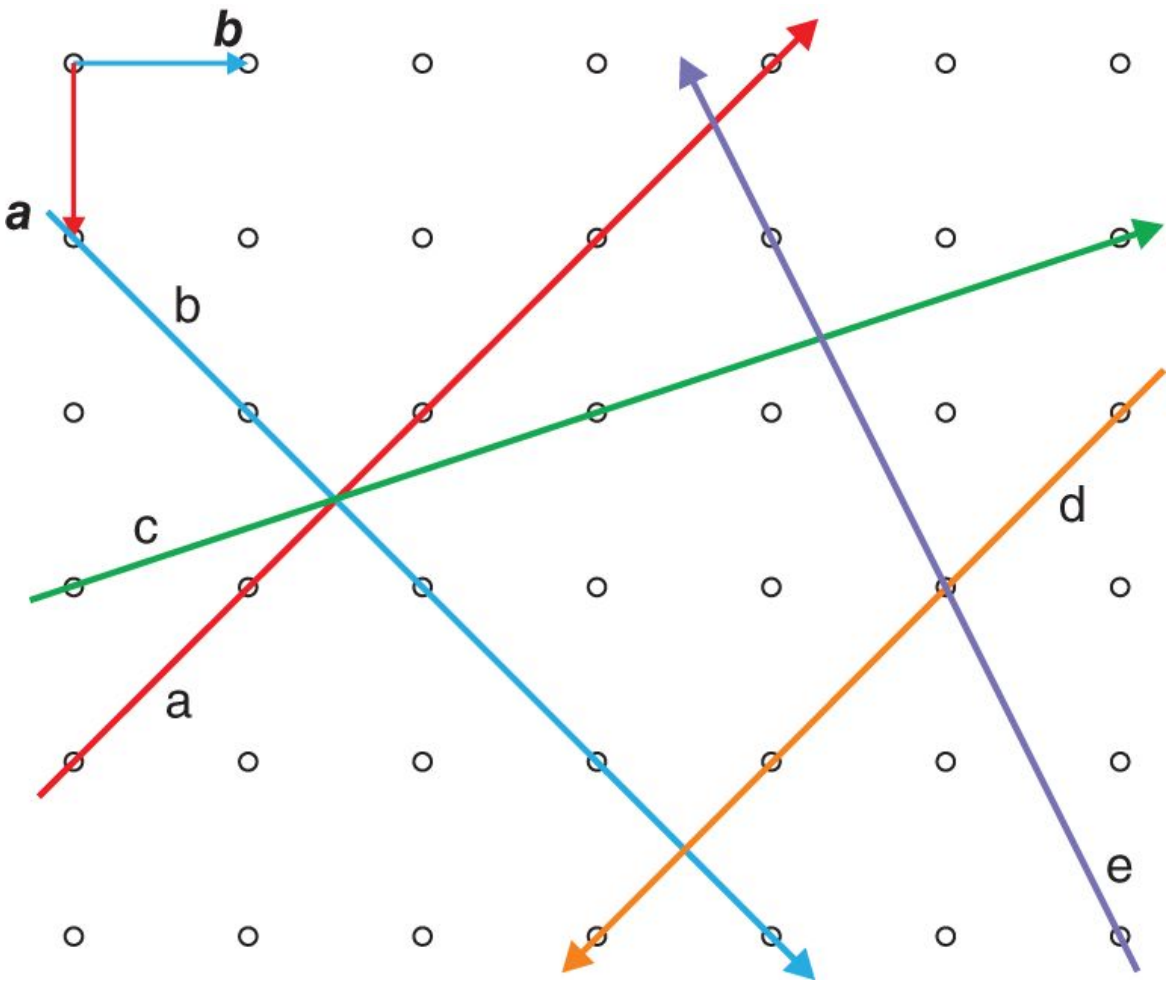
3.6 List the planes belonging to the {110} set in a cubic crystal.

3.7 List the planes belonging to the {111} set in a cubic crystal.

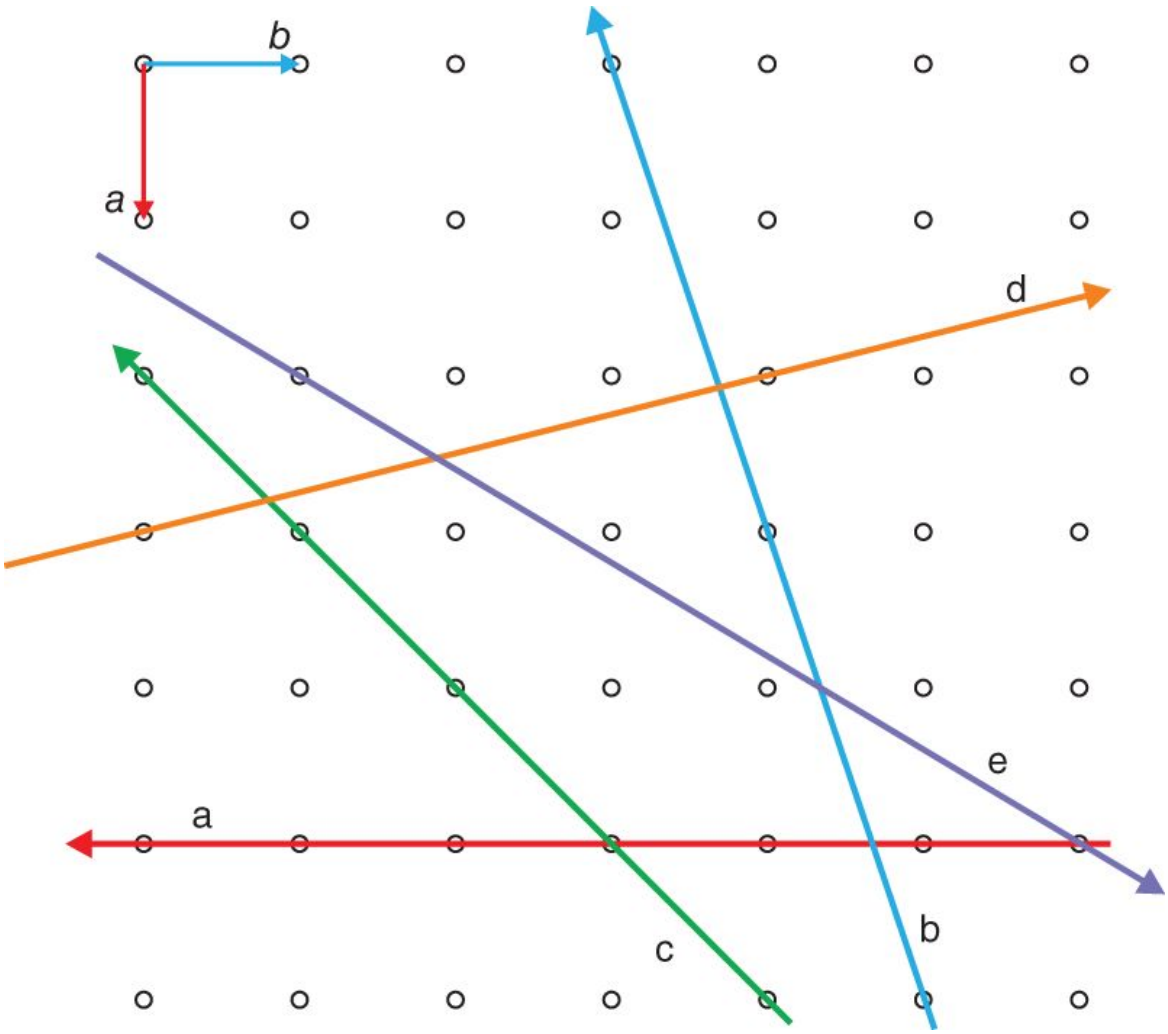
3.8 List the planes belonging to the {*hh*0} set in a cubic crystal.

3.9 List the planes belonging to the {*hk*0} set in a cubic crystal.

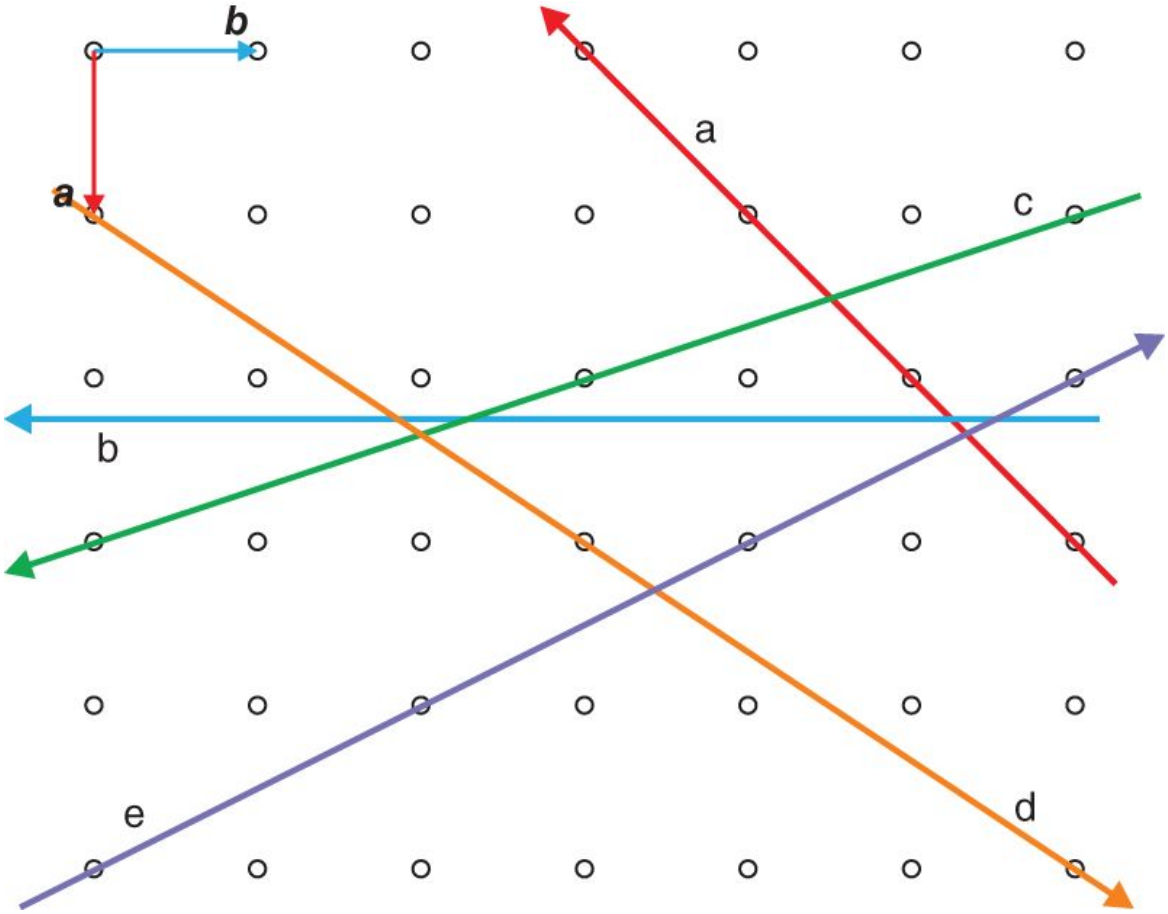
3.10 Index the directions (a)–(e) shown in the figure.



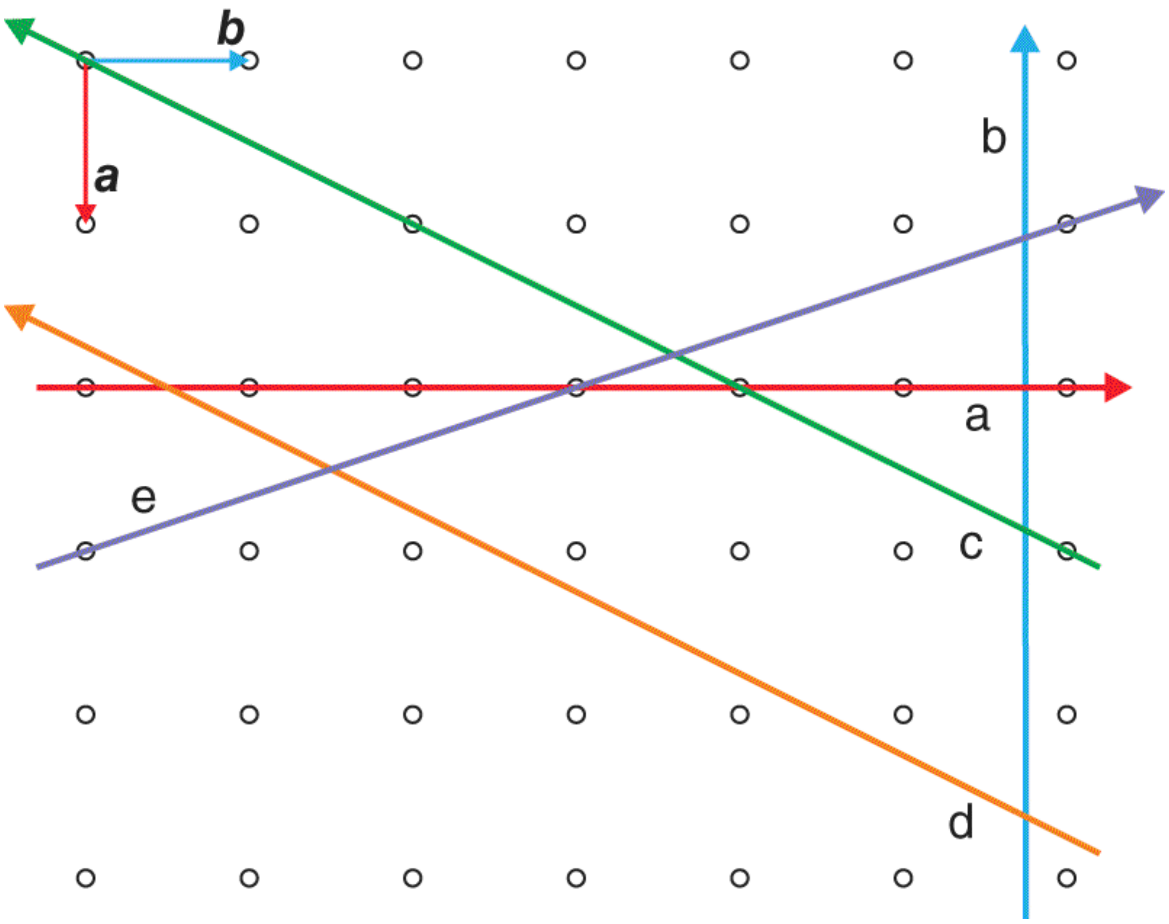
3.11 Index the directions (a)–(e) shown in the figure.



3.12 Index the directions (a)–(e) shown in the figure.



3.13 Index the directions (a)–(e) shown in the figure.



3.14 List the directions belonging to the ⟨100⟩ set.

3.15 List the directions belonging to the ⟨110⟩ set.

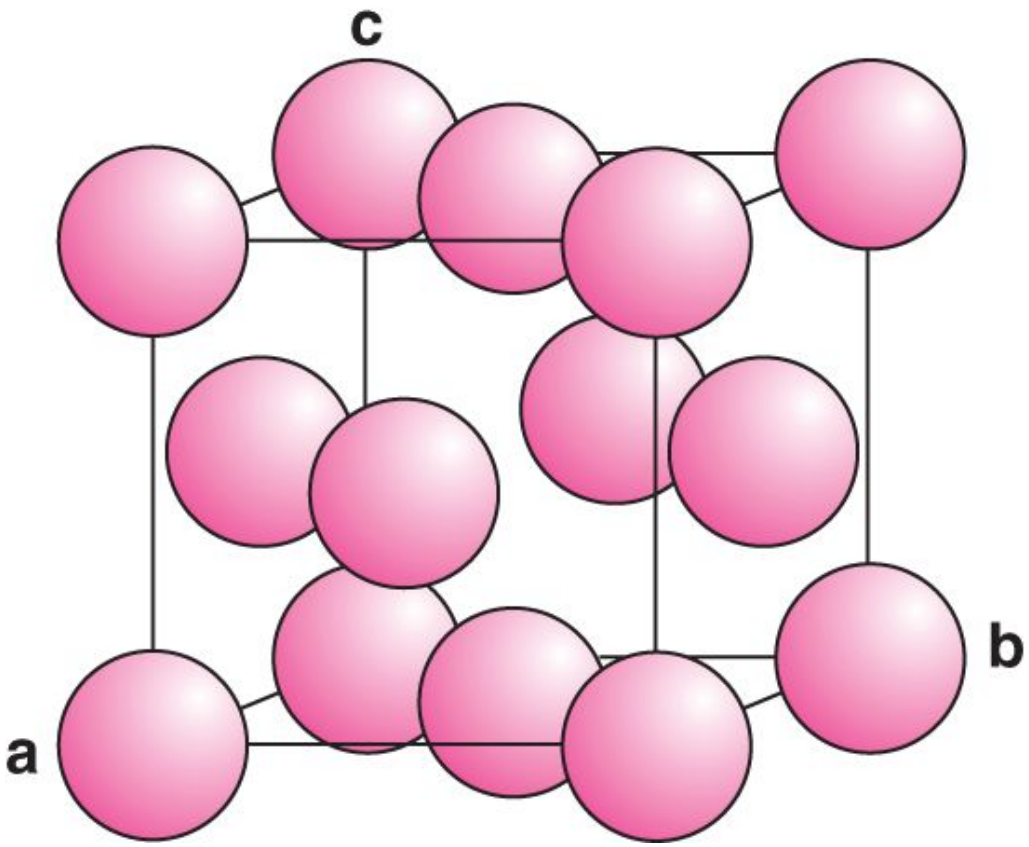
3.16 What is the angle between (110) and [110] in a cubic crystal?

3.17 What is the angle between (132) and [132] in a cubic crystal?

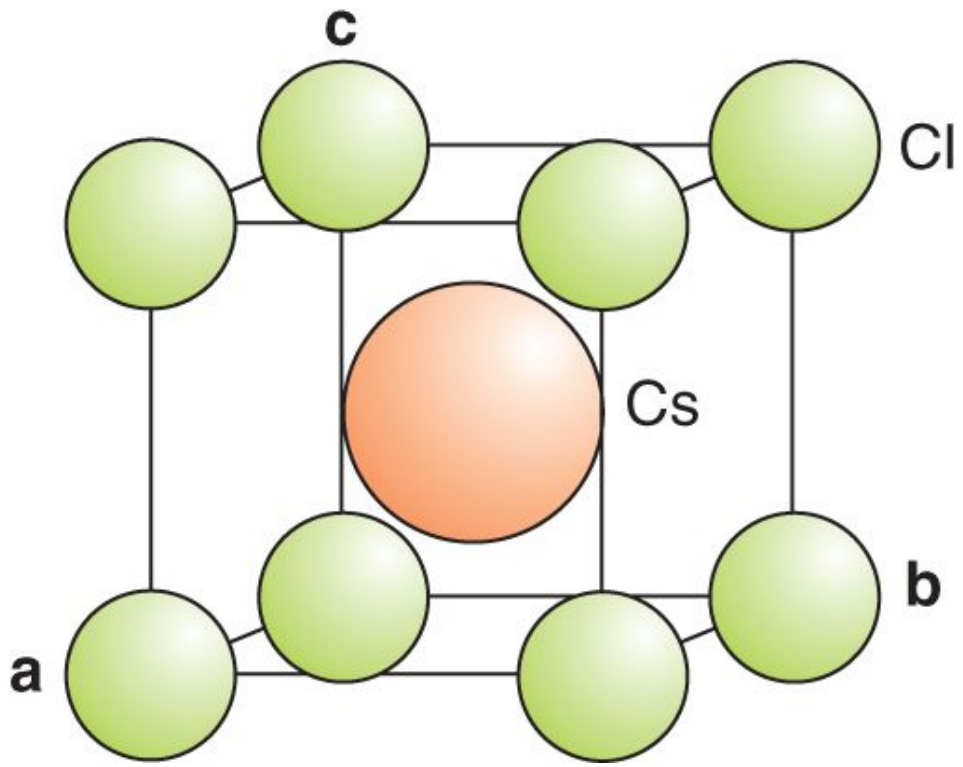
3.18 Sketch the reciprocal lattice of a cubic crystal with $a = 5$ nm.

3.19 The unit cell size of CaO is 0.481 05 nm and that of SrO is 0.516 02 nm. Both adopt the halite (NaCl) structure-type. Estimate the composition of a crystal of formula $Sr_xCa_{1-x}O$, which was found to have a unit cell of 0.5003 nm.

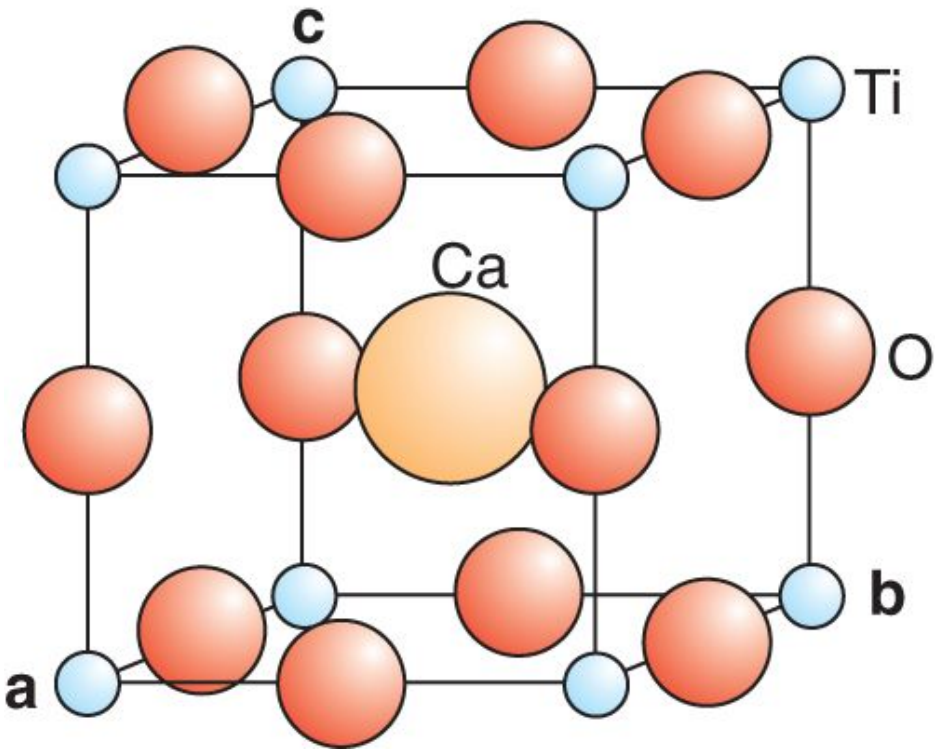
3.20 The cubic unit cell of iridium is drawn in the figure. What are the atomic coordinates? What is the unit cell type?



3.21 The cubic unit cell of CsCl is drawn in the figure. What are the atomic coordinates of each ion?



3.22 The cubic unit cell of perovskite, $CaTiO_3$, is drawn in the figure. What are the atomic coordinates of the atoms?



3.23 The atom positions in cubic nickel oxide are:

Ni : 0, 0, 0 ½, ½, 0 0, ½, ½ ½, 0, ½

O : ½, ½, ½ 0, 0, ½ ½, 0, 0 0, ½, 0

Sketch the unit cell. What is the formula of the oxide? What is the structure type?

3.24 A copper gold alloy has a cubic structure. The atom positions are:

Au : 0, 0, 0

Cu : 0, ½, ½ ½, 0, ½ ½, ½, 0

Sketch the unit cell and determine the formula of the alloy.

3.25 Aluminium has the face-centred cubic structure with a lattice parameter of 0.4050 nm. Estimate the density of the metal.

3.26 Tungsten has the body-centred cubic structure with a lattice parameter of 0.316 51 nm. Estimate the density of the metal.

3.27 Magnesium has the hexagonal structure with lattice parameters of a = 0.320 nm, c = 0.520 nm. Estimate the density of the metal.

3.28 Copper has the face-centred cubic structure and a density of 8.96×10^3 kg m^{-3}. What is the length of the unit cell edge?

3.29 A sample of calcia stabilised zirconia, prepared by heating 85 mol% ZrO_2 with 15 mol% CaO at 1600 °C. The material had a cubic unit cell with a lattice parameter, a, of 0.5144 nm and a measured density of 5485 kg m^{-3}. Calculate the theoretical density of the sample and hence determine whether interstitials or vacancies are more likely to be present in the structure. (The parent structure is fluorite (CaF_2), in which each cubic unit cell contains four metal positions and eight non-metal positions, to give an overall composition of MX_2.)

3.30 The unit cell of a zirconium sulfide, with a measured composition of 77 Zr:100 S, is of the halite (NaCl) type, with a = 0.514 nm. The measured density is 4.80×10^3 kg m^{-3}. Calculate the theoretical density of ideal ZrS with the halite structure, and give an opinion on the defect structure of the real material.

3.31 Will the density of a crystal go up or down if it contains: (a) Schottky defects; (b) Frenkel defects; (c) vacancies; (d) interstitials?

3.32 An iron titanium oxide has the anions in a hexagonal close-packed array. The Fe and Ti atoms each occupy 1/3 of the octahedral sites available in an ordered array. What is the formula of the oxide? What is the likely parent structure of the oxide?

3.33 The enthalpy of formation of vacancies in pure nickel is $\Delta H = 97.3\ \text{kJ mol}^{-1}$. What is the fraction of sites vacant at 1100 °C?

3.34 The enthalpy of formation of vacancies in pure copper is $\Delta H = 86.9\ \text{kJ mol}^{-1}$. What is the fraction of sites vacant at 1084 °C?

3.35 The enthalpy of formation of vacancies in pure gold is $\Delta H = 123.5\ \text{kJ mol}^{-1}$. The density of gold is 19 281 kg m^{-3}. How many atom positions are vacant at 1000 °C?

3.36 The enthalpy of formation of vacancies in pure aluminium is $\Delta H = 72.4\ \text{kJ mol}^{-1}$. The density of aluminium is 2698 kg m^{-3}. How many atom positions are vacant at 600 °C?

3.37 Calculate how the fraction of Schottky defects in a crystal of KCl varies with temperature if ΔH_S is 244 kJ mol^{-1}.

3.38 Calculate the number of Schottky defects in a crystal of KCl at 800 K. The cubic unit cell of this material has a cell edge of 0.629 nm. Each unit cell contains 4 K^+ and 4 Cl^- ions.

3.39 The enthalpy of formation of a Frenkel defect in AgBr, is 1.81×10^{-19} J. Estimate the fraction of interstitial silver atoms due to the Frenkel defect formation in a crystal of AgBr at 300 K.

3.40 AgBr has cubic unit cell with an edge of 0.576 nm. There are four Ag atoms in the unit cell, and assume that there are four interstitial positions available for Ag atoms. Calculate the absolute number of interstitial defects present per cubic metre at 300 K.

3.41 Calculate the enthalpy of formation of Frenkel defects in NaBr, using the data on the number of defects present given in the following table.

Temperature (K)	n_F (m^{-3})
200	1.428×10^{2}
300	7.257×10^{10}
400	1.636×10^{15}
500	6.693×10^{17}

600	3.687×10^{19}
700	6.468×10^{20}
800	5.538×10^{21}
900	2.943×10^{22}

3.42 The energy of formation of Schottky defects in a crystal of CaO is given as 6.1 eV. Calculate the number of Schottky defects present in CaO at 1000 and 2000 °C. How many vacancies are present at these temperatures? CaO has a density of 3300 kg m^{-3}.

3.43 Calculate the number of Schottky defects in a crystal of MgO, at 1500 °C if ΔH_S is 96.5 kJ mol^{-1} and the density of MgO is 3580 kg m^{-3}.

3.44 Calculate the number of Frenkel defects present in a crystal of AgCl, at 300 K, given that the material has a cubic unit cell of edge 0.555 nm that contains four Ag atoms. Assume that the interstitial atoms occupy any of eight tetrahedral sites in the unit cell. ΔH_F is 2.69×10^{-19} J.

3.45 Calculate the number of vacancies in a crystal of NiO containing Schottky defects at 1000 °C, given that ΔH_S is 160 kJ mol^{-1}, and the density is 6670 kg m^{-3}.

3.46 The fraction of Schottky defects in NiO, at 1000 °C is 1.25×10^{-4}. The cubic unit cell contains four Ni atoms, and has a cell edge of 0.417 nm. Calculate the number of nickel vacancies present.

3.47 The number of Schottky defects in LiF, which has a cubic unit cell containing four Li and four Cl atoms, with a cell-edge of 0.4026 nm, is 1.12×10^{22} m^{-3} at 600 °C. Calculate the energy of formation of these defects.

3.48 9 mol% of Y_2O_3 is mixed with 91 mol% ZrO_2 and heated until a uniform product with high oxygen ion conductivity is obtained. The resulting crystal is a stabilised zirconia with the formula $Y_xZr_yO_z$. Determine x, y, and z.

3.49 CaO forms a solid solution with Bi_2O_3 to give a material with a high anionic conductivity. If 10 mol% CaO is reacted with 90 mol% Bi_2O_3;

a. What is the formula of the final solid;

b. What numbers and types of vacancies have been created?

3.50 What defects will form in the crystals made by adding small amounts of compound a to compound b.

a. a. LiBr, b. $CaBr_2$.

b. a. $CaBr_2$, b. LiBr.

c. a. MgO, b. Fe_2O_3.

d. a. MgO, b. NiO.

3.51 What defects will form in the crystals made by adding small amounts of compound a to compound b.

a. a. $CdCl_2$, b. NaCl.

b. a. NaCl, b. $CdCl_2$.

c. a. Sc_2O_3, b. ZrO_2.

d. a. ZrO_2, b. HfO_2.

3.52 Show that the number of metal atom sites N in a crystal of composition MX is given by:

$$N = \rho N_A / \text{molar mass of MX}$$

where ρ is the density of MX and N_A is the Avogadro constant.

3.53 Sodium chloride has a density of 2165 kg m^{-3}. The unit cell, which is cubic, with a cell of edge 0.563 nm, contains four Na and four Cl atoms. Calculate the number of atoms of Na per cubic metre using (a) density; (b) unit cell data.

Notes

1 Vectors are set in **bold** typeface throughout.

2 The unit cell dimensions are often specified in terms of the Ångström unit, Å. 10 Å = 1 nm.

3 This description originally applied to the same external form of the crystals rather than the internal arrangement of the atoms.

4 The data is from the classical paper: E. R. Jette and F. Foote (1933).

4
Solids: Overview

- What factors are important in predicting substitutional alloy formation?
- How is strengthened glass made?
- Why are plastic bags hard to degrade?

This chapter contains an overview of some of the distinguishing structural characteristics of the important classes of solids: metals, silicates and ceramics, polymers and organic materials, and composite materials. Metallic materials tend to have uniform structures that can often be described in terms of the packing of spherical atoms. Ceramics show no such homogeneity of structure, varying from uniform crystals to non-crystalline glasses. Organic solids are just as diverse. Polymers, commonly called plastics, and organic framework structures serve as a broad introduction, both to structures and properties of organic materials. Finally, many important solids are not simply composed of one of the preceding types, but are combinations of various solids, known as composite materials. Reinforced concrete, universally used in building and related urban infrastructure, is but one example.

4.1 Metals

4.1.1 Structures

Roughly speaking, about three quarters of the elements can be regarded as metallic, defined by being ductile, shiny (silvery), giving specular (mirror-like) reflection of light, and being good conductors of electricity. These properties (Described in more detail in Sections, and 12.4) arise because the outer electrons on the individual metal atoms in the solid state are delocalised throughout the structure, so that the atoms that make up the crystal pack together like spheres. In this respect, the majority of metals display one of only three structures at room temperature and standard conditions: face-centred cubic (fcc, A1) (Figure 3.5), body-centred cubic (bcc, A2) (Figure 3.6), and hexagonal close-packed (hcp, A3) (Figure 3.7). The normal room temperature structures of the metals are summarised in Figure 4.1. The difference in energy between these structures is small, and it is surprising that few complex arrangements occur. The major anomaly is mercury, which is the only liquid metal under normal conditions.

Li A2 *a* 0.3509	Be A3 *a* 0.2286 *c* 0.3585												
Na A2 *a* 0.4291	Mg A3 *a* 0.3209 *c* 0.5211											Al A1 *a* 0.4050	
K A2 *a* 0.5321	Ca A1 *a* 0.5321	Sc A3 *a* 0.3309 *c* 0.5268	Ti A3 *a* 0.2951 *c* 0.4686	V A2 *a* 0.3024	Cr A2 *a* 0.2885	Mn	Fe A2 *a* 0.2867	Co **(A3)** *a* 0.2507 *c* 0.4069	Ni A1 *a* 0.3524	Cu A1 *a* 0.3615	Zn A3 *a* 0.2665 *c* 0.4947	Ga	
Rb A2 *a* 0.5705	Sr A1 *a* 0.6084	Y A3 *a* 0.3648 *c* 0.5732	Zr A3 *a* 0.3232 *c* 0.5148	Nb A2 *a* 0.3300	Mo A2 *a* 0.3147	Tc A3 *a* 0.2738 *c* 0.4393	Ru A3 *a* 0.2706 *c* 0.4282	Rh A1 *a* 0.3803	Pd A1 *a* 0.3890	Ag A1 *a* 0.4086	Cd A3 *a* 0.2979 *c* 0.5620	In (A2) *a* 0.3298 *c* 0.5063	Sn
Cs A2 *a* 0.6141	Ba A2 *a* 0.5023	La (A3) *a* 0.3772 *c* 1.2159	Hf A3 *a* 0.3195 *c* 0.5051	Ta A2 *a* 0.3303	W A2 *a* 0.3165	Re A3 *a* 0.2761 *c* 0.4458	Os A3 *a* 0.2734 *c* 0.4392	Ir A1 *a* 0.3839	Pt A1 *a* 0.3924	Au A1 *a* 0.4078	Hg	Tl A3 *a* 0.3457 *c* 0.5525	Pb A1 *a* 0.4950

Figure 4.1 The crystal structures and unit cell parameters (nm) of the metallic elements with the fcc (A1), bcc (A2), and hcp (A3) structure, at 25 °C and atmospheric pressure.

Cobalt is one of the few metals to show appreciable disorder. Below about 435 °C, the structure is a random stacking of A, B, and C planes of metal atoms. It can be transformed into the hexagonal A3 structure by careful annealing at lower temperatures. However, the structure found also depends upon the size of the crystallites. Very small grains, in a sponge-like matrix, possess a face-centred cubic structure, while larger grains are hexagonal close packed.

The normal form of manganese, α-Mn, has a cubic structure that is not directly comparable with the face-centred or body-centred structures. The structure of gallium is also more complex. In the orthorhombic unit cell, the Ga atoms are

arranged in pairs. The structure of indium metal consists of a slightly elongated hexagonal close-packed structure, with a tetragonal unit cell.

Tin has two well-known forms. Grey tin, α-Sn, stable below 13.2 °C has the face-centred cubic structure of diamond, Si and Ge (Figure 3.8). It is brittle and a non-metal. At temperatures above 13.2 °C, the stable structure is β-Sn (white tin), which is the metallic form of tin. The transition from white to grey tin is slow, and the metallic form is stabilised by metallic impurities, so that tin is normally found in the metallic form. Although white tin is metallic, it has a tetragonal unit cell, not directly comparable with the simple face-centred cubic, body-centred cubic, or hexagonal structures described earlier.

The lanthanoid metals lanthanum, praseodymium, and neodymium adopt a double hexagonal structure that has an ...ABAC... repeat, while samarium has a packing repeat ...BABCAC...

Changes in the metal structures are commonly induced by variation of temperature and pressure; the different forms being called *allotropes*. Some of these allotropic forms are listed in Table 4.1.

Table 4.1 Allotropic structures of some metals.

Element	Room temperature structure	High temperature structure	Transition temperature (°C)
Calcium, Ca	fcc, A1	bcc, A2	445
Strontium, Sr	fcc, A1	bcc, A2	527
Scandium, Sc	hcp, A3	bcc, A2	1337
Titanium, Ti	hcp, A3	bcc, A2	883
Zirconium, Zr	hcp, A3	bcc, A2	868
Hafnium, Hf	hcp, A3	bcc, A2	1742
Ytterbium, Yb	hcp, A3	bcc, A2	1481
Iron, Fe	bcc, A2	fcc, A1	912
Cobalt, Co	(hcp, A3)	fcc, A1	435

4.1.2 Metallic Radii

If we assume that the structures of metals are made up of touching spherical atoms, it is quite easy, knowing the size of the unit cell, to work out metallic radii (Figure 4.2). The relationship between the cell dimensions and the radius of the component atoms, *r*, for the three common metallic structures are given.

Face-centred cubic, A1: The atoms are in contact along a cube face diagonal, so that:

$$r = a/2\sqrt{2} = a/\sqrt{8}$$

The separation of the close packed (111) atom planes (along a cube body diagonal) is:

$$d_{111} = a/\sqrt{3} = \sqrt{8}r/\sqrt{3} \approx 1.633\,r$$

Body-centred cubic, A2: The atoms are in contact along a cube body diagonal, which is equal to $4r$ and $\sqrt{3}a$, so that:

$$r = \sqrt{3}a/4$$

Hexagonal close packed, A3: The atoms are in contact along the **a**-axis, hence:

$$r = a/2$$

Li 0.1562	Be 0.1128												
Na 0.1911	Mg 0.1602											Al 0.1432	
K 0.2376	Ca 0.1974	Sc 0.1641	Ti 0.1462	V 0.1346	Cr 0.1282	Mn 0.1264	Fe 0.1274	Co 0.1252	Ni 0.1246	Cu 0.1278	Zn 0.1349	Ga 0.1411	
Rb 0.2546	Sr 0.2151	Y 0.1801	Zr 0.1602	Nb 0.1468	Mo 0.1400	Tc 0.1360	Ru 0.1339	Rh 0.1345	Pd 0.1376	Ag 0.1445	Cd 0.1568	In 0.1663	Sn 0.1545
Cs 0.2731	Ba 0.2243	La 0.1877	Hf 0.1580	Ta 0.1467	W 0.1408	Re 0.1375	Os 0.1353	Ir 0.1357	Pt 0.1387	Au 0.1442	Hg 0.1510	Tl 0.1716	Pb 0.1750

Figure 4.2 Metallic radii, for 12-coordination.

The numerical value of the radius is found to depend upon the number of nearest neighbours (the coordination number, CN, of the atom in question). Atoms in both the fcc, A1 and hcp, A3 structures have 12 nearest neighbours (CN 12), and the radius determined will be appropriate to that coordination. Atoms

in the bcc, A2 structure have eight nearest neighbours (CN 8), and it is necessary to convert the radii measured with respect to this structure into those appropriate to 12 coordination in order to obtain a self-consistent set of values. The conversion can be made using the empirical formula

$$\text{radius [CN 12]} = 1.032\{\text{radius [CN 8]}\} - 0.0006$$

when the radii are measured in nm.

The metallic radii of the elements Mn, Ga, and Sn, with more complex structures, and of Hg, do not allow for the application of the simple rules earlier. The radii for these elements are derived from a comparison of the interatomic distances in alloys with appropriate structures.

There are a number of trends to note. In the well-behaved alkali metals and alkaline earth metals, the radius of an atom increases smoothly as the atomic number increases. The d transition-metals all have rather similar radii as one moves along the period, and these generally increase with atomic number going down a group. The same is true for the lanthanoids and actinoids.

4.1.3 Alloy Solid Solutions

Alloys are important because they often show superior physical properties compared to the pure elements. There are large numbers of alloys, many of which have unusual and complex structures. Here we will mention only two sorts of alloy, both with structures closely related those of the pure metals. *Substitutional solid solutions* have a structure identical to one of the metals involved, the parent structure. The alloy forming (or foreign) atoms simply occupy normally occupied positions in the parent structure. *Interstitial solid solutions* are formed when very small foreign atoms enter the parent structure and sit in normally unoccupied interstitial positions.

The likelihood of forming a substitutional solid solution between two metals will depend upon a variety of chemical and physical properties. A large number of alloy systems were investigated by Hume-Rothery, in the first part of the last century, with the aim of understanding the principles that controlled alloy formation. His findings with respect to substitutional solid solution formation are summarised in the empirical *Hume-Rothery solubility rules*. The likelihood of obtaining a solid solution between two metals is highest when:

1. The crystal structure of each element of the pair is identical.
2. The atomic sizes of the atoms do not differ by more than 15%.
3. The elements do not differ greatly in electronegativity (or else they will form compounds with each other). This implies that they should be near to each other in the Periodic Table.

4. The elements have the same valence. This implies that they should lie in the same Group of the Periodic Table.

Although formulated a century ago, these guidelines remain of value for modelling substitutional alloy formation. The rules predict, for example, that copper–nickel, and copper–gold should form extensive substitutional solid solutions. What they do not predict is the likelihood that the atoms in the solid solution will order. In such cases a new phase, an *ordered solid solution*, will form. This happens in many systems, especially when random substitutional solid solutions are annealed (that is, heated at lower temperatures for some time). For example, a copper–gold alloy held at temperatures close to the melting point (about 890 °C) and then rapidly cooled will have a random distribution of copper and gold atoms over the sites of the face-centred cubic structure (Figure 4.3a). The same material annealed at a temperature of about 400 °C produces the ordered phases Cu_3Au and CuAu. The structure of the copper-rich alloy phase, Cu_3Au (Figure 4.3b) has the gold atoms located at the corners of a cubic unit cell and the copper atoms at the face centres. The ordered structure of the CuAu alloy (Figure 4.3c) has alternating (100) planes composed of either copper or gold atoms. This structure is called the CuAu I structure. Arrangements that are more complex also occur. For example, in the CuAu II structure (Figure 4.3d), a regular set of antiphase boundaries spaced every five CuAu unit cells gives an overall repeat along the **c**-axis of approximately 10 times the cubic cell parameter.

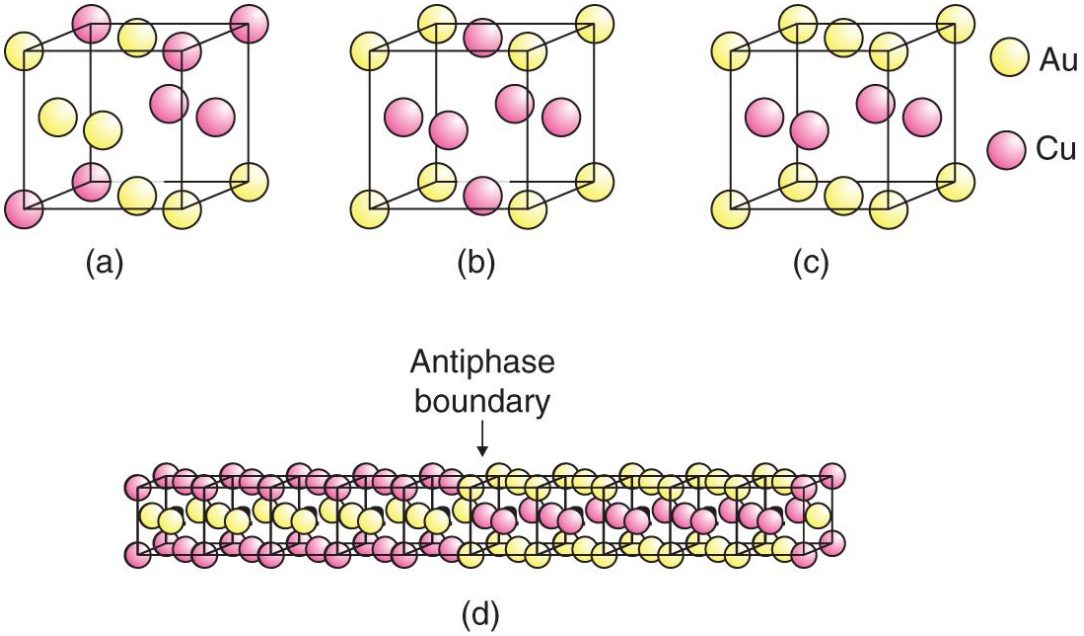


Figure 4.3 The crystal structure of: (a) disordered CuAu; (b) ordered Cu_3Au; (c) ordered CuAu I; (d) ordered CuAu II.

Just as atoms should ideally be of similar sizes to form extensive substitutional solid solutions, so to form an interstitial solid solution, in which atoms fit between the parent atoms in the structure, the radius of the foreign atom should be less than about half of the atomic radius of the parent atoms. Traditionally, the interstitial alloys most studied are those of the transition metals with carbon and nitrogen, as the addition of these atoms to the crystal structure increases the hardness of the metal considerably. Steel remains the most important traditional interstitial alloy from a world perspective. It consists of carbon atoms distributed at random in interstitial sites within the face-centred cubic structure of iron to form the phase *austenite* (Figure 2.18). More recently, hydrogen storage has become important, and interstitial alloys formed by incorporation of hydrogen into metals are of considerable interest.

The sites that are available for foreign atoms in interstitial alloys are of tetrahedral or octahedral geometry. In both the face-centred cubic or hexagonal close-packed structures there are twice as many tetrahedral sites and the same number of octahedral sites present as metal atoms (see Figure 3.25). In terms of the conventional face-centred cubic unit cell, the octahedral sites lie at the mid-points of each of the cell edges, with a further site at the cell centre (Figure 4.4a). Each site on a cell edge is shared by four cells, and there are 12 cell edges, hence each cell contains ¼ × 12 sites at the edges plus one at the cell centre, making four in total. The tetrahedral sites are found in each quarter of the unit cell, making eight in all (Figure 4.4b). The hexagonal close-packed unit cell contains four tetrahedral sites and two octahedral sites (Figure 4.5a,b). In both structures, all of the tetrahedral sites and all of the octahedral sites have identical geometries.

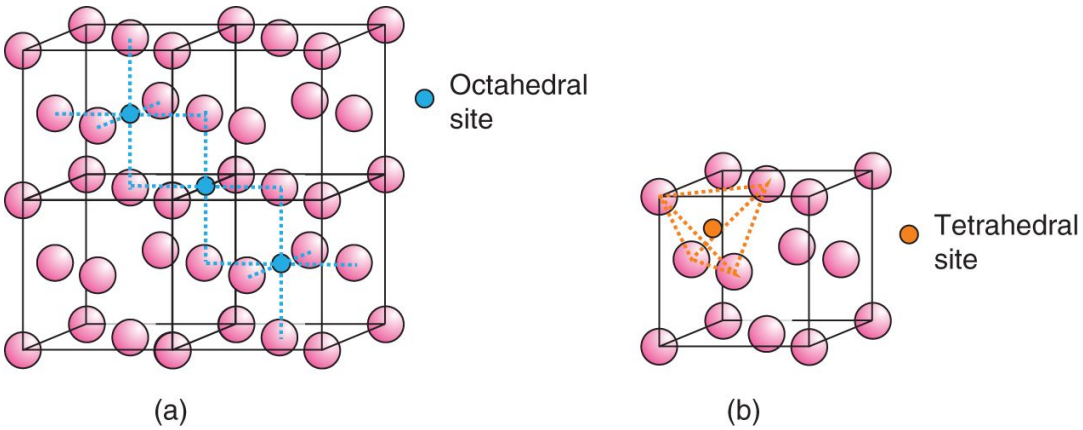


Figure 4.4 Octahedral (a) and tetrahedral (b) sites in the face-centred cubic structure. Not all equivalent sites are marked.

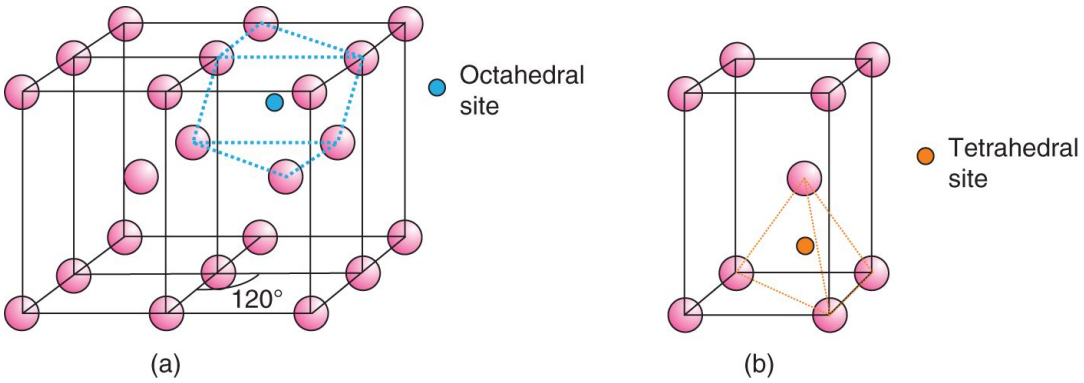


Figure 4.5 Octahedral (a) and tetrahedral (b) sites in the hexagonal close-packed structure. Not all equivalent sites are marked.

The body-centred cubic structure also has 12 tetrahedral and 6 octahedral sites available. The octahedral sites lie on all of the cube faces (Figure 4.6a), and the tetrahedral sites, also on the cube face, are slightly below them (Figure 4.6b). These octahedral and tetrahedral sites are of slightly different geometry than those in the face-centred and hexagonal structures and in addition, there are several different geometries found for the tetrahedral positions. In all three structures the octahedral sites are larger and can accommodate both carbon and nitrogen atoms. The tetrahedral sites are smaller and only hydrogen commonly uses these positions.

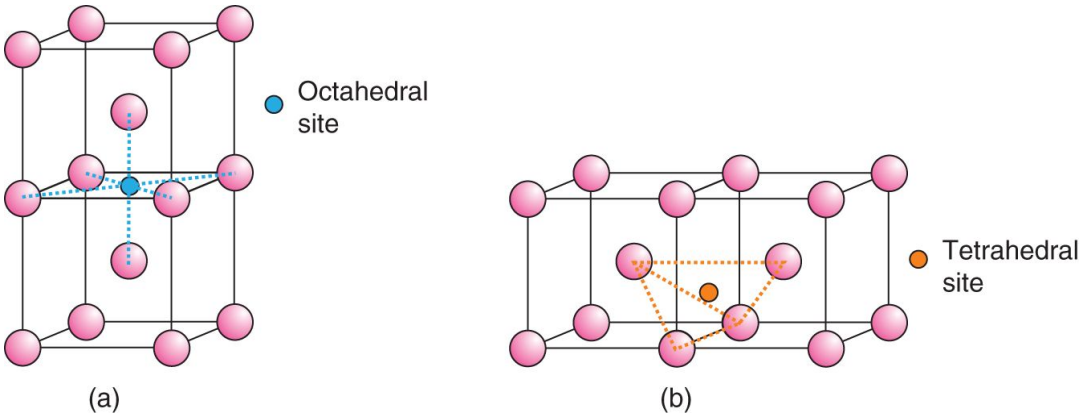


Figure 4.6 Octahedral (a) and tetrahedral (b) sites in the body-centred cubic structure. Not all equivalent sites are marked.

The process by which interstitial alloys form is similar in all systems. A reactive gas, typically hydrogen for hydrides, methane for carbides or ammonia for nitrides, decomposes on the metal surface. The atoms formed can then enter the structure, to occupy sites at random. The phases formed are often called α-phases. Continued reaction leads to the formation of new structures, either by the ordering of the impurity atoms, as described for substitutional alloys, or by more extensive structural rearrangements, as in cementite, Fe_3C.

Although the size of the foreign atoms is of importance, the sites that are occupied, and the degree of occupancy of the available sites also depends critically on chemical interactions between the species. For example, in the α-phase, NbH_x, only a fraction of the tetrahedral sites, with a particular geometry, are occupied. Moreover, the limiting composition of the α-phase, approximately $NbH_{0.1}$, is achieved when only a fraction of these particular sites is filled. At higher hydrogen concentrations chemical interactions lead to the nucleation of the hydride NbH.

4.1.4 Metallic Glasses and Quasicrystals

Metallic glasses are of interest because they have a number of superior mechanical properties compared with polycrystalline metals; in particular, high strength and corrosion and wear resistance. However, metals have simple crystal structures that readily nucleate as melts are cooled, so that in order to make a metallic glass by cooling, it is necessary to disrupt the natural tendency of the liquid to crystallise. This can be achieved using the following strategies. Firstly, a system having a deep eutectic point is helpful, because then the liquid can cool significantly before solidification starts. Secondly, a system containing three or more metals, each of which crystallises in a different structure, will deter crystallisation. Finally, mixtures of atoms with widely differing sizes crystallise less readily. These considerations suggest a general formula for metallic glasses of $T_{70-90}(SM)_{0-15}(NM)_{10-30}$, where T is one or more transition metals or lanthanoids (SM) is one or more semimetals such as Si or Ge and (NM) is one or more non-metals such as P or C.

The earliest metallic glasses were produced by squirting a jet of liquid metal against a rapidly spinning copper disc cooled with liquid nitrogen, to achieve a cooling rate of approximately 10^5–10^6 K s^{-1}. The first non-crystalline metallic material to be made in this way had a composition of $Au_{75}Si_{25}$. Ribbons of metallic glass were explored in the 1970s and 1980s as soft magnetic transformer cores. A widely used magnetic material for this purpose is METGLAS, with a composition of $Fe_{40}Ni_{40}P_{14}B_6$. Other examples of transformer core materials include $Fe_{86}B_8C_6$ and $FeB_{11}Si_9$.

At the end of the 1980s, studies of more complex systems made it possible to fabricate glassy metals using much slower cooling rates, down to 10 K s^{-1}. This opened the door to the production of glassy metals in bulk forms via traditional casting methods. Bulk glasses have unique chemical and physical properties that make them superior to polycrystalline alloys for many applications. Glasses with complex formulae such as $Fe_{72}Al_5Ga_2P_{11}C_6B_4$ find applications as magnetic materials. Glasses with superior mechanical properties are typified by a group of alloys known as Vitraloy (or Vitralloy), with compositions close to $Zr_{46.75}Ti_{8.25}Cu_{7.5}Ni_{10}B_{27.5}$, used, among other things, for golf-club heads.

Related alloys in the Fe-Cr-Mo-C-B system (Figure 4.7) are used as hard surface coatings as initial oxidation results in a corrosion resistant thin oxy-hydroxide film with an overall composition $Cr_x(OH)_y$.

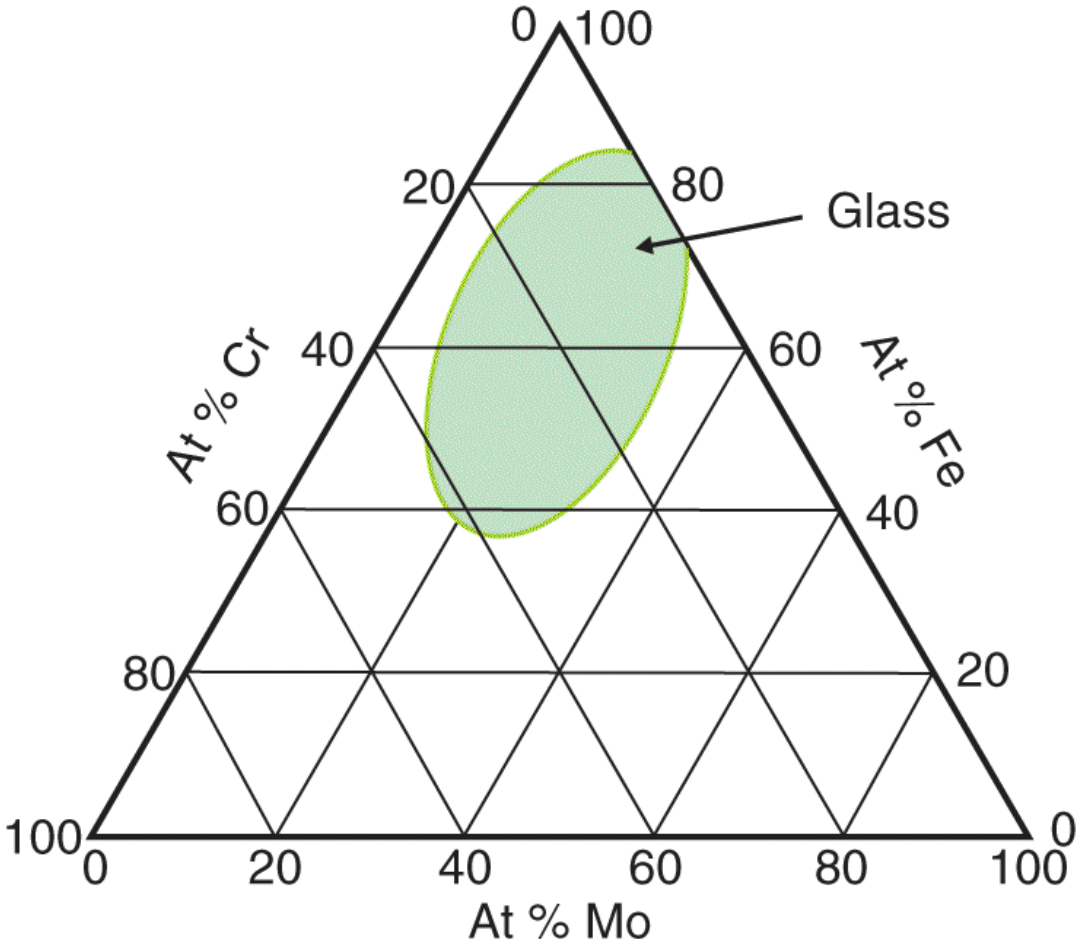


Figure 4.7 The glass forming region in the $(Fe_{1-x}Cr_xMo_y)_{75}C_{15}B_{10}$ phase diagram.

As described earlier (Chapter 3) a crystal can only be built from a unit cell consistent with the seven crystal systems and the 14 Bravais lattices. The unit cell can be translated to build up the crystal, but no other transformation, such as reflection or rotation, is allowed. This totally rules out a unit cell with overall fivefold, or greater than sixfold, rotation symmetry. In 1984, a metallic alloy of composition approximately $Al_{88}Mn_{12}$ was synthesised, which gave a sharp electron diffraction pattern that clearly showed the presence of a 10-fold symmetry axis. The crystallographic problem is that the sharp diffraction pattern indicated long-range translational order. However, this was incompatible with the 10-fold rotation axis, which indicated that the unit cell possessed a forbidden rotation axis.

Since then, many other alloys that give rise to sharp diffraction patterns and that show 5-fold, 8-fold, 10-fold, and 12-fold rotation symmetry have been discovered. They resemble the metallic glasses in that they are generally alloys with a complex formula. The resulting materials became known as *quasiperiodic crystals*, or more compactly, as *quasicrystals*. Figure 4.8 shows the electron diffraction pattern from a quasicrystalline alloy of composition $Al_{72}Ni_{20}Co_8$, clearly showing 10-fold rotation symmetry.

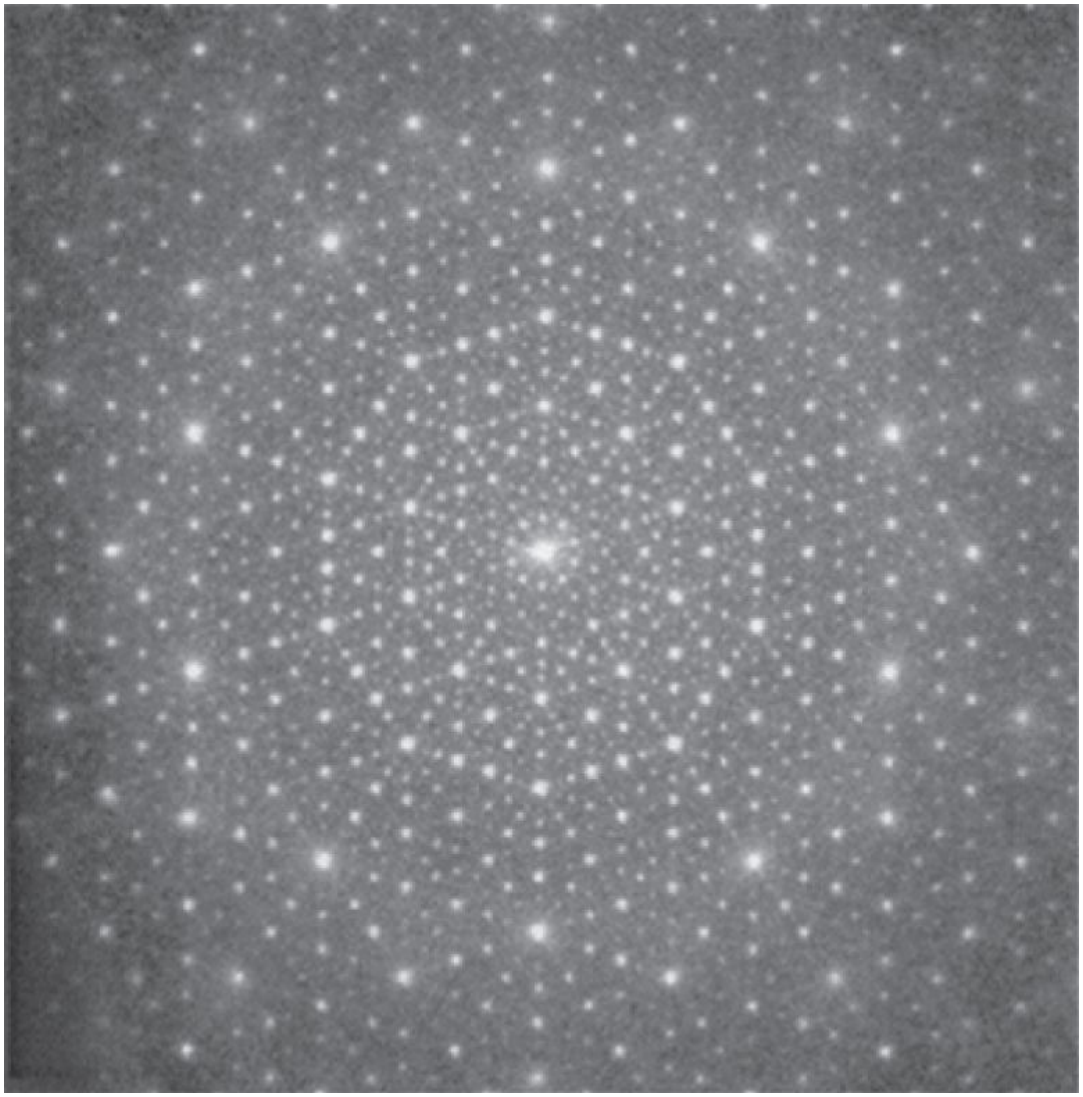


Figure 4.8 Electron diffraction pattern from the quasicrystalline alloy $Al_{72}Ni_{20}Co_8$, showing 10-fold rotation symmetry.

Source: Photograph courtesy of Dr. Koh Saitoh, see K. Saitoh et al. (2004), reproduced with permission.

The simplest way that the contradiction between classical crystallography and the structures of quasicrystals can be resolved is to consider a quasicrystal as an imperfectly crystallised liquid. The liquid state of many metal alloys contains

metal icosahedra that form in a transitory fashion, because this polyhedral arrangement represents a very efficient way of packing a small number of spheres together (Figure 4.9a). Thus, a liquid metal can often be regarded as being composed of icosahedral clusters of metal atoms that are continually forming and breaking apart. On solidifying a liquid metal, the icosahedra freeze. If the cooling rate is slow enough, the icosahedra order onto a lattice, say cubic or hexagonal, which then extinguishes the overall icosahedral symmetry of the structure. A large number of metal alloy structures are known, which contain ordered icosahedra, including $CoAs_3$, with the cubic skutterudite structure, in which As atoms take on this geometry (Figure 4.9b).

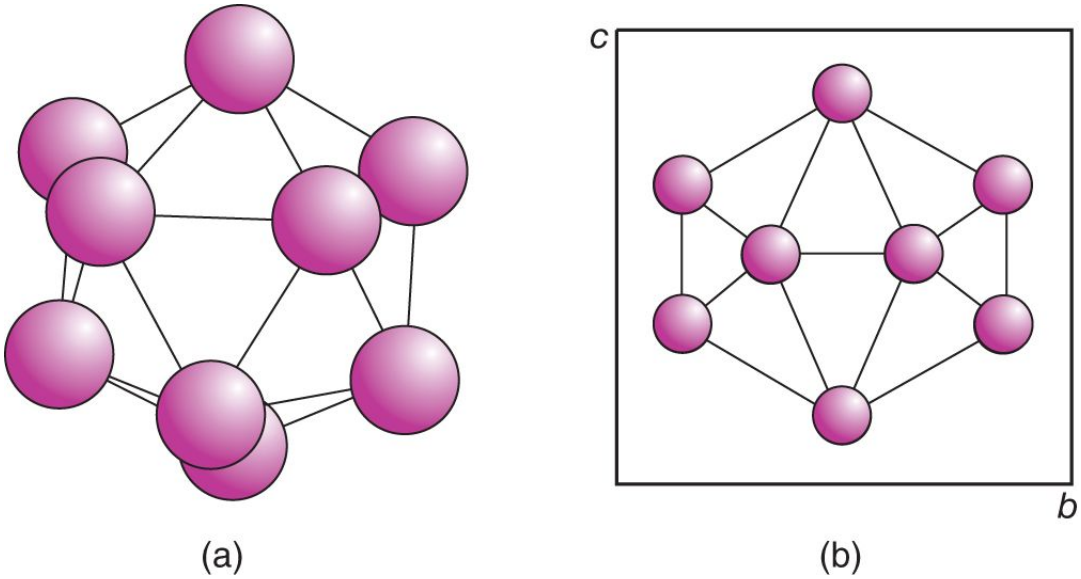


Figure 4.9 Icosahedra: (a) icosahedral geometry; (b) part of the As_3Co skutterudite structure, projected down (100) showing an icosahedron of As atoms. (Note that only a fraction of the As atoms and none of the Co atoms in the As_3Co unit cell are shown. The atoms in both (a) and (b) touch. They are drawn smaller here to show the icosahedral geometry.)

If the cooling rate of a liquid metal is very fast, the icosahedra freeze at random, to give an icosahedral glass (Figure 4.10). Over a range of intermediate temperatures, the icosahedra all freeze in the same orientation, but they do not have the energy or time to order on a lattice. A quasicrystal can then be regarded as made up of icosahedral clusters of metal atoms, all oriented in the same way, and separated by variable amounts of disordered material (Figure 4.10). That is, the materials show orientational order but not translational order.

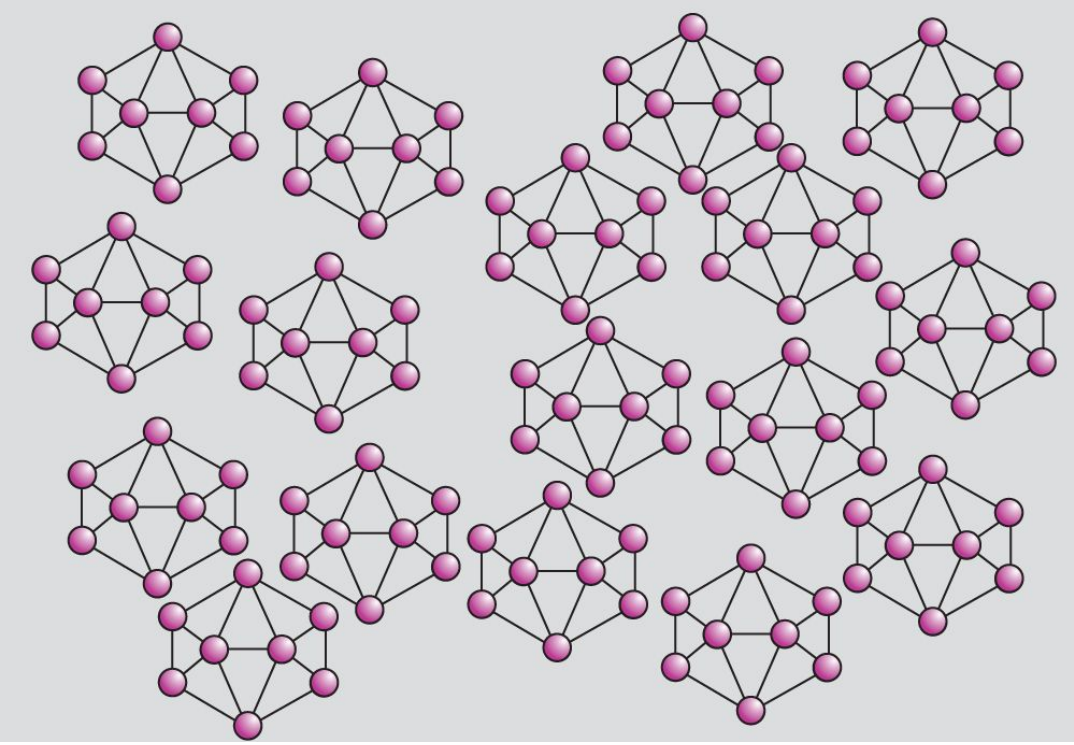


Figure 4.10 Icosahedra, arranged with the same orientation, but not on a lattice, can be taken as one model of a quasicrystal alloy. If the icosahedra are disordered further the structure can be taken as one model of a bulk metallic glass.

4.1.5 The Principal Properties of Metals

Since the earliest times, metals have been recognised by their shiny appearance and the fact that they could be bent, twisted, or beaten into foils and retain the shape so formed. Metals show good electrical and thermal conductivity and dissimilar metals when joined can be formed into thermocouples in which a current flow is engendered when the two ends are at different temperatures. The relationship between the thermal conductivity and electrical conductivity is given by the *Wiedemann–Franz law*:

$$\text{Thermal conductivity/}$$
$$\text{electrical conductivity} = 3\,Tk_B{}^2/e^2$$

where T is the temperature (K), k_B is the Boltzmann constant, and e is the charge on the electron.

These properties arise via the interaction of crystal structure and microstructure, all of which are consequences of the metallic bond. In a metal the electrons, which are non-localised in bonds, can move throughout the metal under the imposition of only a very low thermal or electrical driving force. The

shiny appearance of metals is also due to the fact that when light photons strike a metal surface, the highest energy electrons, those near to the Fermi surface, can absorb the photons, as plenty of empty energy states lie nearby. However, the electrons can just as easily fall back to the lower levels originally occupied, and the photons are re-emitted. A detailed explanation of reflectivity of a metal requires knowledge of the exact shape of the Fermi surface and the number of energy levels (density of states) at the Fermi surface. In the case of gold, relativistic effects have also been found to play a part to play in the generation of the gold colour, as opposed to the silver tone exhibited by most metals. Similarly, when two dissimilar metals are joined, electrons will flow from the higher Fermi energy to the lower. This gives rise to thermoelectric effects and to the operation of thermocouples. Less directly, the Fermi energy is related to the extent to which metals corrode.

Alloying with other metals or non-metals in small amounts will not change these physical attributes drastically but the foreign atoms generally impede electron transport. Alloys therefore tend to have poorer electrical conductivity than pure metals. If a new phase forms, the Fermi surface and the density of states at the Fermi surface will change and the electrical conductivity will alter abruptly. For example, the incorporation of hydrogen within magnesium will initially produce an interstitial alloy with metallic properties, but an inferior electronic conductivity. Additional hydrogen leads to the formation of the hydride MgH_2, which is transparent and non-metallic.

An important mechanical property of a metal is that of ductility, meaning that a metal can be deformed easily without breaking, and can retain the deformed shape indefinitely. Pure metals in particular are soft and easily drawn into wires or hammered into foils. This property can be attributed to the crystal structures of the metals, which consist of packing of more or less spherical atoms. These can readily roll over each other to produce the deformation. There are no strong localised bonds to be broken, and so the metallic bonding does not hinder the deformation. However, metallic bonding does occur, and generally, calculations show that a pure metal deforms more easily than the metallic bonds should permit. The conflict is resolved by recognition of the role of dislocations, which allow deformation to occur without the necessity of breaking significant numbers of metallic bonds at any moment.

Associated with the easy deformability of a metal is the fact that metals can be easily hardened. The explanation again lies with to the dislocations present in the crystals. The trapping of these dislocations, called *pinning*, prevents easy deformation, and hence the metal becomes harder. At its simplest level, this can come about by the introduction of impurities and the formation of substitutional or interstitial alloys, a fact used empirically since the Bronze Age. However, more effective hardening comes about if precipitates form in the crystal. These, brought about by alloying and annealing, produce very hard metals. Steels, already far harder than pure iron, are hardened further by the incorporation of

carbon or nitrogen, followed by heat treatment that causes these elements to combine and form precipitates in the parent crystals. Alloy compositions that result in large amounts of second phase formation are hardest as dislocation movement becomes almost impossible, with the consequence that the metal becomes brittle. Thus, cast iron, which contains numerous precipitates of cementite and graphite, is very brittle, as are metals that contain large quantities of hydrogen. Recently much interest has been shown in alloys consisting of nanoscale crystallites, as these can offer enhanced mechanical properties compared to similar alloys with larger grains.

4.2 Crystalline Silicates and Inorganic Ceramic Materials

The term ceramics originally referred to solids that were fabricated from naturally occurring ingredients such as china clay, which had been moulded into a desired shape and then heated at high temperatures until the soft mass was transformed into a hard brittle solid. Bricks, earthenware pots, porcelain, and similar items fall into this grouping. Ordinary glass, originally made by melting sand (SiO_2), is also usually included. The major properties of these solids are that they are hard, brittle, chemically inert dielectrics (electrical insulators) that are stable at high temperatures.

During more recent years, the term ceramics has expanded to include a large number of inorganic materials, which display a wide variety of other important properties, including electronic conductivity, superconductivity, magnetism, piezoelectric effects, ferroelectric effects, and a number of important optical properties. It is difficult to specify a small number of defining characteristics of such a diverse group, except that they are generally inorganic solids, most often oxides, but also including carbides, silicides, nitrides, oxy-nitrides, hydrides, and others, which all tend to be hard, inflexible, and stable at high temperatures.

Because of this diversity, it is convenient to consider ceramics that are essentially silicates, called *traditional ceramics*, separately from *engineering ceramics*, with important mechanical properties, *electroceramics* when electronic properties are emphasised, or *glasses* and related non-crystalline ceramics. Traditional ceramics are used in utility applications such as brickwork and drainage pipes, as well as porcelain and other fine decorative ceramic objects. Engineering ceramics are used to extend the operating range available to metallic components. They are valued for high temperature stability and for extreme hardness. Typical uses include hard surface coatings on metallic components (titanium nitride, tungsten carbide); inert high temperature components (valves, cylinder liners, ceramic shielding and blankets, furnace linings); high speed cutting tool inserts (transition metal carbides); and abrasives (alumina, silicon carbide). Electroceramics are very high purity materials that possess unique electronic properties, varying from insulating to superconducting. These form the active element in many gas sensors, temperature sensors, batteries, fuel cells, and solar cells. Ceramic magnets are widely used in motors. Ceramics are also important in fluorescent lighting and as components of computer displays.

4.2.1 Silicate Structures

Silicon is a small atom with the electronic configuration [Ne] $3s^2\ 3p^2$, which mainly links to other atoms via strong tetrahedral sp^3-hybrid covalent bonds analogous to those that form in many carbon compounds. In silicates, each silicon atom is usually linked to oxygen to form an [SiO_4] tetrahedron (Figure 4.11a). The bonds are strong and silicon–oxygen tetrahedra are stable

and vary very little in size. The Si–O distances are always close to 0.162 nm and the O–O distances to 0.27 nm. In terms of an ionic model, the $[SiO_4]$ tetrahedral group has an overall charge of −4, and is written $[SiO_4]^{4-}$.

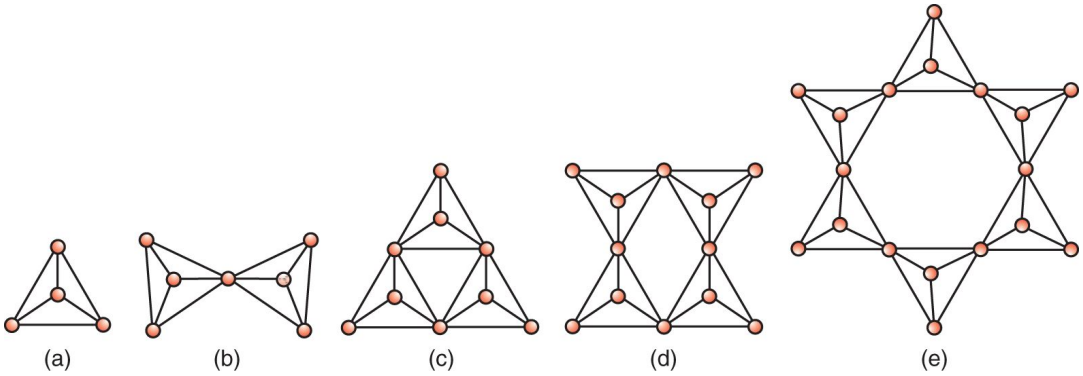


Figure 4.11 Linked $[SiO_4]$ units found in ionic silicates: (a) isolated (SiO_4); (b) (Si_2O_7); (c) (Si_3O_9); (d) (Si_4O_{12}); (e) (Si_6O_{18}).

These $[SiO_4]$ tetrahedra form the basic structural unit in silicates and dominate silicate chemistry and physics. They are found as isolated units, or condensed to form $[SiO_4]$ chains, sheets or three-dimensional (3D) networks. In all these structures, only the vertices of the tetrahedra are shared, never edges or faces.

Silicates that contain isolated $[SiO_4]^{4-}$ units, pairs $[Si_2O_7]^{6-}$, three-membered rings, $[SiO_9]^{6-}$, four-membered rings, $[Si_4O_{12}]^{8-}$, or six-membered rings, $[Si_6O_{18}]^{12-}$ (Figure 4.11)1 are often called *ionic silicates*. Silicates that contain chains or sheets of silicate tetrahedra are *extended anion silicates*. Even the simplest structure, a single chain of tetrahedra linked by corners (Figure 4.12a), can adopt several different configurations. The composition of a single chain is $[SiO_3]^{2-}$, leading to compounds of formula $MSiO_3$, the *pyroxenes*. Two single pyroxene chains can join by linking half of the free vertices to form a double chain, of formula $[Si_4O_{11}]^{6-}$ (Figure 4.12b). These double chains are found in the *amphiboles*, a group including several forms of asbestos. Linkage of the free vertices on the amphibole chains leads to single silicate layers (Figure 4.13a,b), and linking the free vertices that lie at the apex of each tetrahedron will lead to double silicate layers (Figure 4.13c). Silicate layers in which the free vertices link to a layer of magnesium or aluminium hydroxide octahedra to produce composite layers form an enormous group that includes many *clays*. These composite structures are possible because the positions of the oxygen atoms at the vertices of a silicate layer matches well the geometry of the close-packed array of oxygen atoms that occurs in the hydroxides $Mg(OH)_2$ and $Al(OH)_3$ (Figure 4.14a,b). In both of these compounds, the Mg^{2+} and Al^{3+} cations are in octahedral sites. The match between the oxygen arrays allows a single silicate

layer to link to a single hydroxide layer (Figure 4.14c) or two silicate layers to sandwich a hydroxide layer (Figure 4.14d). The single silicate plus hydroxyl layer is found in clays such as *kaolinite* and minerals such as *chrysotile asbestos* (Figure 4.15a). Double silicate sandwich layers enclosing a hydroxyl layer are found in minerals such as *talc* (Figure 4.15b). If aluminium replaces some of the silicon, the silicate layers take on a negative charge. This is counterbalanced by cations placed between the layers, to form the *micas* (Figure 4.15c).

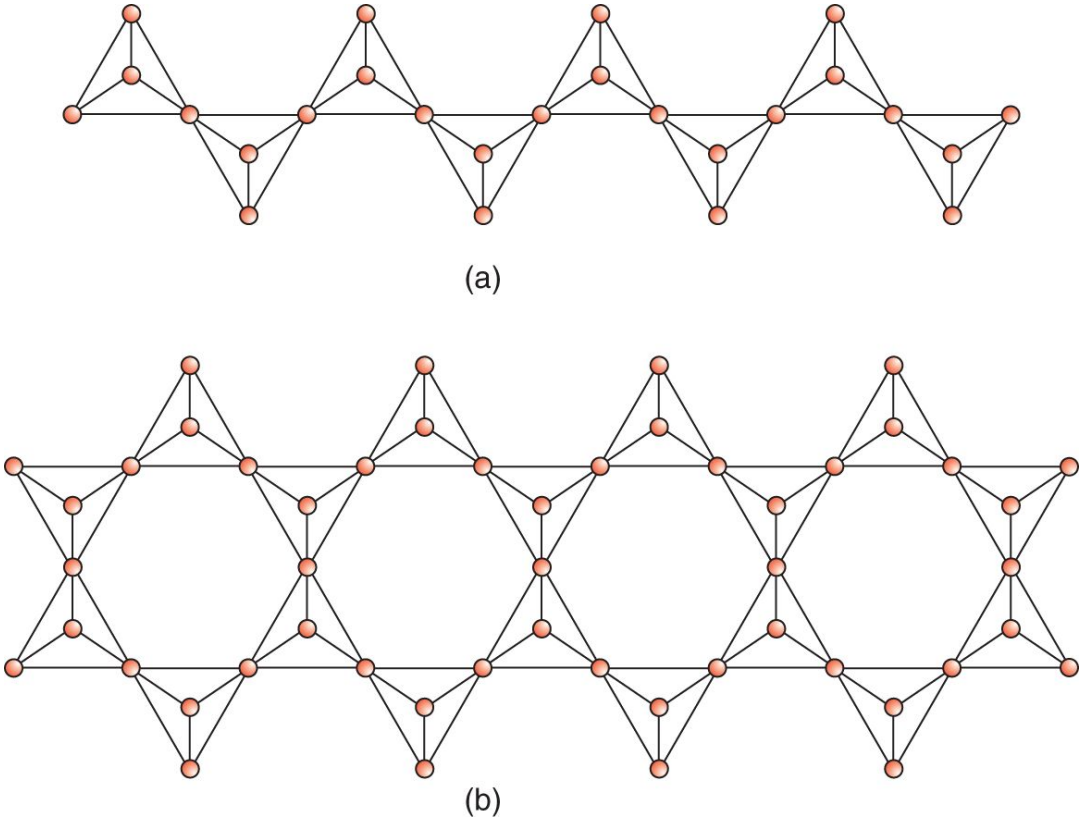


Figure 4.12 (a) Single (SiO_3) and (b) double (Si_4O_{11}) chains.

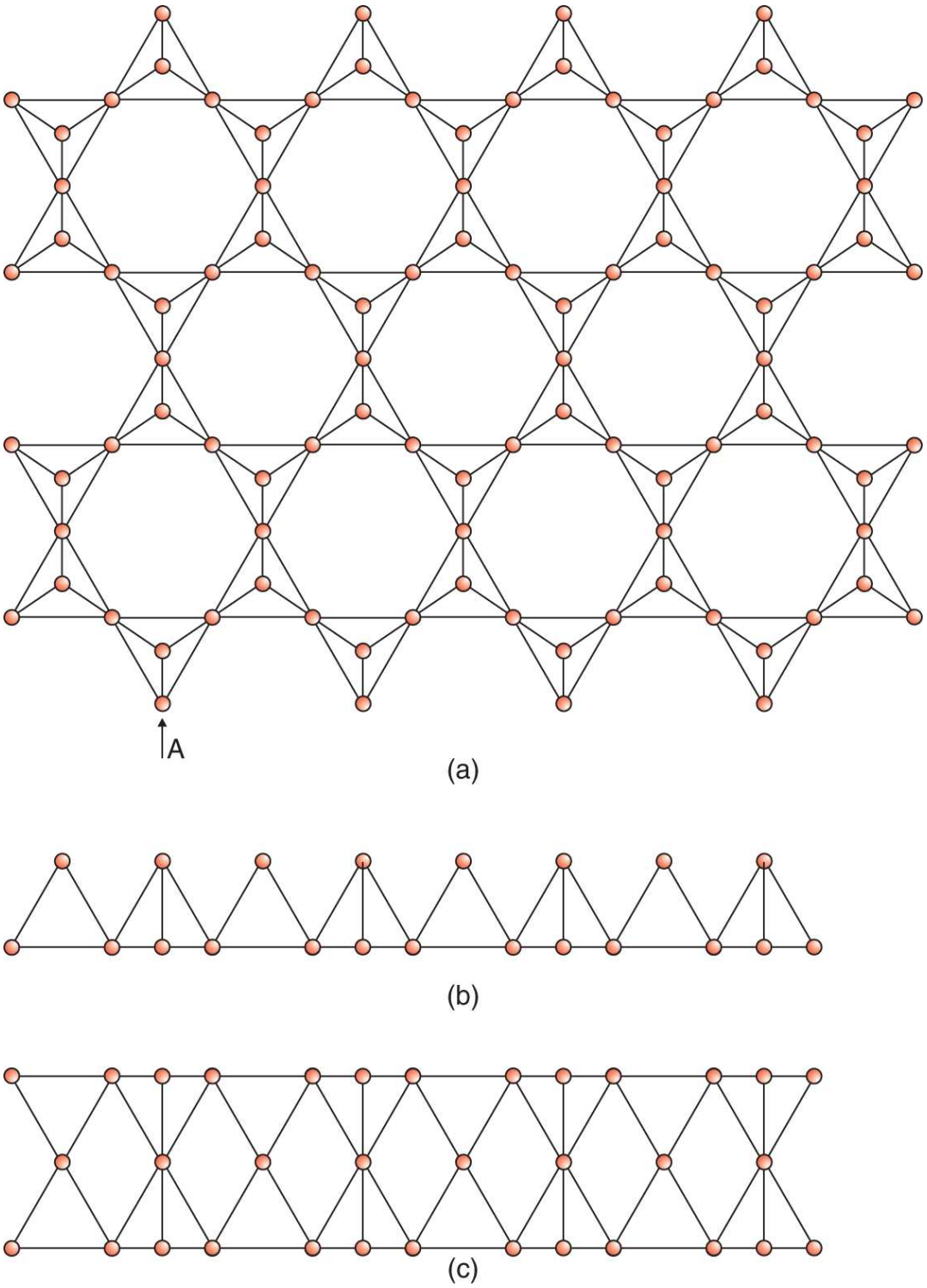


Figure 4.13 (a) Single sheet of corner-linked (SiO_4) tetrahedra; (b) the same sheet viewed along A; (c) double sheets formed by linking two layers of the type shown in (a) viewed along A.

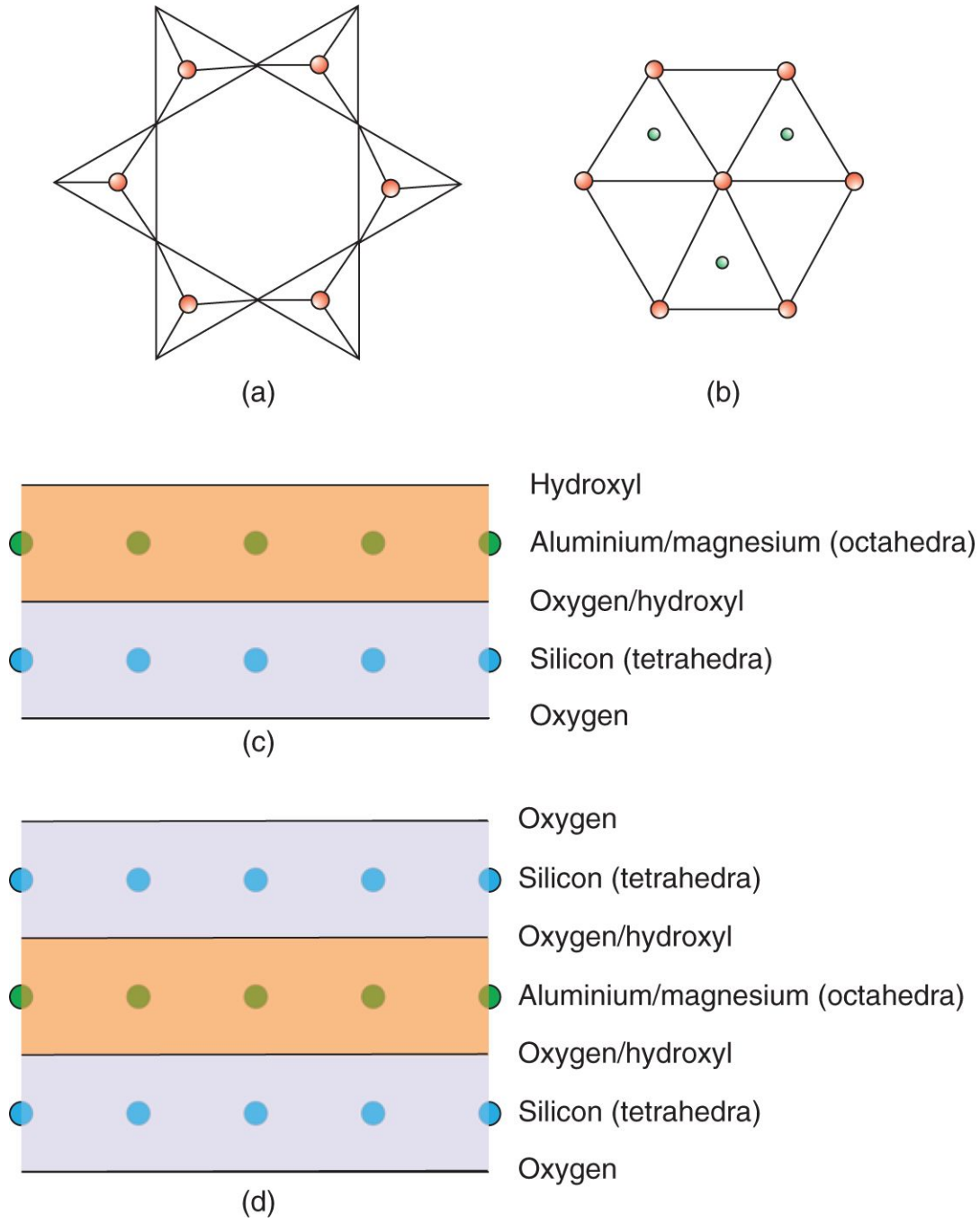


Figure 4.14 The structure of clays and related minerals: (a) single layer of corner-linked (SiO_4) tetrahedra; (b) basal close-packed layer of O^{2-} (large circles) in $Mg(OH)_2$ or $Al(OH)_3$, with cations in octahedral sites above the layer (small circles); (c) composite layer formed by uniting a single silicate layer (a) with a single hydroxide layer (b); (d) composite structure formed by linking two silicate layers (a) with a single hydroxide layer (b).

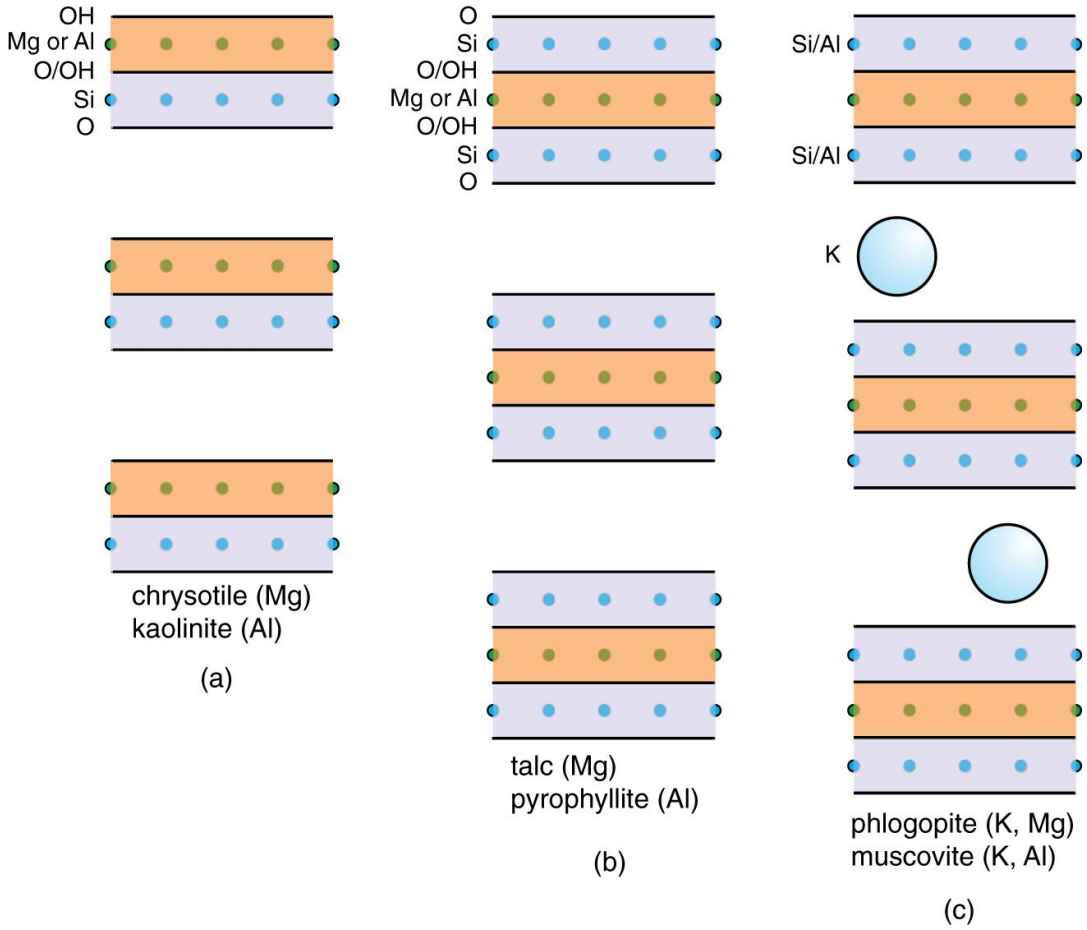


Figure 4.15 The structures of some clay- and mica-related minerals formed from composite silicate-hydroxide layers.

In a number of important minerals, the [SiO_4] tetrahedra link by all vertices to form a 3D covalent network. This can adopt a number of different conformations, the simplest of which are typified by the polymorphs of silica, SiO_2, the commonest of which are the various forms of quartz. When aluminium replaces some of the silicon, the framework takes on a negative charge, usually compensated for by the insertion of large cations into the network. This produces the large group of minerals the *aluminosilicates*, which include the important *feldspars*, which make up many strong rocks.

The related *zeolites* and *ultramarines*, all of which are of considerable industrial importance, have a porous structure. The zeolites, of which approximately 250 structures have been characterised, are known as molecular sieves and are widely used in industry for the separation of water and other small molecules in gas mixtures. Although there are large numbers of structural variations found in these framework silicates, many can be described in terms of the structure of the mineral sodalite, $Na_8Al_6Si_6O_{24}Cl_2$. This contains a framework of corner linked AlO_4 and SiO_4 tetrahedra, which can be envisaged as a truncated octahedron

(Figure 4.16). The resulting polyhedron, called a *sodalite unit* or β-*cage*, has Si and Al alternately at vertices of the truncated octahedron with O near to the mid-points of each edge. The charge imbalance due to the substitution of Al^{3+} for Si^{4+} is compensated by the inclusion of additional Na^{+} at the centre of each hexagonal ring. The framework encloses a large cavity that contains the Cl^{-} anions. These truncated octahedral frameworks are linked via the square faces in sodalite itself, and via cubes located on these square faces in the commercial zeolite Linde-A. The use of pentagonal prisms so as to link the sodalite units will produce the structure of the mineral faujasite, $NaCa_{0.5}(Al_2Si_5O_{14}).10H_2O$, and forms the basis of a number of synthetic zeolites such as Linde-X.

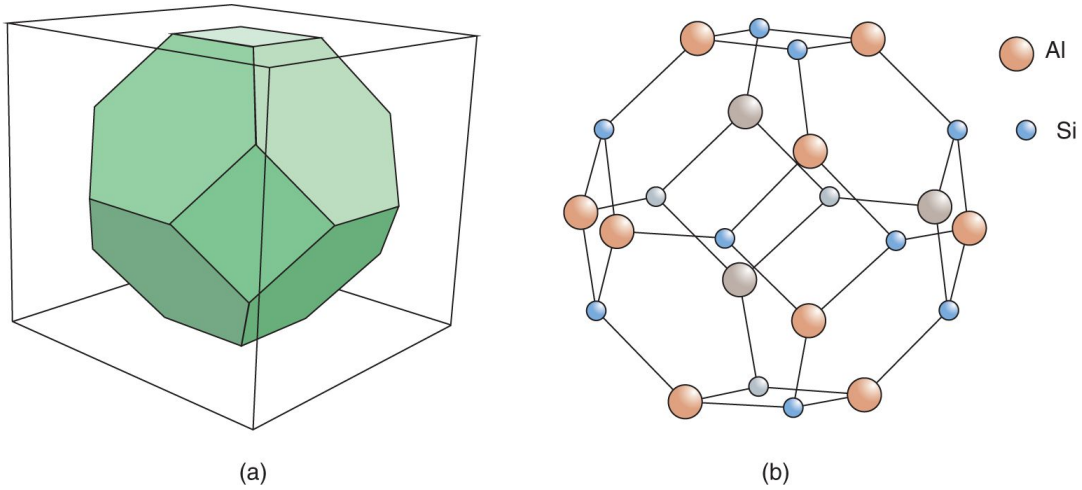


Figure 4.16 (a) Truncated octahedron; (b) sodalite unit, with oxygen atoms omitted.

Silicate structures of most relevance to ceramics are summarised in (Table 4.2).

Table 4.2 A summary of silicate structures.

Structure	Basic formula	Hardness (Mohs)	Examples
Isolated			
Monomer	$[SiO_4]^{4-}$	8–5	Mg_2SiO_4, forsterite (*olivines*) $Ca_3Cr_2(SiO_4)_3$, uvarovite (*garnets*)
Dimer	$[Si_2O_7]^{6-}$	5	$Sc_2Si_2O_7$, thortveitite
3-ring	$[Si_3O_9]^{6-}$	7–4	$BaTi(Si_3O_9)$, benitoite
4-ring	$[Si_4O_{12}]^{8-}$	7–4	$Ca_3Al_2(BO_3)(Si_4O_{12})(OH)$, axinite
6-ring	$[Si_6O_{18}]^{12-}$	6–4	$Be_3Al_2(Si_6O_{18})$, beryl $NaMg_3Al_6(BO_3)_3(Si_6O_{18})(OH)_4$, tourmaline
Chains			
Single	$[SiO_3]^{2-}$	7–4	$MgSiO_3$, enstatite (*pyroxenes*)
Double	$[Si_4O_{11}]^{6-}$	5	$Ca_2Mg_5Si_8O_{22}(OH)_2$, tremolite (*amphiboles*)
Sheets			
Single silicate layer	$[Si_2O_5]^{2-}$	3–1	$Na_2Si_2O_5$, nitrosilite
Double silicate layer	$[SiO_2]$	3–1	$CaAl_2Si_2O_8$ (half Si replaced by Al), misteinbergite
Single silicate + single hydroxide layer	$[Si_2O_5]^{2-}$ + $[OH]^-$	3–1	$Al_2(OH)_4Si_2O_5$ kaolinite (*clays*) $Mg_3(OH)_4SiO_5$ chrysotile (*clays*)
Double silicate + single hydroxide layer	$[Si_4O_{10}]^{4-}$ + $[OH]^-$	3–1	$Al_2(OH)_2Si_4O_{10}$, pyrophyllite (*clays*) $Mg_3(OH)_2Si_4O_{10}$, talc (*clays*)

Structure	Basic formula	Hardness (Mohs)	Examples
	$[(Si,Al)_4O_{10}]^{n-}$ + $[OH]^-$	3–1	$KAl_2(OH)_2Si_3AlO_{10}$, muscovite (*micas*) $KMg_3(OH)_2Si_3AlO_{10}$, phlogopite (*micas*)
Network			
Silicate	$[SiO_2]$	8	SiO_2, quartz
Alumino-silicate	$[(Si,Al)_4O_8]$	7–5	$KAlSi_3O_8$, orthoclase (*feldspars*)
Sodalite, zeolite related	$[Si_6Al_6O_{24}]$ + $[Na^+, Ca^{2+}$ ...]	3–1	$Na_8Al_6Si_6O_{24}Cl_2$, sodalite $NaCa_{0.5}(Al_2Si_5O_{14}).10H_2O$, faujasite

4.2.2 Some Non-silicate Ceramics

Almost every inorganic oxide that does not contain silicon, as well as many carbides and nitrides, can be thought of, to some extent, as a non-silicate ceramic. A large number of important ceramics adopt the halite (NaCl) structure. These include the oxides MgO and NiO and many carbides and nitrides such as TiC and TiN. The oxides are ionic insulators, while the carbides and nitrides have metallic properties.

Aluminium oxide, alumina, able to withstand high temperatures, is used in laboratory furnace ware and other applications where resistance to high temperatures in air are important. A number of other oxides, including Fe_2O_3 and Cr_2O_3, adopt the same structure. This is most easily described as a (slightly distorted) hexagonal close packed array of oxygen ions, with Al^{3+} ions distributed in an ordered fashion over 2/3 of the available octahedral sites (Figure 4.17). A number of well-known gemstones consist of Al_2O_3 doped with small amounts of transition metal impurities. Ruby owes its colour to about 0.5% Cr^{3+} substituted for Al^{3+}, while sapphire contains a small amount of Ti^{4+} and Fe^{2+} substituted for Al^{3+}.

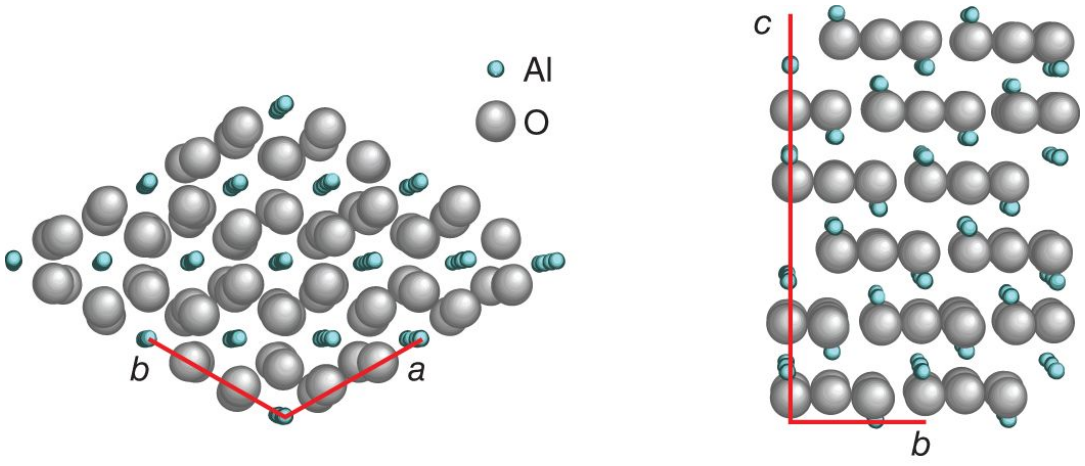


Figure 4.17 The structure of alumina, Al_2O_3: (a) projected close to [001]; (b) projection close to [100].

Zirconia, ZrO_2, is an important ceramic because it is also able to withstand high temperatures. At room temperature, the structure contains irregular polyhedra formed by seven oxygen ions surrounding each Zr^{4+} cation, contained in a monoclinic unit cell. At a temperature of about 1100 °C, the unit cell becomes tetragonal, a change caused by the coordination polyhedra becoming more regular. At temperatures of about 2300 °C, the structure becomes cubic and adopts the CaF_2 structure (Figure 4.18). In this, each Zr^{4+} ion is surrounded by eight oxide ions at the corners of a cube and the whole can be visualised as an array of ZrO_8 cubes, each linked by all edges, in a 3D chessboard array. The monoclinic to tetragonal crystallographic transformation at approximately 1100 °C involves a significant volume change. This causes cracking and weakness in ceramic components and makes it impossible to use zirconia in its pure state for high temperature applications. However, the cubic form can be stabilised by the addition of impurities such as calcia, CaO, in which Ca^{2+} ions substitute for Zr^{4+} ions in the structure. The resulting *calcia-stabilised zirconia* remains cubic from room temperature to the melting point, above 2300 °C, allowing problem-free high-temperature use.

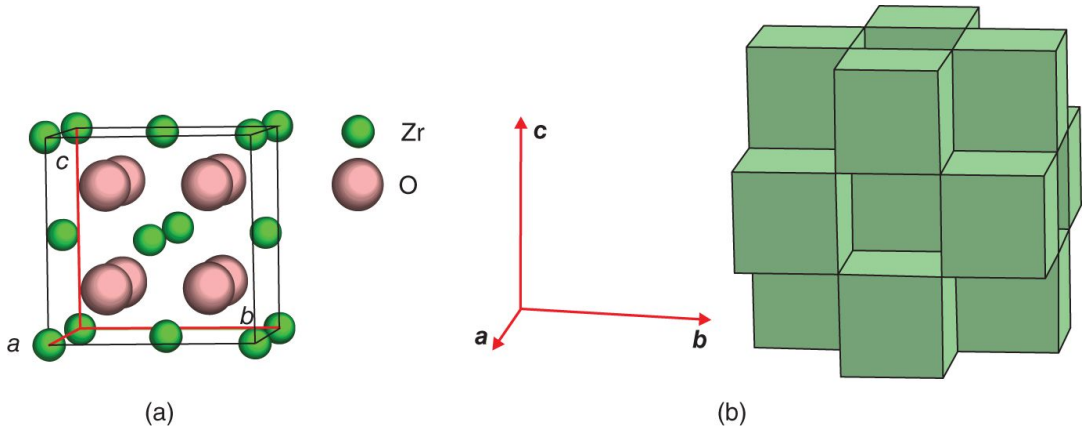


Figure 4.18 Cubic calcia-stabilised zirconia: (a) unit cell; (b) the structure as a stacking of edge-shared ZrO_8 cubes.

4.2.3 The Preparation and Processing of Ceramics

Many of the characteristic properties of ceramics arise during manufacturing. Ceramic bodies are made by *high temperature firing* (heating) routes that induce chemical reactions to take place, during which the final material is produced.

Traditional ceramics are mainly made from mixtures of clays, silica (often extracted as flint), and feldspars (especially $K_2Al_2Si_6O_{16}$ and $Na_2Al_2Si_6O_{16}$).

Low-quality structural products, such as bricks, pipes, are made directly from the appropriate clay. Higher quality ceramics, for example, porcelain, are made from carefully controlled amounts of specific clay, flint, and feldspar. Because of the use of these three major components, such materials have been given the name of *triaxial whitewares*. Clay based traditional ceramics are first formed into the shape desired, using traditional potters' techniques, or automated processes. The resulting shapes are then heated to 1000 °C or higher. At these temperatures, the clays initially react to lose water and subsequently hydroxyl ions. The residues react to give a mixture of new phases, including glasses. Considerable shrinkage usually occurs during these reactions, and the production of an object with a precise size and shape requires careful controls.

Engineering ceramics and electroceramics are usually pure single phases. Many are made by milling components to produce a fine powder, pressing into the desired shape and finally firing. This third step is critical and heating cycles are often complex. Initial heating is usually at a low temperature, 100–400 °C, to remove water and to burn off binder (an organic material used to cement the dry grains together). At a later stage of processing, the material is heated to a final temperature, in excess of 1000 °C, to sinter the particles and allow any chemical reactions to occur. In this final stage, some components may melt, and in some cases (i.e. porcelain) one of the products is a glass. This is called *vitrification*.

Hard ceramic surface coatings on metallic components can be made by heating the metal in an appropriate gaseous atmosphere. Reaction takes place at the metal surface and atoms from the gaseous component diffuse into the surface layer. Thus, heating titanium in nitrogen gas will cause a layer of titanium nitride TiN, to form on the surface as a hard layer. For more information on sintering of ceramics, see Section 6.1.

4.2.4 The Principal Properties of Ceramics

The principal properties of ceramics arise from a combination of chemical bonding and the atomic defects and microstructure that result during fabrication. The bonding, whether described as ionic or covalent, is strong, which ensures that the solids are chemically inert and often *refractory* (stable to high temperatures). Refractory ceramics are widely used in furnaces and other high temperature equipment. In addition, refractory components are used both externally, to protect the outside of space vehicles and satellites, and internally in rocket motors.

The lack of free electrons endows basic ceramics with poor thermal and electronic conductivity. The chemical flexibility of ceramics, however, allows them to be selectively doped with other atoms. In particular, doping with transition metal or lanthanoid ions generates ceramics with a wide variety of optical, electronic, and magnetic properties. Insulators, for example, can be transformed into superconductors in this way.

Ceramics do not deform very easily at ordinary temperatures, as strong chemical bonds must be broken, and unlike metals, dislocation movement is severely hampered. Generally, stress results in brittle fracture, especially under impact (Section 8.3). However, in large part, the brittle nature of ceramics arises not from the bonding or dislocation density, but in the microstructure. Surface flaws cause many ceramics to fail catastrophically when flexed. Voids, pores, large grains, and foreign inclusions in the structure, resulting from the chemical reactions that take place during high temperature fabrication, are sources of weakness. Although ceramics have a rather low tensile strength, and readily fracture when stretched, they are much stronger when compressed. This property, coupled with the strong bonding, makes ceramics hard, and they are used as abrasives, cutting tools, and hard coatings.

4.3 Silicate Glasses

Silicate glasses are hard, transparent, fairly strong, corrosion resistant materials, in which the main component is silica, SiO_2 (Table 4.3). There are a number of naturally occurring silicate glasses, including obsidian (a volcanic rock which is black due to iron oxide impurities), pumice (a glassy froth), flint and opal. These all show the typical glass properties of hardness and brittleness.

Table 4.3 Some silicate glasses.

Name	Typical composition	Important property	Principal uses
Soda glass	15%Na_2O : 85% SiO_2	Transparent	Window glazing
Soda-lime glass	72% SiO_2 : 14% Na_2O : 14% CaO	Transparent	Window glazing
Borosilicate (Pyrex)	80% SiO_2 : 13% B_2O_3 : 7% Na_2O	Low coefficient of expansion	Cooking ware, laboratory ware
Crown glass	9% Na_2O : 11% K_2O : 5% CaO : 75% SiO_2	Low refractive index	Optical components
Flint glass	45% PbO: 55% SiO_2	High refractive index	Optical components, ‘crystal’ glass
Lead glass	up to 80%PbO : 20% SiO_2	Absorbs radiation	radiation shielding
Silica	100% SiO_2	Very low coefficient of thermal expansion	Optical components, laboratory ware, optical fibres

Silicate glass production marked an early stage in civilisation. *Faience* was fabricated by the Egyptians thousands of years ago from moulded sand, coated with *natron*, a residue of minerals left after flooding of the river Nile, consisting mainly of calcium carbonate, sodium carbonate, common salt, and copper oxide. The object was heated to about 1000 °C, at which point the alkali coating reacted to form a glassy exterior with a blue colour imparted by the copper oxide.

Improvements in glass technology were carefully guarded secrets of medieval guilds, and it is only relatively recently that high-quality transparent glass has been readily available. This was brought about firstly with the development of high-quality lenses in the nineteenth century and then with the development of

optical-fibre based communication systems in the second half of the twentieth century.

4.3.1 Bonding and Structure of Silicate Glasses

The main structural unit in silicate glasses is the tetrahedral [SiO_4] group. The covalent bonds between the central silicon atom and the oxygen atoms are strong, and [SiO_4] tetrahedra persist in the liquid states of silica and silicates. These tetrahedra link to one another by sharing corner oxygen atoms to form discrete or interpenetrating chains and rings. During solidification the entangled chains are difficult to reorganise into an ordered crystalline arrangement, so that cooling a melt, even at fairly slow rates, can result in the formation of a glass. Thus, structure of a silicate glass was envisaged by Zachariasen, in 1932, as an irregular intertwining of chains of corner-linked [SiO_4] tetrahedra to form a loose *random network* (Figure 4.19).

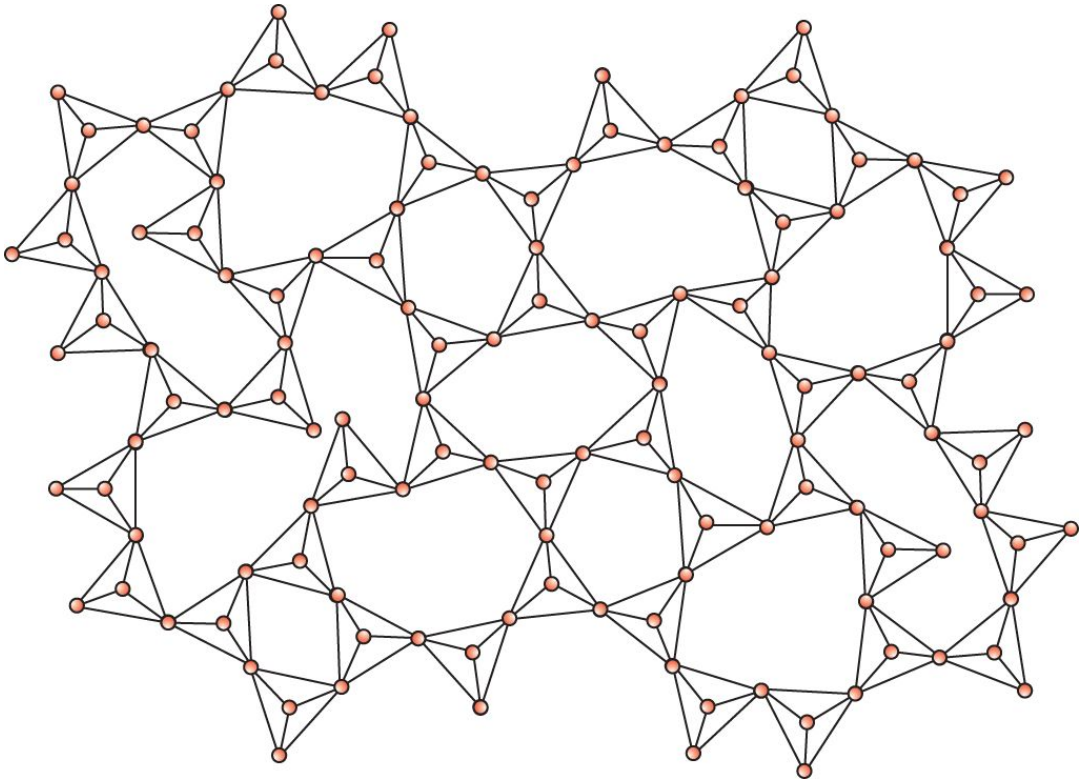


Figure 4.19 The random network model of corner-linked (SiO_4) tetrahedra in a silicate glass.

A silicate glass is often described as a *supercooled liquid*. However, it is important to keep in mind that glass is *not* a liquid, but a solid with a structure that does not show any long-range order. This status of glass is revealed by the behaviour on heating, because, unlike crystalline solids, glasses do not have a sharp melting point. Instead they continually soften from a state, which can be

confidently defined as solid to a state that can be defined as a viscous liquid. In place of a melting point, glasses are characterised by a *glass transition temperature*, T_g. The glass transition temperature is determined by plotting the specific volume (the volume per unit mass) of the glass as a function of temperature (Figure 4.20). Both the high temperature and low temperature regions of such a plot are usually linear. The value of T_g is given by the intersection of the extrapolated high and low temperature lines. Above the glass transition temperature, the material can be considered a liquid, while below the glass transition temperature, it is considered a solid. The glass transition temperature is not a fixed material property, but varies with the cooling rate and as a number of different definitions of T_g are used, values must be treated cautiously. Nevertheless, it is a useful materials parameter that gives guidance concerning the softening and working temperature of a particular glass.

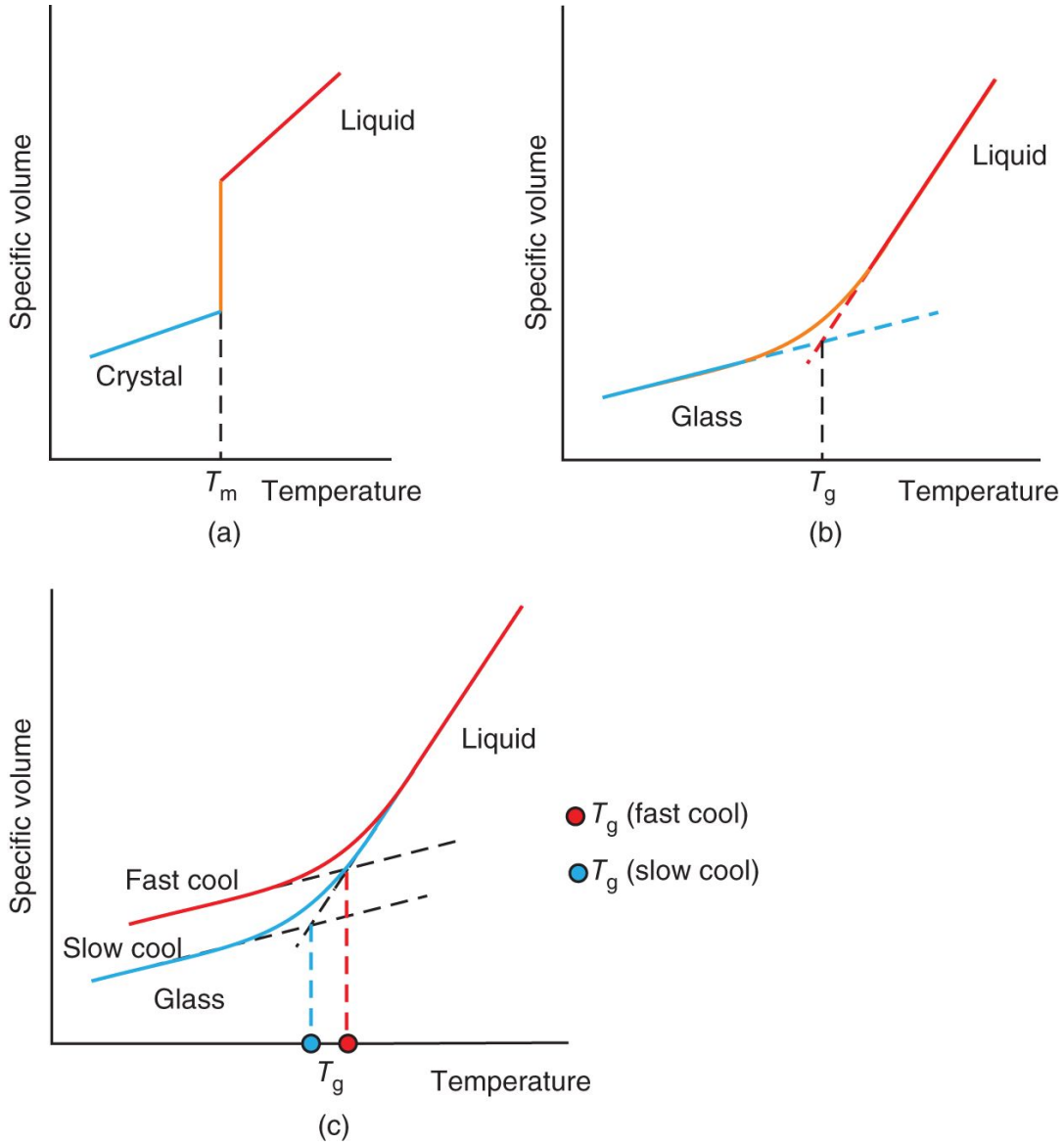


Figure 4.20 Specific volume vs. temperature curves: (a) crystalline solid with melting point T_m; (b) glass with glass transition temperature T_g; (c) the effects of cooling on the glass transition temperature, T_g.

The random network model works well for silicate glasses. Apart from $[SiO_4]$ tetrahedra, cations that occupy tetrahedral or triangular coordination with similar metal–oxygen bond lengths to Si□O bonds can also fit neatly into the chains. These ions, typically B^{3+}, Ge^{4+}, Al^{3+}, Be^{2+}, and P^{5+}, are known as *network formers*. Large cations, which tend to disrupt the ability of the $[SiO_4]$ tetrahedra to crystallise in regular arrays, also enhance glass formation. These, known as *network modifiers*, are typified by K^+, Na^+, Mg^{2+}, or Ca^{2+}. Other cations, those with higher valence and coordination typically of 6, are called *intermediates*. Typical examples are Ti^{4+}, Cu^{2+}, and Zn^{2+}. Although

intermediate cations alter the properties of a glass, by adding colour, for example, they do not have a direct role to play in glass formation. Note that these three divisions are not mutually exclusive and some ions fall into more than one category. For example, Al^{3+} is regarded both as a network former, when in triangular or tetrahedral positions, and an intermediate when in octahedral coordination.

4.3.2 Glass Deformation

One of the most curious properties of glass, and one that differentiates glass from both metals and crystalline ceramics, is the fact that it can be deformed readily in a semi-molten state, by traditional techniques such as glass blowing. This is because glasses behave as very viscous liquids at moderate temperatures and can then be manipulated into the desired shape, which is retained in solid form on cooling.

The viscosity of a glass, especially the variation of viscosity with temperature (Figure 4.21a), is an important physical parameter, as it defines the temperature regimes over which the glass can be worked (Table 4.4). The *working range* of a glass is the range of viscosities over which normal processing takes place, usually between viscosity2 values of 10^3 and $10^{6.6}$ Pa s. Note that the temperature at which any glass reaches the requisite viscosity depends upon its composition, which itself is related to its final use.

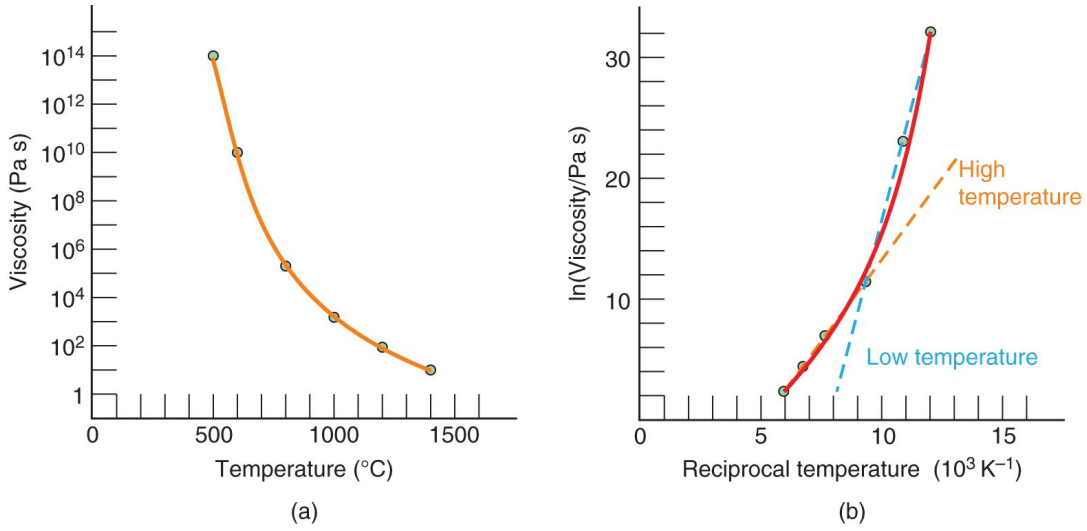


Figure 4.21 Viscosity of a soda-lime glass: (a) viscosity vs. temperature; (b) in(viscosity) vs. reciprocal temperature.

Table 4.4 Glass viscosity.

Glass condition	Approximate viscosity (Pa s)	Comment	Approximate temperature (°C) for soda-lime glass
Melting temperature	10	Glass becomes fluid and a homogeneous melt is achievable	1450
Working point	10^3	Glass easily deformed, but retains its shape	1000
Softening point	$10^{6.6}$	Glass deforms under its own weight	700
Annealing point	10^{12}	Residual stress in a thin plate can be removed in 15–20 minutes	550
Strain point	$10^{13.5}$	Fracture – plastic deformation boundary	500

Because of the practical importance of viscosity, there have been several attempts to define the relationship between viscosity and temperature by a mathematical equation. The simplest starting point is an Arrhenius equation:

$$\eta = \eta_0 \exp(E/RT)$$

where η is the viscosity, η_0 is a constant, E is the activation energy for viscous flow, R is the gas constant, and T the temperature (°K). This equation describes a large number of processes, including diffusion, in which the transformation is possible only when the participants have sufficient energy to overcome a barrier E. In such reactions, Arrhenius behaviour is confirmed by a plot of ln η *vs.* $1/T$:

$$\ln \eta = \ln \eta_0 + E/RT$$

which is linear with a slope of E/R. The temperature variation of the viscosity of most glasses does not obey this simple law, which are then said to show *non-Arrhenian* behaviour. Instead, the plot of ln η *vs.* $1/T$ is generally a smooth curve in which the high temperature slope is usually lower than that of the low temperature part (Figure 4.21b). However, it is sometimes convenient to approximate portions of this curve to straight line *Arrhenian* behaviour, as in the case of pure silica, which has an activation energy (slope) of approximately 500 kJ mol^{-1} above 1400 °C and an activation energy of approximately 700 kJ mol^{-1} below 1400 °C.

Several other equations have been suggested to overcome the shortcomings of the Arrhenius equation. The most widely used is the empirically derived *Vogel–*

Fulcher–Tammann equation (also called the *Vogel–Fulcher* equation or VFT equation):

$$\log \eta = A + [B/(T - T_0)]$$

where A, B, and T_0 are empirical constants, and T is the temperature (K). This equation is able to correlate viscosity measurements over very wide ranges, especially at higher temperatures. At lower temperatures, though, the predicted viscosities are generally higher than experimental values. To correct for this, several other empirical equations have been suggested, the most recent being the Mauro-Yue-Ellison-Gupta-Allan (MYEGA) equation:

$$\log \eta = \log \eta_\infty + (K/T) \exp (C/T)$$

where η_∞, K, and C are empirical constants. This equation agrees well with the VFT equation at higher temperatures but the two tend to diverge at lower temperatures, where the MYEGA equation provides the best fit.

Viscosity is often represented on an *Angell plot* (Figure 4.22) of log η vs. T_g/T. The (straight) diagonal corresponds to ideal Arrhenius behaviour. Curves to the right of this are said to represent *fragile liquids* while the opposite side of the diagonal refers to *strong liquids*. The *fragility* of a liquid thus measures the departure from Arrhenian behaviour. It is related to the various atomic interactions present contributing to the viscosity, including weak bonding such as van der Waals and hydrogen bonding, but is not simply an expression of chemical bonding. Silica, for example, falls into the fragile region, but has strong covalent Si□O bonds present within the SiO_4 tetrahedra.

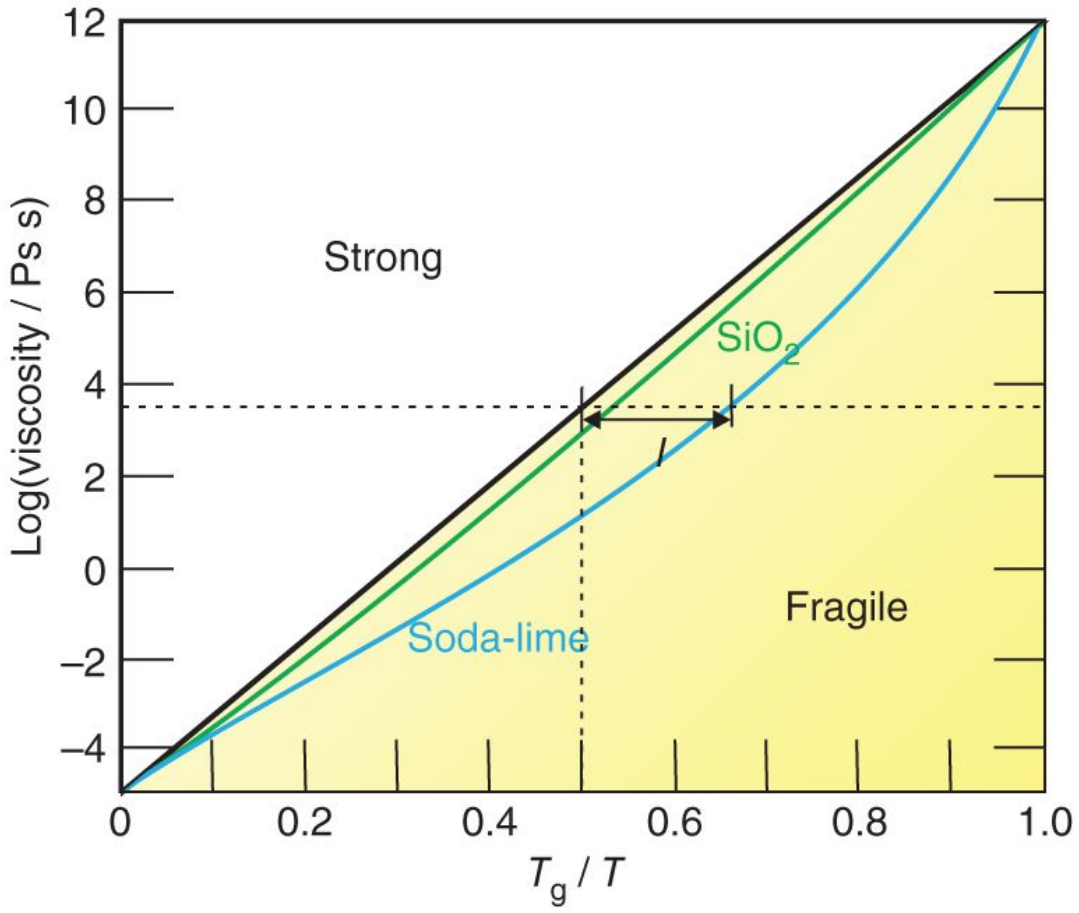


Figure 4.22 Angel plot of fragility. The curves are for silica and soda-lime glass.

The fragility has been quantified by measuring the slope of the Arrhenius plot as the temperature tends towards the glass transition temperature, T_g:

$$m = [d \log \eta / d(T_g/T)]_{T \approx T_g}$$

where m is called the fragility or also the *steepness index*. The value of m for a liquid obeying the Arrhenius law is given by the slope of the diagonal on the Angell plot, equal to 17. The value of m is not always easy to obtain from experimental data and an alternative *kinetic fragility* $F_{1/2}$ has been proposed. The $F_{1/2}$ value is given by:

$$F_{1/2} = 2\,[T_g/T - 0.5]$$

where T_g/T is the value where a horizontal line drawn half-way up the log η cuts the requisite glass curve. The quantity in brackets corresponds to the value l (for soda-lime glass, Figure 4.22) and the factor 2 ensures that $F_{1/2}$ values fall between 0 and 1.

The concept of fragility is far reaching and has applications in many areas of science apart from silicate glasses, including the study of water-based solutions and pharmaceutical products.

4.3.3 Strengthened Glass

Glass is strong in compression but notoriously weak in tension. Freshly drawn glass fibres, for example, are stronger than steel, but attack of the surface by water vapour in the atmosphere causes the strength of the fibre to fall dramatically. This weakness is usually attributed to small flaws in the glass surface, called *Griffith's flaws* (Section 8.3). Under tension the flaws generate cracks allowing the material to fracture. Many mechanisms for strengthening glass are ways to prevent cracks from propagating through the solid.

Tempered glass is about four times stronger than ordinary glass, and when it fractures it breaks into small, blunt pieces rather than jagged shards. The glass is strengthened by rapidly cooling the hot surface with air jets. Initially, the glass is cut to shape, and surface flaws and rough edges removed by grinding and polishing. The glass is then heated to 620 °C. High-pressure jets of air, in a predetermined array, rapidly cool the surface in several seconds. During this quenching process, the outside of the glass sheet becomes rigid, as the outside temperature drops below the glass transition temperature. The inside, though, is still above this temperature as it cools more slowly than the outside. As the inside cools, it tends to shrink away from the solid outer surfaces. This results in the centre of the sheet being in tension, while the outside of the sheet is being pulled in and so is in compression. These opposing tensile and compressive stresses are the strengthening mechanism. As glass usually fails due to the generation of surface cracks, a surface under compressive force is much harder to break. Carefully controlled patterns of stresses are generated, dependent upon the final use of the material. (These stress patterns are can be seen as coloured shapes in the windshield of a car driven in sunlight when the driver wears sunglasses with polarising lenses, such as Polaroid sunglasses.)

Chemical strengthening aims to mimic the tension and compression distribution just described, by using chemical means. The methods used are successful, but tend to be more expensive than air-cooling, and are mainly used in applications in which cost is a secondary consideration. The principle of the method is to selectively replace some of the metal ions in the glass to achieve tension or compression. For example, if a soda-lime glass is placed in a bath of molten potassium nitrate the Na^+ in the glass surface is replaced by the larger K^+. This causes a surface compressive stress and an interior tensile stress, and the glass is strengthened. This process is used to produce aircraft glazing and lenses. Similarly, if the Na^+ is replaced by Li^+, which is smaller, the surface is under tension and the centre under compression. This process is used in the fabrication of glass for use as laser materials.

4.3.4 Glass-Ceramics

A glass-ceramic is a solid that is largely crystalline, made by the partial *devitrification* (crystallisation) of a glass object of the desired shape. Glass-ceramics are therefore composite materials that consist of crystals and some glass. They combine the ease of production of glass with much-enhanced thermal and mechanical properties. The initial step is to fabricate the object in a suitable glass. The transformation of this piece into a largely crystalline body of the same shape and size is carried out by a controlled thermal treatment. This induces the precipitation of crystal nuclei, the growth of crystals on the nuclei, and the development of an almost fully crystalline final product (Figure 4.23). In addition, the crystallisation must occur in a glass that is viscous enough not to sag or distort during the transformation.

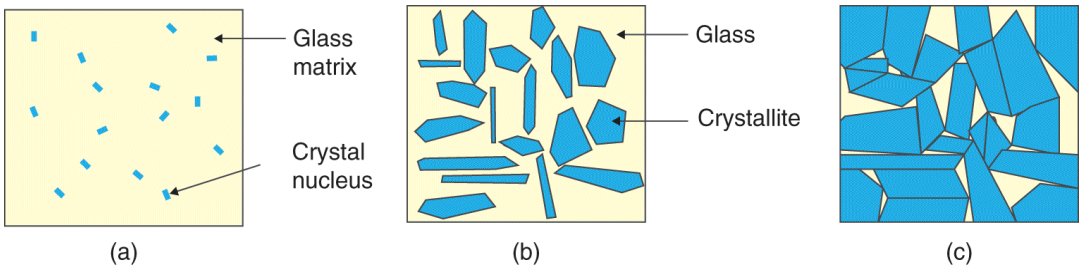


Figure 4.23 Crystal nucleation (a) and growth (b, c) in a glass ceramic, schematic.

The two most important factors in glass-ceramic production are the composition of the melt and the microstructure of the final product. These are inter-related, of course. The composition controls the ability of the substance to form a glass with the correct viscosity and workability, as the starting solid is completely glassy in nature. Composition also controls which nuclei can form in the glass, and the types of crystal that can grow. Most crystals have a definite crystal habit, and this factor greatly influences the microstructure of the final solid.

Good mechanical properties are achieved because the solid consists of a mass of interlocking crystals. Any surface flaws cannot easily propagate through the solid, as the passage of a crack is blocked by crystals its path. Good thermal properties are achieved by ensuring that the crystals that form have very low coefficients of expansion, thus making the material resistant to thermal shock. The optical properties of the solid can also be manipulated. If the crystals are kept to a dimension below that of the wavelength of light, the solid will be transparent to visible light. If the crystal dimensions are larger, the solid can be, for instance, opaque to visible light but transparent to microwaves or radar waves. Crystallites of a modest size that are well dispersed in a glass residue give a translucent solid, typified by porcelain.

The simplest glass-ceramics, from a microstructural viewpoint, are the transparent ultra-fine-grained materials that consist essentially of the high

temperature form of quartz. The preparation of these materials starts with a silica melt that also contains some ZrO_2, TiO_2, Al_2O_3, and MgO. The melt is formed into the desired shape and then heat treated while the viscosity remains at a value great enough to prevent sag or deformation. At this temperature, $ZrTiO_4$ crystals are the first to nucleate, and on these, crystals of the high temperature form of quartz, stable between 573 and 870 °C, grow. The presence of the aluminium in the melt means that the quartz is not pure SiO_2, but some aluminium substitutes for silicon in the crystals. In order to maintain charge neutrality, some Mg^{2+} is also incorporated into the quartz, and this stabilises the structure to room temperature. The final solid is a mass of crystals less than 60 nm in size, together with some residual glass, which cements them together. The small crystal size means that the solid is transparent, and the very low thermal expansion of quartz means that the material is resistant to thermal shock. The many small grains in the structure prevent crack growth, so that the ceramic is also strong.

Other compositions and heat treatments result in glass ceramics with different microstructures. The amount of glass present and the type of crystalline phase present will have a profound effect upon the resultant properties. For example, in materials used as rocket nose cones, the structure contains mainly cordierite, $Mg_2Al_2Si_5O_{18}$, an aluminosilicate with a very low thermal expansion coefficient. The residual glass is in the form of isolated volumes, so that flow is not easy when the material is heated. Machinable glass-ceramics can also be prepared by suitable control of chemistry and crystallisation. These materials contain a crystalline phase consisting of interlocking plates of mica, with a composition approximating to fluorphlogopite, $KMg_3AlSi_3O_{10}F_2$. The presence of the interlocking plates prevents the solid from splintering when it is cut. These are used for precision ceramic components that can only be manufactured via a machining stage.

4.4 Polymers and Organic Materials

4.4.1 Polymers

The majority of polymers, and the only ones considered here, are compounds of carbon. Polymers are very widespread and can be synthetic, for example, nylon, or natural, for example, rubber. Ordinary artificial polymeric materials, colloquially often called plastics, are made up of tangled and interlocking polymer chains. At the other extreme, the polymers that form vital components of living organisms, such as DNA, which is a polymer of amino acids, is precision assembled in living cells to have controlled length, sequences of subunits, stereochemistry and chemical reactivity.

Within both of these groups, the carbon backbone of the molecule can be linear, branched or cross-linked, and a great variety of chemical groups can be attached to the chain spine. Relatively simple polymers are long chain-like macro (giant) molecules, often called linear polymers, made by the linkage of large numbers of small repeating molecules called *monomers*, each of which has two reactive sites that can link the molecules head to tail. Short chain lengths formed in the course of synthesis or degradation of such polymers are called *oligomers*. More complicated structures form when molecules with more than two reactive sites can couple together or link with linear chains to produce 3D configurations. The chains are most often tangled in an amorphous mass, and these structures are described in Sections 4.4.2, 4.4.3, and 4.4.4. However, highly crystalline network structures can also be synthesised and these covalent organic frameworks (COFs) and metal-organic frameworks (MOFs) are described Section 4.4.6.

There are four main divisions of synthetic and generally non-crystalline polymeric materials. *Thermoplastic* materials can be melted and reformed a number of times. These consist of networks of long chains linked by weak non-covalent bonds to give an amorphous mass. They are used below the glass transition temperature of the melt. On being heated through the glass transition temperature, these polymers transform into viscous fluids that can be remoulded. Typical examples are polyethylene, polypropylene, poly(methyl methacrylate), polycarbonates, and polyamides (nylons) (see Table 4.5 and Section 4.4.2). *Elastomers* (Section 4.4.4) are a subgroup of thermoplastics and can be deformed a considerable amount and return to their original size rapidly when the force is removed.

Table 4.5 Addition polymers.

Monomer source	Polymer formula	Names	Uses
CH_2=CH_2 (ethene, ethylene)	$(CH_2–CH_2)_n$	Polyethylene, polythene, PE	Squeeze bottles, food bags, dishes, insulation, coatings
CH_2=CH Cl (vinyl chloride)	$(CH_2–CHCl)_n$	Poly(vinyl chloride), PVC	Pipes, floor covering, insulation, adhesives, films, credit cards
CH_2=CCl_2 (vinylidene chloride)	$(CH_2–CCl_2)_n$	Poly(vinylidene chloride)	Food wraps, fibres, cling film
CH_2=CH CH_3 (propylene)	$(CH_2–CH\ (CH_3))_n$	Polypropylene, PP	Pipes, valves, carpets
CH_2=CH C_6H_5 (styrene)	$(CH_2–CH\ (C_6H_5))_n$	Polystyrene, PS	Jugs, cups, packaging, styrofoam, appliance parts
CH_2=CH CN (acrylonitrile)	$(CH_2–CH\ (CN))_n$	Polyacrylonitrile, PAN, Orlon, Acrilan	Fabrics, carpets, high-impact plastics
CH_2=CH $COOCH_3$ (vinyl acetate)	$(CH_2–CH\ (CH_3\ COO))_n$	Poly(vinyl acetate), PVA	Wood adhesives, paper coatings, latex paints

Monomer source	Polymer formula	Names	Uses
$CF_2=CF_2$ (tetrafluoroethene)	$(CF_2–CF_2)_n$	Poly(tetrafluoroethylene), PTFE, Teflon	Non-stick coating, electrical insulation, bearings
$CH_2=C(CH_3)COOCH_3$ (methyl methacrylate)	$(CH_2–C(CH_3)COOCH_3)_n$	Poly(methyl methacrylate) Perspex, Lucite, Plexiglass	Substitute glass, acrylic paints, pipes
$CH_2=CH–CH=CH_2$ (1,3-butadiene)	$(CH_2–CH–CH–CH_2)_n$	Polybutadiene, buna rubber	Tires, hoses, pond liners

Thermosetting materials are amorphous polymers that can only be formed once. The junction molecules and the strands that they link are joined into a network by strong covalent bonds. The materials are used well below the glass transition temperature of the phase. They cannot be melted and tend to degrade upon heating. These solids are typified by the epoxy, polyurethane, and formaldehyde resins. They are generally used above the glass transition temperature.

Gels are network polymers that can be penetrated by liquids such as water or organic solvents, to give a soft but flexible solid that holds the liquid in the cavities within the network of the polymer. Examples include polyacrylamide hydrogel and naturally occurring gelatine. Gels are also described as colloidal materials.

4.4.2 Polymer Formation

Polymers that are produced by the adding together of a single monomer unit are called *addition polymers*, and those that eliminate one or more small molecules during this reaction are called *condensation polymers*. If the molecular reactants have only two reactive sites (or are *bifunctional*) generally at either end of the molecule, then linear polymer chains result.

Many small molecules are, in a normal state, unreactive. In order to link a large number of monomers, it is necessary for each end of the molecules to be made chemically reactive. Generally, this involves breaking chemical bonds to yield two reactive half-bonds. The simplest starting point for a discussion of this process is the monomer molecule ethene (ethylene), $CH_2=CH_2$. Schematically, the monomer ethene can be linked to other ethene monomers if the double bond is opened and the resulting broken bonds are linked together, in an *addition*

reaction (Figure 4.24). The chemical formula of the resulting polymer, polyethylene, or polythene, is $[CH_2–CH_2]_n$, where n takes a value of several thousand. (Note that the industrial preparation of polyethylene and other polymers is generally quite different from the schemes illustrated.) The polymer chain is constructed from (CH_2) units. In these, each carbon atom is bonded to two hydrogen atoms and two carbon atoms using strong sp^3-hybrid bonds, which have a tetrahedral geometry (Figure 4.25). The carbon–carbon bonds are free to rotate, which allows the polymer chain to coil into ordered or disordered regions. Note that polythene, like all polymers, does not have a definite chemical formula. The number of (CH_2) units in the chain is influenced by preparation conditions. Polymers with low average values of n have different physical properties than those in which n is large.

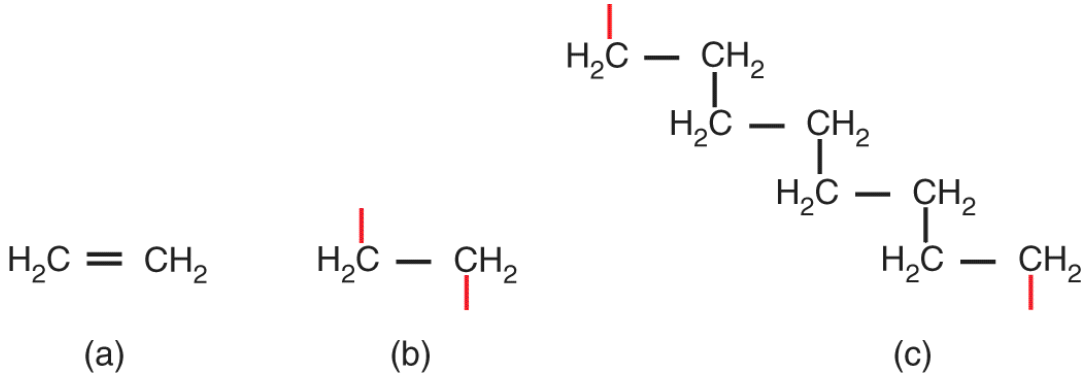


Figure 4.24 The polymerisation of ethylene, schematic: (a) a single monomer molecule; (b) double bond opening; (c) monomer linkage to form the polymer chain.

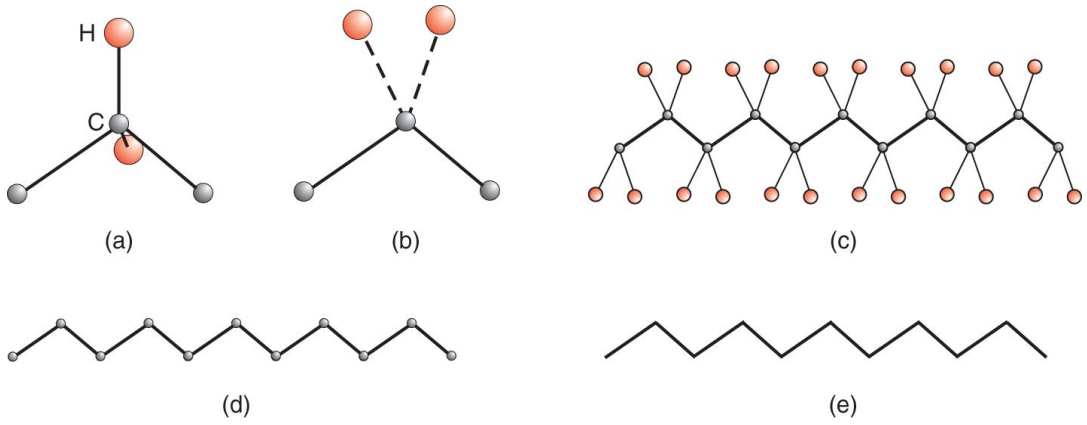


Figure 4.25 (a, b) The structure of a tetrahedral CH_2 unit in a polymer chain. (c) A chain of linked CH_2 units. (d, e) Simplified representations of the chains with H atoms omitted.

Related polymers can be formed by replacing one or more of the hydrogen atoms in the monomer ethene with a chemical group of atoms X (Table 4.5).

Hence the formula of the monomer becomes CH_2=CHX when one hydrogen is replaced. The principle of polymerisation remains precisely the same, however, although the details are modified by the size and location of the side-group X.

A typical *condensation reaction* is one between an acid group, –COOH, and a hydroxyl group, –OH, to form a larger molecule and split out water:

$$\mathrm{R{-}COOH + R'{-}OH \rightarrow R{-}COO{-}R' + H_2O}$$

where R and R′ represent different carbon chains. In molecules with only one acid or hydroxyl group, as written, the reaction stops after the first step. In order to create a polymer, the monomers of condensation polymers need *two* reactive groups on each monomer. The polyesters, which generally start from terephthalic acid and the alcohol ethylene glycol, are an important group of polymers made by reaction is between acid groups and hydroxyl groups (Figure 4.26). The initial reaction that links a terephthalic acid molecule with an ethylene glycol molecule occurs at approximately 210 °C. Polymerisation and the regeneration of ethylene glycol takes place towards 270 °C. The product, polyester poly(ethylene terephthalate), or PET, is widely used for shatterproof bottles and wrappings.

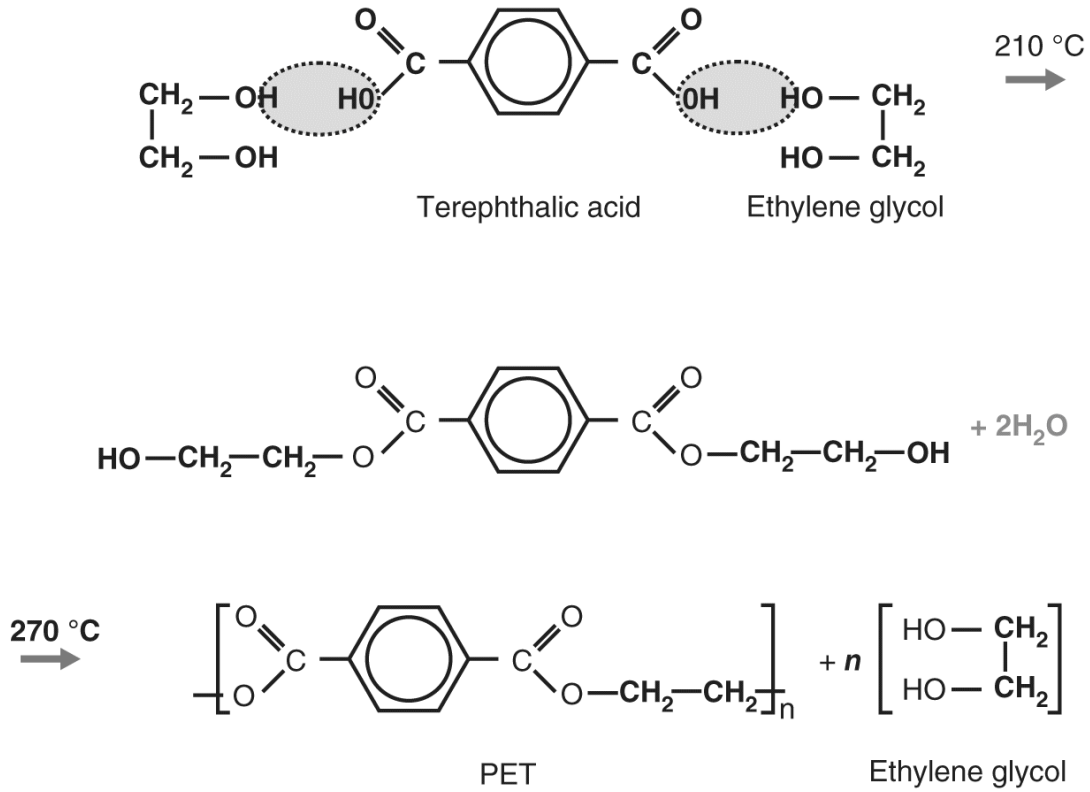


Figure 4.26 The formation of the polyester PET, schematic.

A group of polymers that form in a similar way are the thermoplastic polyamides, better known as *nylons*. The principal reaction is between an acid group, –COOH, and an amide group, $–NH_2$:

$$R–COOH + R'\text{-}NH_2 \rightarrow R–CON–R' + H_2O$$

where R and R′ represent different carbon chains. Once again, in order to form a polymer, each monomer must have a reactive group at each end of the molecule. The formation the commonest type of nylon, nylon 66, is an example (Figure 4.27a). The name, nylon 66 indicates that there are six carbon atoms in each section of the repeat unit of the polymer. Amino acids are molecules with an acid group at one end and an amide group at the other, separated by a carbon chain, $HOOC–R–NH_2$. These molecules can also polymerise by linking head to tail, to generate a nylon. The amino acid analogue of nylon 66 is nylon 6 (Figure 4.27b). (Note that nylon 6 is not actually made from the amino acid shown, but from ε-caprolactam.)

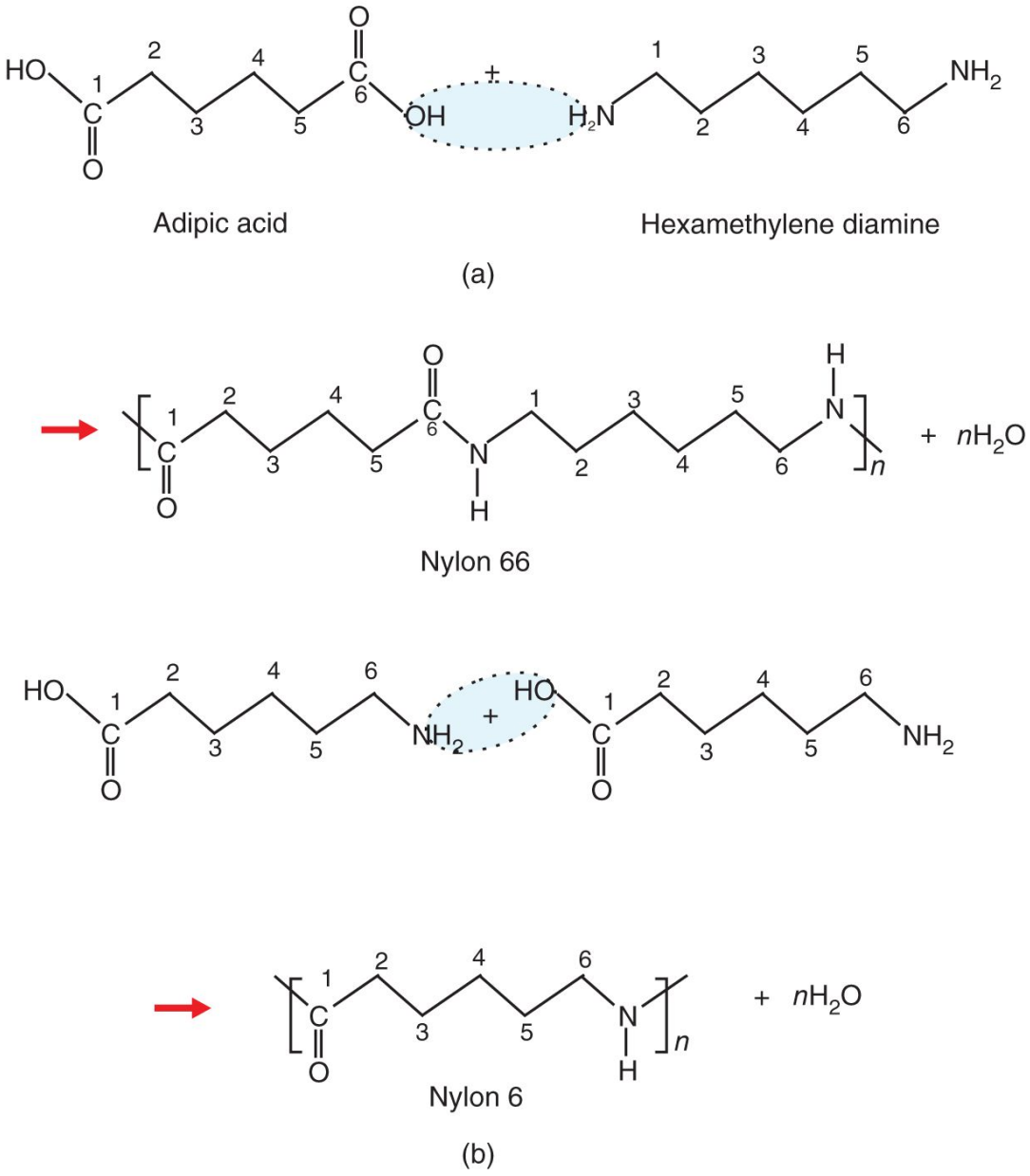


Figure 4.27 Formation of nylon, schematic: (a) nylon 66; (b) nylon 6. The carbon skeleton of the polymer chains is drawn as a zig-zag and the number of C atoms between N atoms is shown.

Nylon 6,6 is an example of an *even–even* nylon, sometimes just abbreviated to *even*. It is easy to envision other even–even polymers, such as nylon 44. Similarly, it is possible to think of *odd–odd*, or just *odd*, nylons, such as nylon 55. Although these nylons have very similar chemical properties, they differ in important electrical aspects and the nature of the chain determines whether these plastics can be used to make piezoelectric (Section 9.2) or other electronic components.

A rather similar chemical reaction produces widely used thermosetting polymers typified by epoxy resins. In this group of materials, the reactive *epoxide* group, which consists of an oxygen atom in a (usually) three-membered ring (Figure 4.28a), is opened and used to link monomers. Epoxy resin adhesives normally come as two-part mixes. The resin component contains small and medium size molecules with an epoxide group at each end, called di-epoxy molecules (Figure 4.28b,c). The resin is set by adding a cross-linking agent or *catalyst*, which is a diamine (Figure 4.28d) that joins the epoxy-containing molecules together to form a strong cross-linked network. Once the network has been formed, it is very difficult to disrupt, and epoxy resins are typical *thermoset* polymers. Note that there is no reaction product on polymerisation, which means that only a small dimensional change occurs as the precursors harden. This is of importance in applications where shrinkage or expansion would create difficulties, as in the original application of these materials, dental fillings.

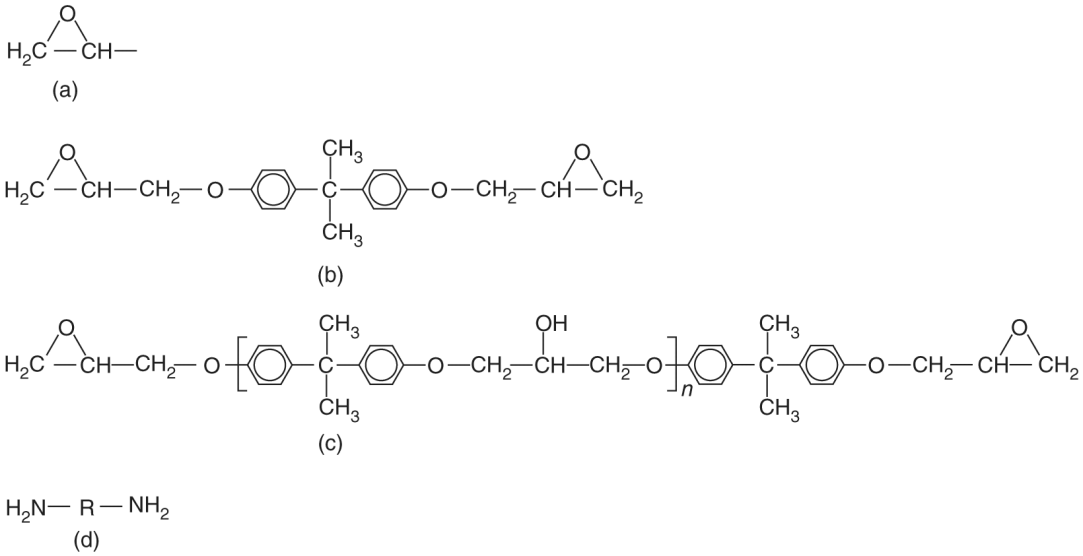


Figure 4.28 Epoxy resins: (a) epoxide group; (b) a small diepoxy molecule; (c) a small polymer molecule, with *n* up to approximately 25, found in the resin part of a two-part epoxy adhesive mix; (d) a diamine linking group (catalyst), where R represents a short chain of CH_2 groups.

The degree of cross-linking between chains changes properties dramatically. Weak cross-links tend to soften materials, while extensive cross-linking turns the material into a hard resin. Materials such the epoxy resins are heavily cross-linked into a hard mass. Cross-linking is the key to elastomeric properties. All cross-linked polymers tend to be difficult to disrupt, and they can be regarded as one giant molecule. This makes them of use when durability is needed. Many of the polymers mentioned earlier can be modified so as to form extensive cross-linking by increasing the number of reactive groups on the monomers. For example, the polyester formed by phthalic acid and glycerol (Figure 4.29) can cross-link to form a polymer that is used in bake-on car paints.

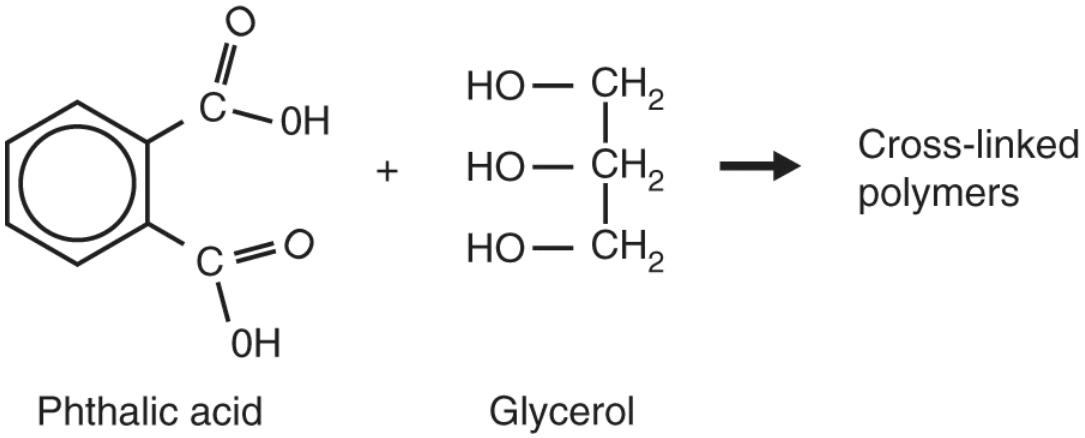


Figure 4.29 Reaction of molecules to give cross-linked polymers.

4.4.3 Microstructures of Polymers

Although the chemical makeup of polymers influences properties, by far the most important aspect is that of microstructure. The most significant features of polymer microstructure are chain length, chain branching, chain side-groups (which contribute to chain stiffness), and strength of cross-links between chains. The degree of crystallinity of the polymer, which depends on the factors just listed, is also of considerable importance. In fact, the strength of most thermoplastic polymers depends critically upon this latter factor.

Unlike simple organic molecules, polymers consist of long chains of varying length that *cannot* be characterised by a constant molar mass. Indeed, many cross-linked polymers can be thought of as a single molecule, so that the molar mass is the total mass of the polymer. However, it is helpful to have a measure of the degree of polymerisation or the distribution of chain lengths in a polymer, and this is given in terms of an *average* molar mass. Ideally, the distribution of chain lengths in a sample of a polymer will take on the shape of a bell-shaped curve. As the chain length is reflected in the mass of the molecule, this information is often given as a graph of number of molecules against the molar mass (Figure 4.30a). Generally, the distribution departs from a simple bell shape (Figure 4.30b) and the form of the resultant curve is often characteristic of a particular reaction mechanism.

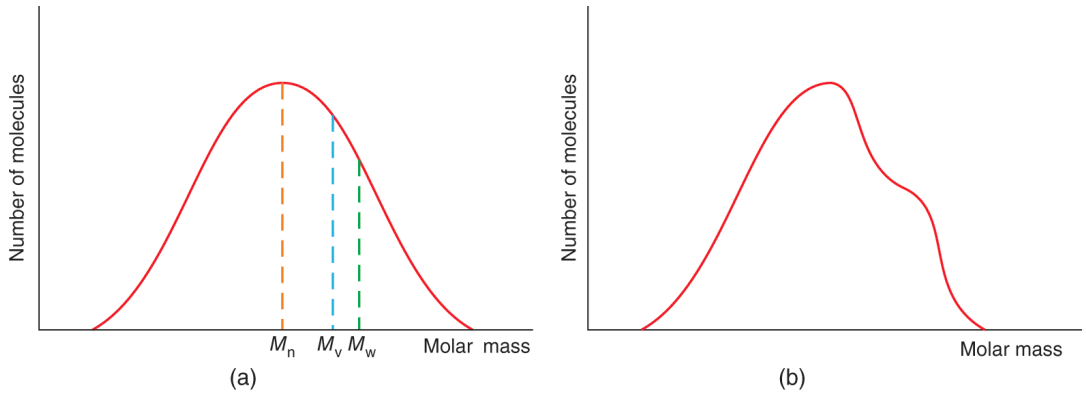


Figure 4.30 Molar mass of a polymer. (a) Idealised distribution: M_n, number average molar mass; M_v, viscosity molar mass; M_w, weight molar mass and (b) example of a real mass distribution.

In order to quantify the chain length, the molar mass must be defined statistically. The *number average molar mass*, M_n, is given by:

$$M_n = \sum_i x_i M_i$$

$$x_i = n_i \Big/ \sum_i n_i$$

where x_i is the fraction of the total number of chains within the chosen molar mass range, and n_i is the number of molecules with a molar mass M_i. The number average molar mass corresponds to the peak in a bell-shaped distribution curve. An alternative measure, the *weight average molar mass*, M_w, takes into account the fact that most of the mass of the sample resides in bigger molecules.

$$M_w = \sum_i w_i M_i$$

$$w_i = \frac{n_i M_i}{\sum_i n_i M_i}$$

where w_i is the mass fraction of each type of molecule present within the chosen molar mass range. The molar mass can also be determined experimentally. A method frequently used is to measure the viscosity of a solution containing the

polymer. A greater viscosity indicates longer chains and a higher molar mass. The molar mass determined in this way, M_v, lies between M_n and M_w.

The degree of polymerisation, N, the number of monomer units in an average chain, is given by the molar mass divided by the mass of the monomer, m. The value obtained depends upon which of the various molecular mass values are chosen.

$$N_n = M_n/m$$

$$N_w = M_w/m$$

$$N_v = M_v/m$$

Polymerisation involves the breaking and reforming of large numbers of chemical bonds and this process results in a number of broadly different molecular structures (Figure 4.31). Frequently chains are not simply linear but also have side branches, which can cross-link molecules. Growth of several chains can also start from a small nucleation centre, and the resulting chains can branch to give a *dendritic* structure. All of these molecular geometries impart different physical properties to the resultant polymer.

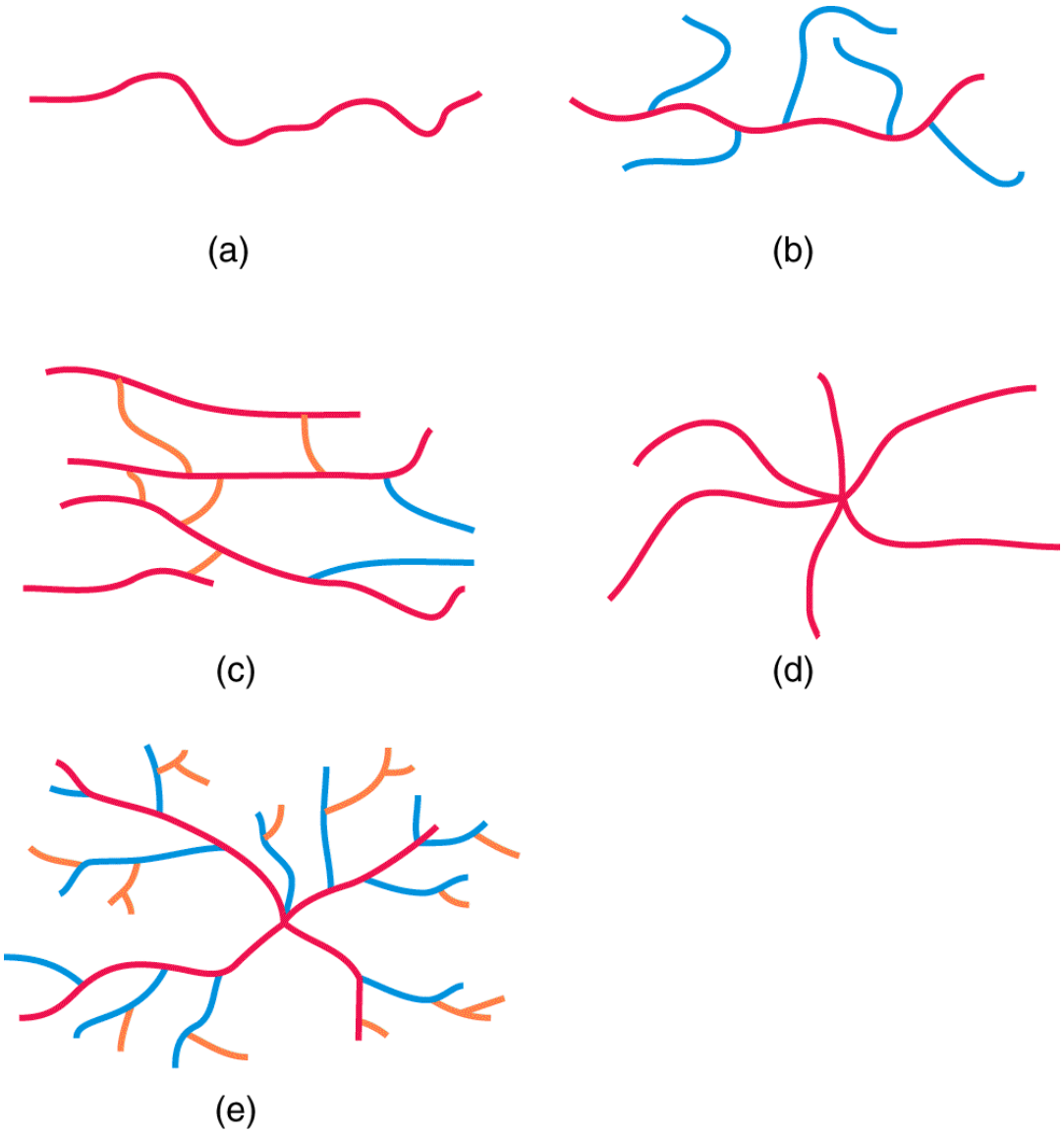


Figure 4.31 Polymer chain geometries, schematic: (a) linear; (b) branched; (c) cross-linked; (d) star; (e) dendritic.

Polyethylene (polythene) is a chain polymer 10^4 ethene (ethylene) units or more in length. If these chains are relatively short and highly branched, the material has a low density, a low refractive index and is very flexible but weak. It is referred to as *low-density polyethylene*, LDPE. *High-density polyethylene*, HDPE, consists of linear molecules, and has a molecular weight of between 200 000 and 500 000. It is much stronger than LDPE. *Ultra-high molecular weight polyethylene*, UHMWPE, with a molecular weight of the order of 5 000 000, is stronger still. The mechanical properties of polyethylene are influenced further by the degree of crystallisation that occurs. On cooling slowly from the melt, some chains can order into crystalline regions 10–20 nm thick. These crystalline

regions are of high density and of high refractive index. Most polythene is a mixture of crystalline and amorphous regions, which is why it appears milky.

Straight chains rarely occur in polymers. More often, as in solid polythene and many similar polymers, the chains fold back on themselves with a characteristic fold length, of the order of 10 nm in polythene itself (Figure 4.32a). The folded chains aggregate into blocks that have a regular structure similar to a crystal. This unit of microstructure is called a *lamella* (Figure 4.32b). The lamellae are not usually made from a single folded chain, but are formed by a variety of neighbouring chains that can be united by unfolded stretches of polymer (Figure 4.32c). The parts of the chains not incorporated into the lamellae then link one lamella to another in the partly crystalline material. During crystallisation of the melt, lamellae form in three dimensions, from a nucleation site, to form a *spherulite*. This feature consists of a set of spokes, called *lamellar fibrils*, radiating out from a common centre into the amorphous inter-spoke regions (Figure 4.33).

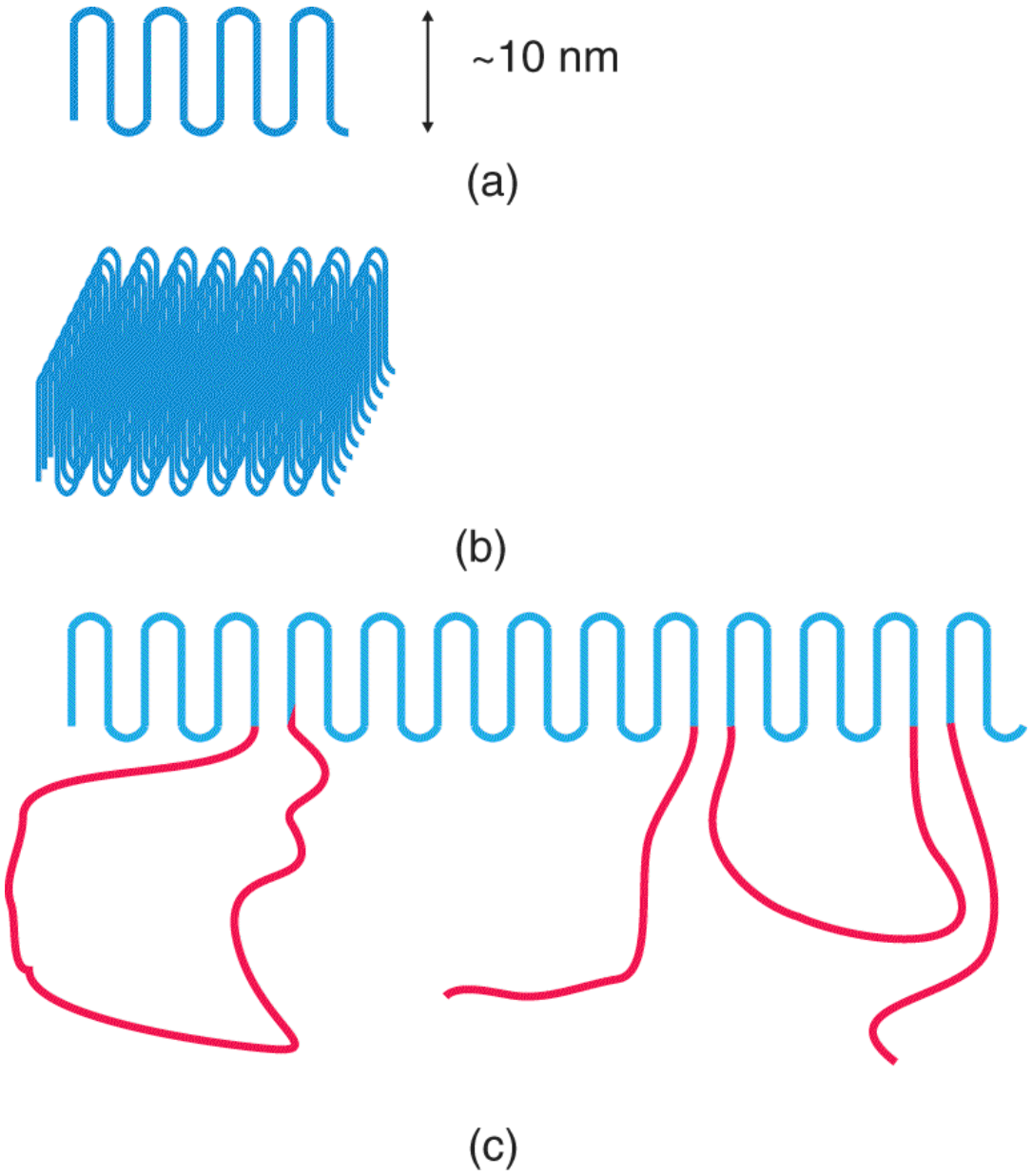


Figure 4.32 (a) Polymer chains in polyethylene, schematic; (b) folded chains forming a lamella; (c) lamella consisting of several polymer chains.

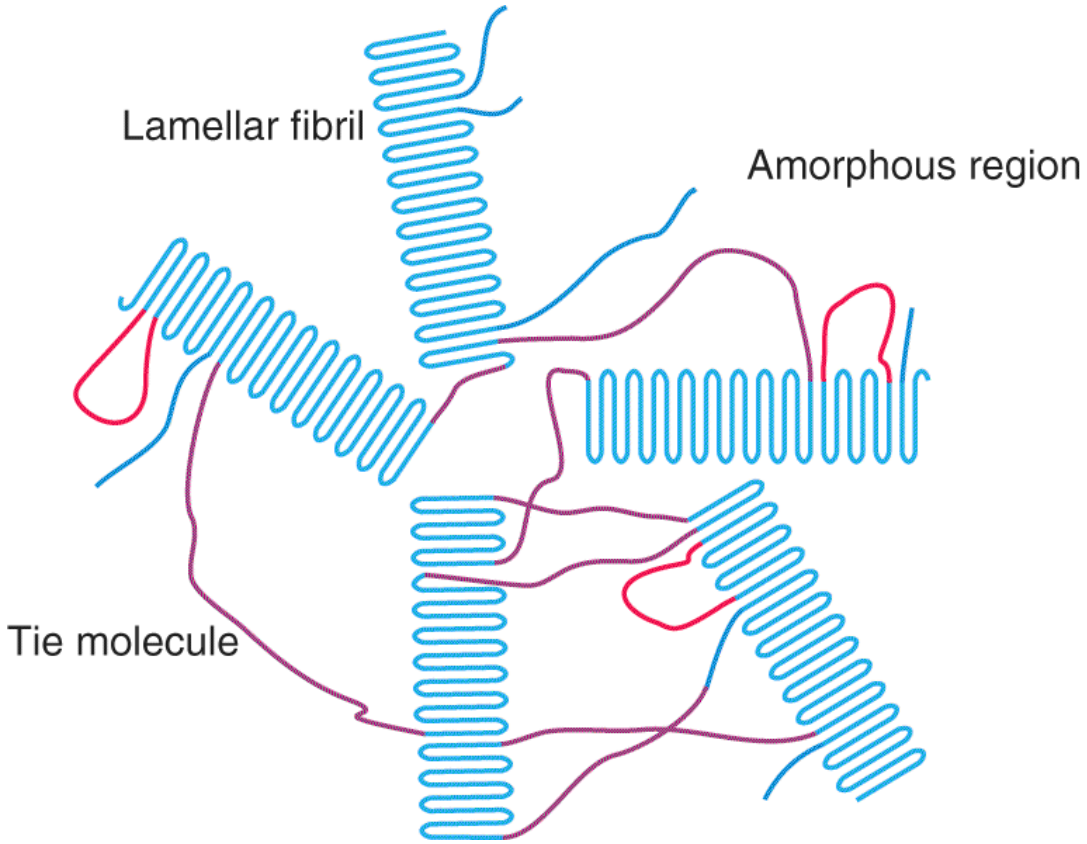


Figure 4.33 Schematic structure of a spherulite in a polymer such as polythene.

The degree to which a polymer can crystallise depends upon the details of the chain. For example, the amide group is a polar unit, and forms hydrogen bonds with the carboxyl oxygen, in nylons. These intermolecular forces hold the chains together, which produces a highly-crystalline polymer with excellent strength. The degree of crystallinity of any polymer can be determined if the density of polymer crystals and purely amorphous material is known. The reasoning is exactly the same as used in Vegard's law, for the determination of the lattice parameter of a solid solution (Section 3.2). The fraction of crystalline polymer, x_c, is:

$$x_c = (\rho_s - \rho_a)/(\rho_c - \rho_a)$$

where ρ_s is the density of the sample, ρ_c is the density of the crystals, and ρ_a is the density of completely amorphous polymer.

The disposition of side groups on a polymer chain has a great influence on properties, especially the flexibility of the chain, thus changing the melting point of the polymer and ability of the chains to pack together. These modify both strength and optical properties. Three different arrangements have been characterised; *atactic* polymers have a random arrangement of side groups,

isotactic polymers have the groups all on one side of the chain, and *syndiotactic* polymers have the side groups alternating (Figure 4.34). For example, atactic polypropylene is largely amorphous and weak. The *stereo-regular* material syndiotactic polypropylene is crystalline, transparent, and hard, while isotactic polypropylene also crystallises and readily forms fibres. Polystyrene is similar. The atactic material is amorphous whereas syndiotactic polystyrene is crystalline. Poly(methyl methacrylate), used as a replacement for glass, is almost completely amorphous.

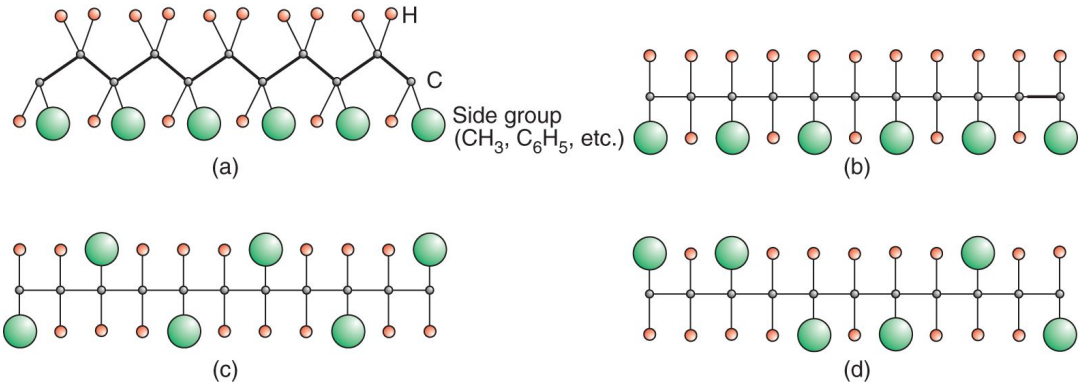


Figure 4.34 Polymer chains: (a) perspective view of an isotactic polymer chain; (b) projection of (a) from above; (c) a syndiotactic chain, viewed as in (b); (d) an atactic polymer chain viewed as in (b).

Polymer properties are considerably modified by polymerising two different monomers together to produce *copolymers*. The two components can be arranged in an alternating way, at random, in extensive blocks of just one or other polymer, or one polymer can attach as branches to the main chain of the other (Figure 4.35). The physical and chemical properties of these materials can be regarded as a combination of the properties of the two parent polymers. For example, pure atactic polystyrene is transparent and brittle. Polybutadiene (synthetic rubber) is resilient but soft. High impact polystyrene, a graft copolymer of these two materials, is durable, strong, and transparent.

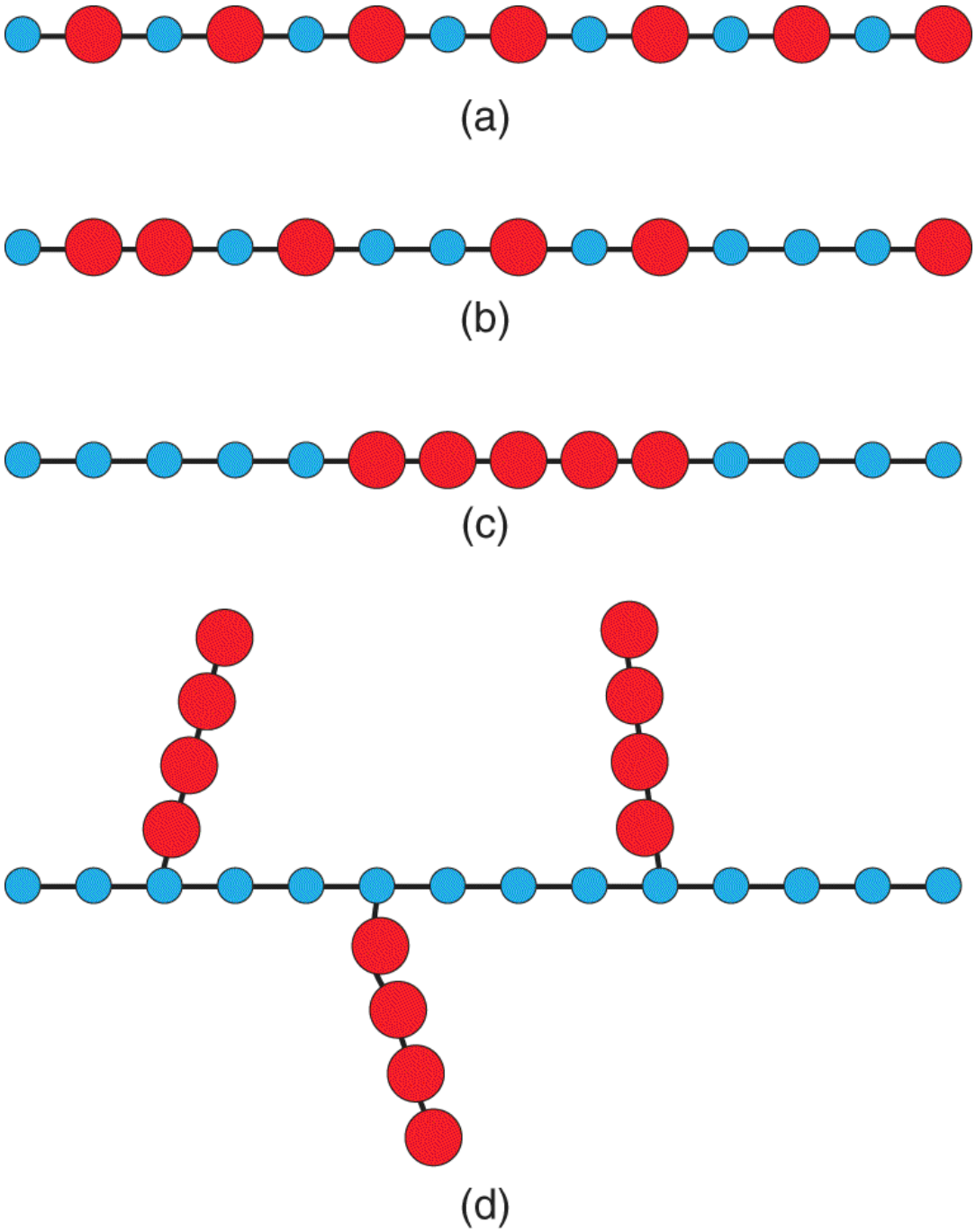


Figure 4.35 Schematic polymer geometry: (a) alternating copolymer; (b) random copolymer; (c) block copolymer; (d) graft copolymer.

4.4.4 Elastomers

Elastomers are materials that behave like rubber. They can be stretched to many times their original length, and when the force is released, they spring back to their original size and shape. Elastomers are a subgroup of amorphous thermoplastics with the distinction dependent upon the glass transition

temperature, T_g, of the polymer. Normal thermoplastics have a glass transition temperature well above room temperature and behave as hard, brittle solids. Those with a glass transition temperature below room temperature are soft and can be easily deformed at room temperature. These are elastomers. Cooling an elastomer well below its glass transition temperature will make it hard and brittle. For example, a piece of rubber tubing dipped in liquid nitrogen becomes brittle and can easily be shattered with a hammer blow.

In its normal form, the microstructure of an elastomer is a mess of jumbled coiled polymer chains (Figure 4.36a). On being stretched, the chains tend to partly align parallel to one another (Figure 4.36b). The stretched state is thermodynamically less stable than the coiled state, and the material will revert to the coiled state when the deforming stress is removed. The entropy of the coiled state is greater than that of the stretched state and provides the driving force for the material to return to the coiled configuration. However, entropy alone does not control the key 'snap back' property of elastomers. This is provided by cross-linking *a few* of the elastomer molecules using other molecules. When the elastomer is now stretched, some of the bonds in the cross-links are stretched, and these spring-back rapidly when the tension is released.

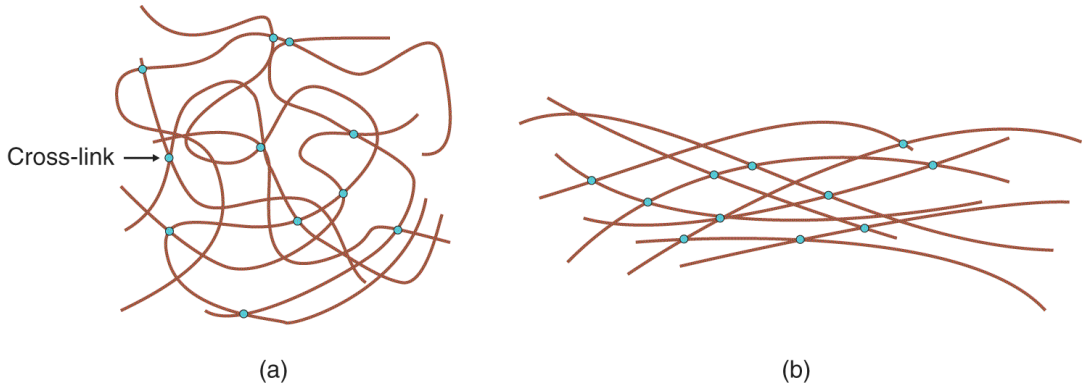


Figure 4.36 Elastomer, schematic: (a) unstretched; and (b) stretched.

The best-known elastomer is natural rubber, polyisoprene. Isoprene (Figure 4.37a) is a liquid at room temperature, which polymerises readily to give the elastomer polyisoprene (Figure 4.37b). The polymerisation produces two main geometrical isomers (molecules with the same formulae but different bond arrangements). Natural rubber is the *all cis-* form of polyisoprene (Figure 4.37c), in which the methyl ($-CH_3$) groups and hydrogen (H) atoms are on the same side of the carbon–carbon double bond. The other isomer, the *all trans-* form, called *gutta-percha* (Figure 4.37d) is also found in nature. It is harder and finds use in golf balls and dentistry.

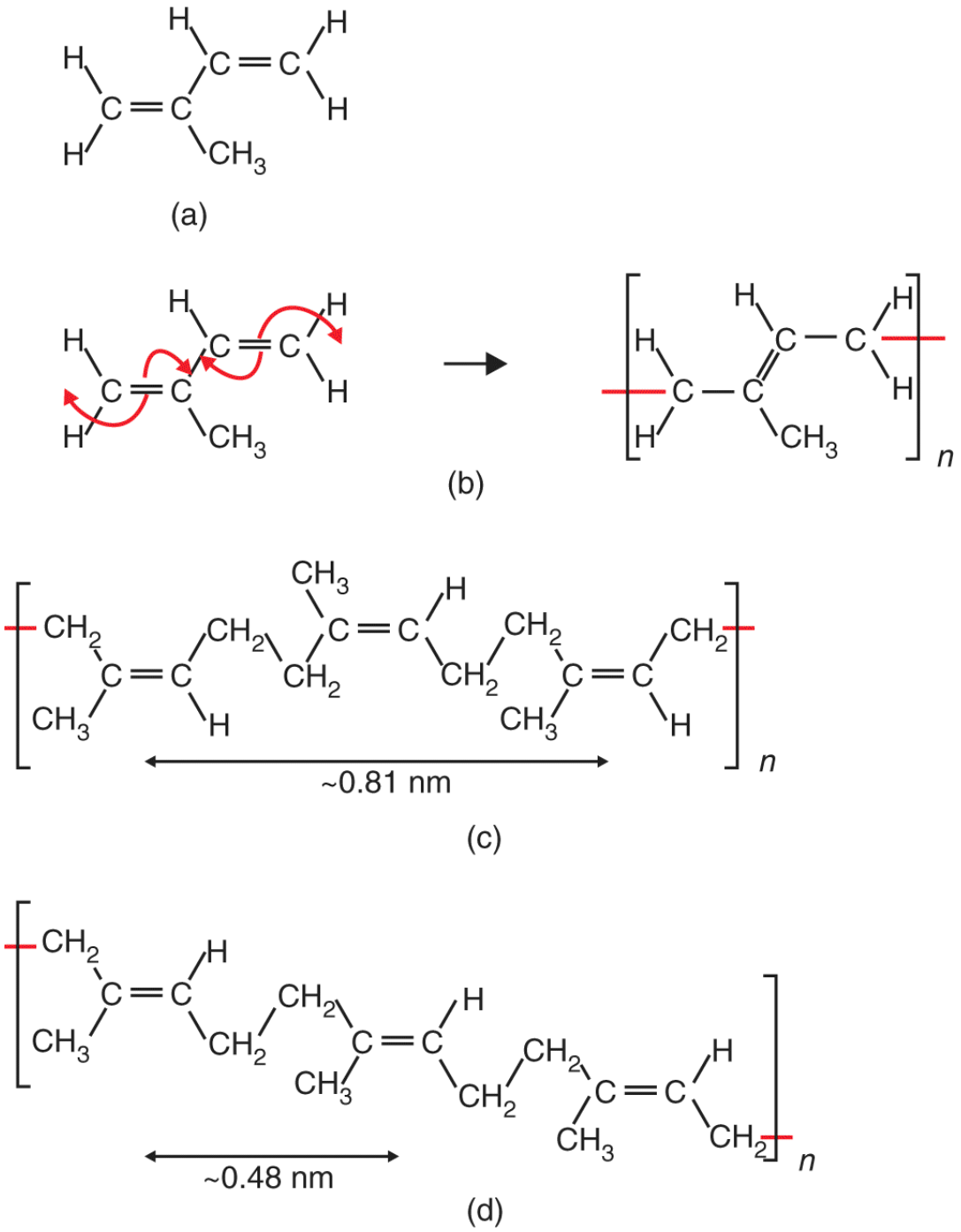


Figure 4.37 Rubber: (a) isoprene; (b) bond redistribution and polymerisation; (c) natural rubber, *all-cis*-polyisoprene; and (d) gutta-percha, *all-trans*-polyisoprene.

Rubber *latex*, the milky liquid tapped from rubber trees, is a suspension of rubber in water. (It is found in many plants, including dandelions.) The rubber in latex can be coagulated by the addition of acetic acid, to give a soft, easily oxidised material called *crepe* or *gum rubber*. In its natural state, this rubber is

sticky, like most amorphous thermoplastics at a temperature above T_g, and does not possess the typical elastomer properties of snap-back when stretched.

The practice of cross-linking the polyisoprene chains in natural rubber to form a usable elastomer was discovered by Goodyear, in 1839. He heated natural rubber latex with sulphur, a process called *vulcanisation*. This transforms the sticky runny natural material into a product in which the elastic properties are retained while the stickiness is lost. The cross-linking (Figure 4.38) utilises the remaining double bonds in the elastomer chains. These are opened by the sulphur molecules to form the ties.

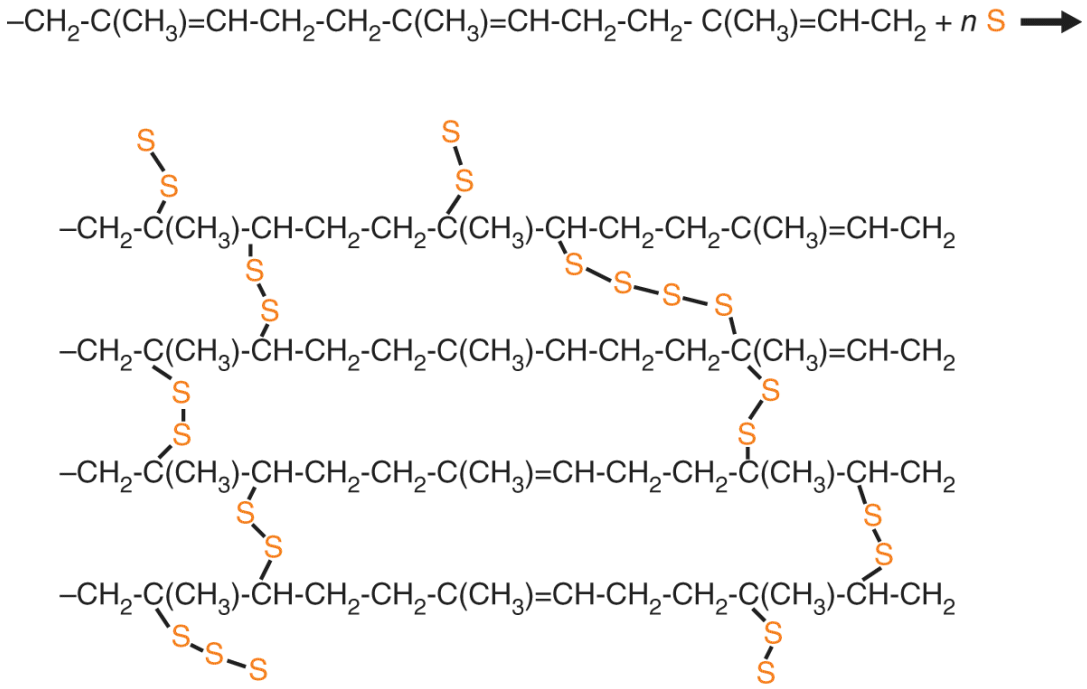


Figure 4.38 The formation of cross-links between molecules as a result of vulcanisation.

Cross-linking makes the polymer more rigid. For a soft rubber, 1% or 2% of sulphur is added. If this process is carried too far, the whole mass of polymer turns into a solid block. Hard rubbers contain up to 35% sulphur, and in essence are transformed into thermosetting polymers similar to epoxy resins by this extensive cross-linking.

A large family of artificial rubbers related to natural rubber are now produced. One of the earliest was obtained by the polymerisation of butadiene in the presence of sodium (Na), to give buna (*Bu*tadiene + *Na*) rubber. Two other widely used rubbers are neoprene (Figure 4.39a), a rubber resistant to organic solvents, prepared from a chlorinated hydrocarbon precursor and nitrile rubber (Figure 4.39b), which is a copolymer of butadiene (CH_2=CH–CH=CH_2), and propenenitrile (CH_2=CH–CN), containing a nitrile (cyanide, –CN) group.

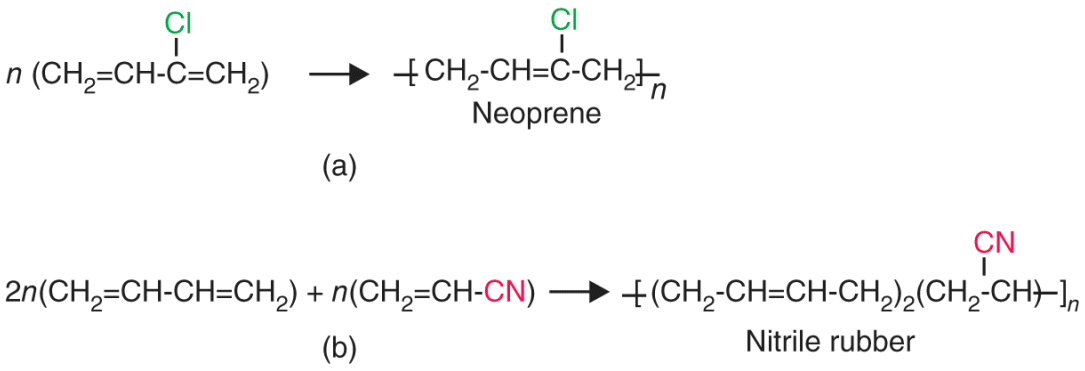


Figure 4.39 Synthetic elastomers: (a) neoprene; and (b) nitrile rubber.

4.4.5 Production of Polymers

There are two principal mechanisms for the formation of a polymer, *step growth* and *chain growth*. In step growth, growing chains can link together to form longer chains. In chain growth, monomers are added, one by one, to the growing end of the chain, which must contain an atom or group of atoms in a reactive state. This is often a *free radical*, a molecule or molecular fragment that has an unpaired electron present, as a consequence of which is extremely reactive chemically (Figure 4.40). Nylons, for example, form by a step-growth mechanism. In addition, monomer molecules or chains can link head-to-tail (the most common way), head-to-head, or tail-to-tail. Random linkage can also occur, so increasing the complexity of the reaction.

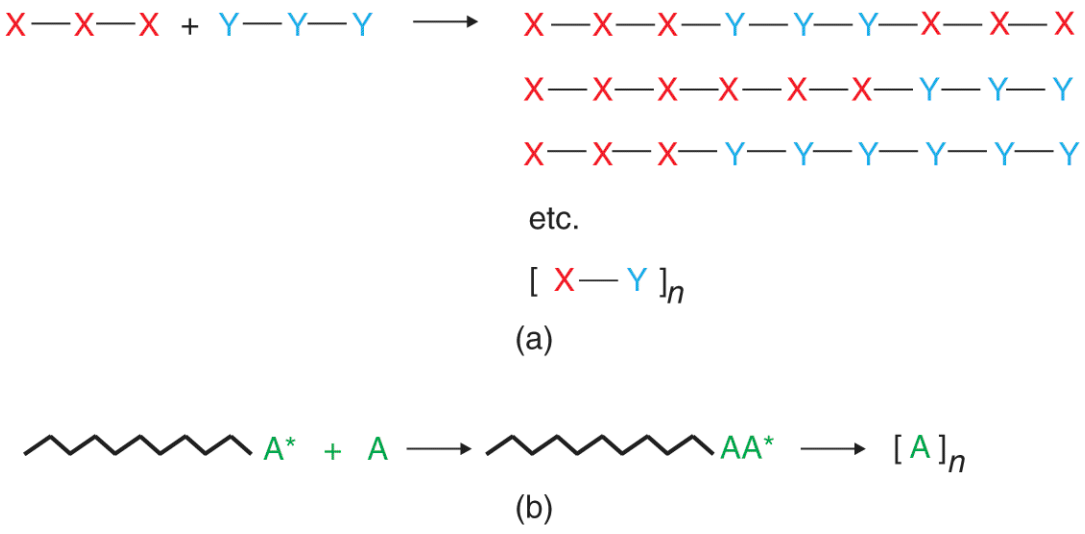


Figure 4.40 Polymerisation mechanisms: (a) step growth; (b) chain growth, A* represents a reactive centre.

To form a polymer, the initial monomers must be activated in some way to start the reaction, a step called *initiation*. This can be accomplished by heat or high-energy radiation such as ultraviolet light. These processes, which are not reproducible enough for industrial production, contribute to the degradation of

polymers in normal use. Industrially, the initiation stage is achieved by mixing a wide variety of active molecules with the monomers.

The continuing production of the polymer chains by linkage of the monomers, the second stage of the reaction, is called *propagation*. The mechanisms of propagation are complex and not all of the reaction steps are fully understood for all reactions. Nevertheless, the propagation stage is of key importance in the production of special polymers, and for this, catalysts are usually employed, described in the following text. In several cases, catalysts not only increase the rate of reaction, but also ensure that the addition of the monomers to the growing polymer chain takes place in a constrained way. That is, the reacting molecules only approach the growing chain in one direction. This leads to the production of polymers with a single tacticity, such as isotactic polystyrene. *Ziegler–Natta* and *metallocene* catalysts fall into this group and are used to prepare polymers with controlled structures.

Free radical polymerisation combines initiation and propagation into one process. This method is used to produce low density branched polyethylene, poly(methyl methacrylate), and poly(vinyl acetate). Taking ethene (ethylene) as an example, the gas is pressurised to 100 atm. at 100 °C and a small amount of an unstable initiator molecule, typically an organic peroxide or azide is added (Figure 4.41a,b). These decompose to form extremely reactive free radicals, which attack double bonds, thus initiating polymerisation (Figure 4.41c). The reaction is able to satisfy the bonding requirements of the initial free radical, but creates a new free radical in the process, which can attack another ethene molecule, leading to continued chain growth (Figure 4.41d,e). The reactions can be written as:

$$R^{\cdot} + CH_2 = CHX \rightarrow R{-}CH_2{-}CHX^{\cdot}$$

$$R{-}CH_2{-}CHX^{\cdot} + CH_2 = CHX$$

$$\rightarrow R{-}CH_2{-}CHX - CH_2{-}CHX^{\cdot}$$

where $R^{\cdot}$ represents an organic free radical. Free radical polymerisation can occur with many, but not all, ethene derivatives. Because there are no constraints on how the chain grows, the polymers are atactic.

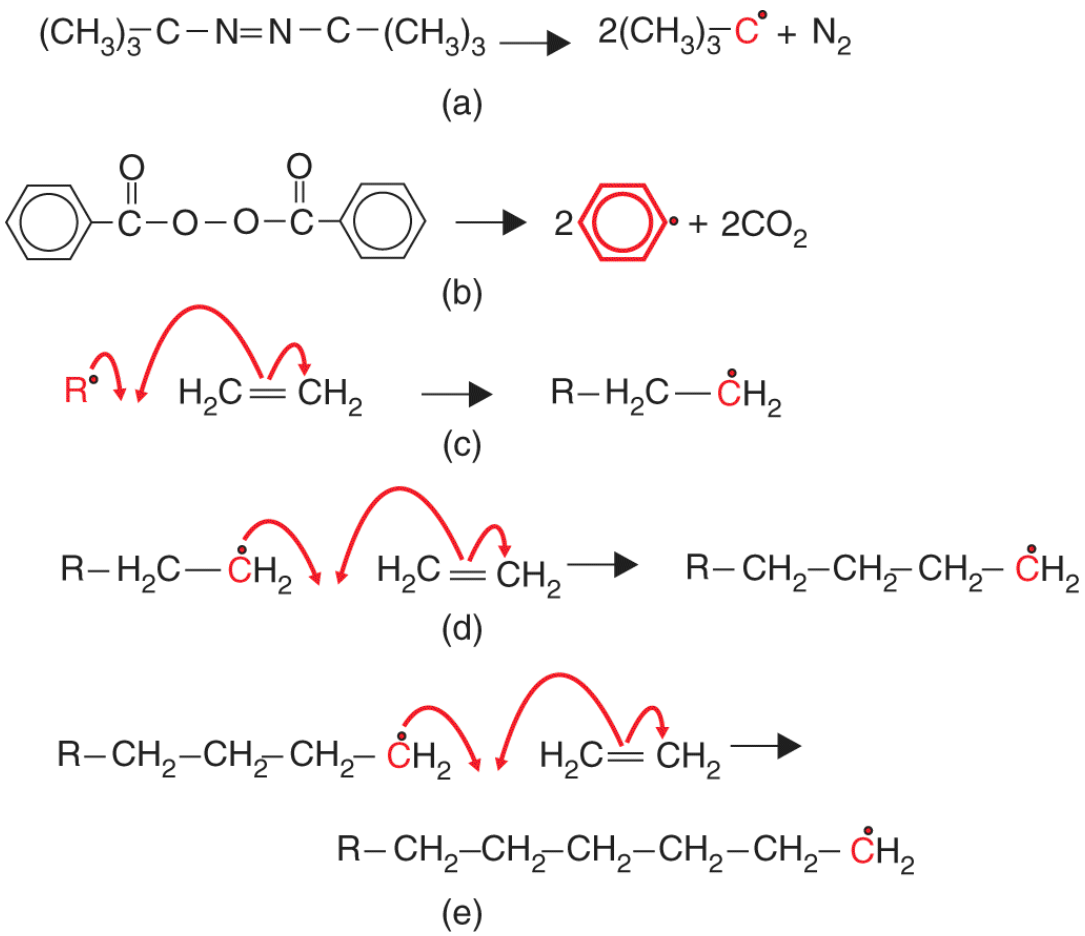


Figure 4.41 The generation of free radicals (*) by bond breaking: (a) an azide; (b) a peroxide. (c, d, e) Reactions of a free radical R* with ethene.

Growing polymer chains must ultimately stop growing. This is brought about by the process of *termination*. Chain termination can come about in two ways. The simplest is for two growing chains to meet and join. In such cases, the reactive end of the chains, often free radicals with unpaired electrons, unite when the free electrons form bonds joining the chains (Figure 4.42a). A second mechanism, *disproportionation*, allows one growing chain to take a hydrogen atom from another chain. The resultant is that one chain terminates with a $-CH_3$ group and one in a double bond (Figure 4.42b). Further reaction at this double bond can continue to lengthen this chain. The hydrogen can also be extracted from the middle of a chain. This will terminate one chain in a $-CH_3$ group, and create a free radical in the interior of the chain. This reactive region can act as a centre for new chain growth and continued polymerisation will produce a branch (Figure 4.42c). This is a very common occurrence in polyethylene polymerisation and results in the highly branched low density (LD) polythene, described earlier. In order to make much stronger non-branched high density (HD) polythene, different preparation methods need to be used.

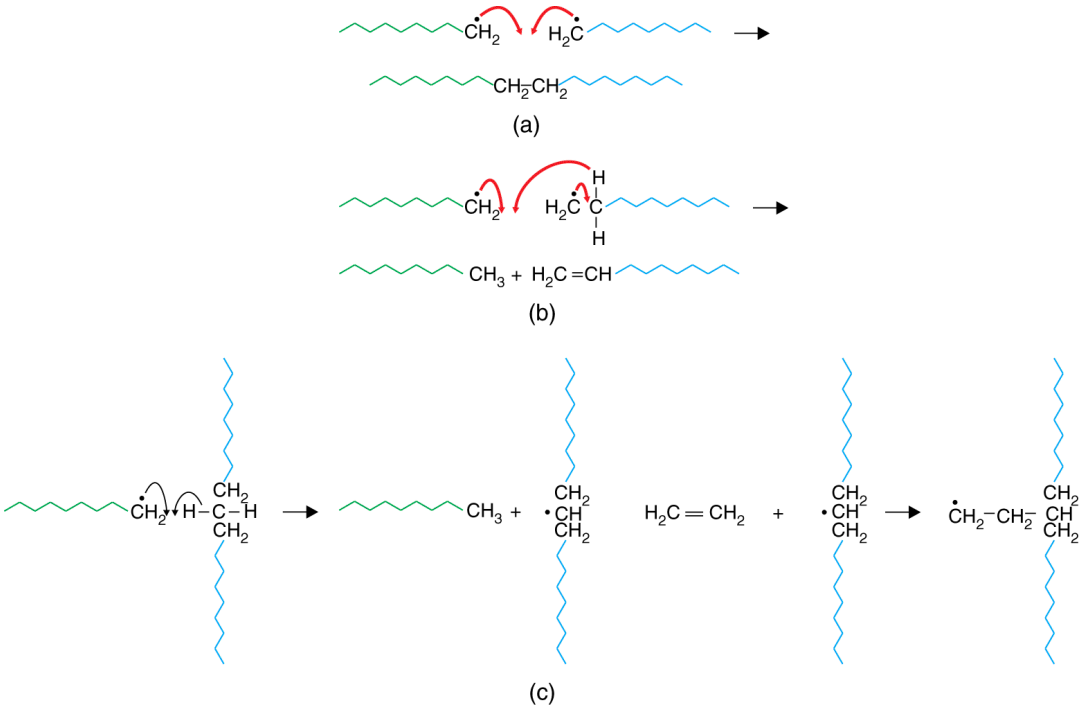


Figure 4.42 Chain termination: (a) two chains meeting; (b) disproportionation; (c) disproportionation leading to branching.

Metallocenes are derived from the cyclopentadiene anion, $[C_5H_5]^-$ (Figure 4.43a). These are stable units, in which a delocalised π molecular orbitals lie above and below the plane of the pentagon of carbon atoms, allowing strong chemical bonds to form with metal cations. The first metallocene investigated, *ferrocene*, contained Fe^{2+} sandwiched between the anions (Figure 4.43b). Catalysts for polymer production are derived from a molecule called *zirconocene*, containing Zr^{4+} cations, in which the Zr is bonded to two $[C_5H_5]^-$ anions and two Cl^- anions (Figure 4.43c). In order to produce polymers with a precise structure, the components surrounding the Zr^{4+} ion are carefully modified. A change in the cyclopentadiene anions alters the geometry of the approach of monomers to the cation. The point where the polymer chain grows is controlled by replacement of the Cl^- ions with methyl (CH_3) groups, as the methyl group acts as the point of attachment of successive monomer molecules.

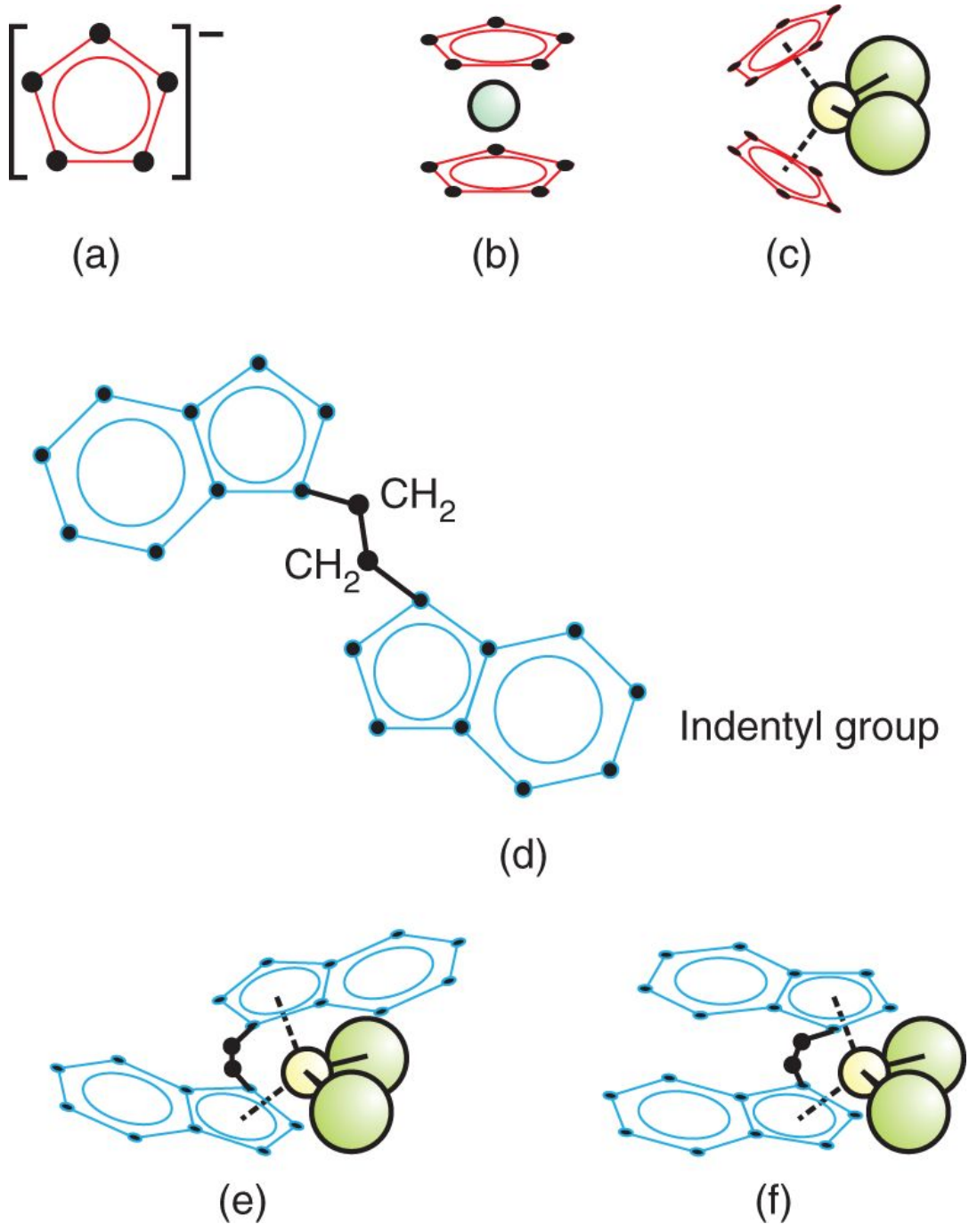


Figure 4.43 Metallocene catalysis: (a) the C atom framework of $[C_5H_5]^-$; delocalised orbitals represented by a circle inside the framework; (b) ferrocene molecule; (c) zirconocene molecule; (d) two identyl groups linked by two CH_2 groups; (e) molecular geometry required to produce isotactic polypropylene; and (f) molecular geometry required to produce atactic polypropylene.

In the molecule used to produce polypropylene (Figure 4.43d), a complex structure called an *indentyl group* (a benzene ring linked to a cyclodiene)

replaces each of the cyclopentadiene groups. The indentyl groups are actually linked by a short chain of two CH_2 groups. When these are opposed (Figure 4.43e), the incoming monomers are guided into a position where only isotactic polypropylene can form. In molecules in which the bulky groups are on the same side of the Zr^{4+} cation (Figure 4.43e,f), the approach of the monomers is variable, and atactic polypropylene is produced.

4.4.6 Organic Framework Structures: MOFs and COFs

In recent years a large effort has gone into the synthesis of highly crystalline porous organic solids. The first of these to be developed were MOFs. These materials are polymers formed by the self-assembly of a wide range of metal cations, including the 3d transition metals and lanthanoids, linked by organic molecules. These latter have to have at least two or more reactive binding sites in order to build 2D and 3D frameworks. Typically, they are carboxylic acids, azoles, and pyridines, although many others have also been used. The resulting materials, besides being crystalline, have pores which, depending upon the starting molecules, can range from the order of 0.1 to 50 nm in diameter. Vast numbers of MOFs are now known including those with structures that mimic the silicate zeolites.

In outline, these structures can be represented by a regular arrangement of metal nodes, joined by connecting molecules to make a framework (Figure 4.44a). Many thousands of framework structures have now been synthesised. The first well characterised phase, named MOF-5, was built from Zn_4O units (Figure 4.44b) in which four Zn atoms surround a central O atom. Each Zn atom is, itself, tetrahedrally coordinated by oxygen arising from the linking molecules 1,4 benzene dicarboxylic acid (Figure 4.44c,d). The Zn_4O groups are thus linked into a cubic array by the amalgamation with the organic linker molecules (Figure 4.44e).

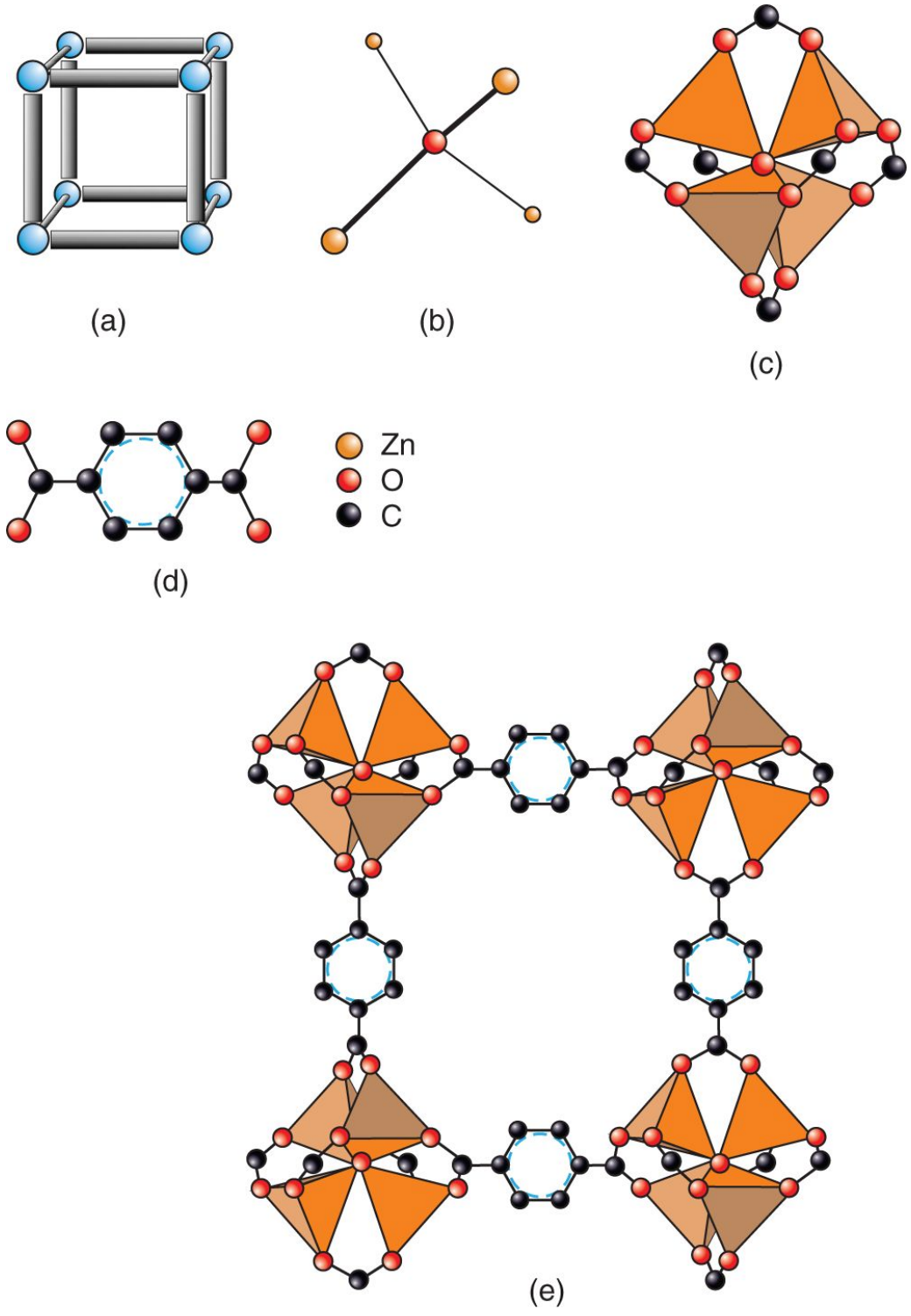


Figure 4.44 MOFs: (a) schematic structure, consisting of metal nodes and organic linkers; (b) Zn_4O unit; (c) Zn_4O unit and neighbouring atoms in MOF-5; (d) 1,4 benzene dicarboxyl unit; and (e) MOF-5 single layer, schematic. The H atoms are omitted for clarity.

COFs are similar to MOFs, but the metal units are replaced by covalent organic molecules. The production of these open frameworks uses similar mechanisms to that in conventional polymerisation, condensation reactions between esters, amides, and ethers, via stepwise growth. The condensation of a bifunctional monomer or the condensation of two different monomers with two reactive groups will produce linear polymers, described earlier. The reaction of molecules with more reactive groups will produce two-dimensional (2D) sheets or 3D solids. In both cases, careful design of the reacting molecules can produce effectively 2D porous sheets or 3D porous frameworks, with pore sizes ranging from nanoporous, with pore sizes less that approximately 0.2 nm, to mesoporous, with pore sizes much greater than this.

For example, the condensation of phenyldiboronic acid (1,4 benzene diboronic acid), $C_6H_4[B(OH)_2]_2$ (DBA) and 2,3,6,7,10,11-hexahydroxytriphenylene (hexahydroxytriphenylene) (HHTP) (Figure 4.45a,b) yields the 2D crystalline framework COF-5, with pore of approximately 2.7 nm in diameter (Figure 4.45c). The basic architecture of the porous framework can be maintained while increasing the pore size by increasing the length of the linking molecules; demonstrated by the condensation of HHTP with 4,4′-diphenylbutadienebis(boronic acid) (DPB), to form HHTP-DPB COF (Figure 4.45e) with a pore diameter of 4.7 nm.

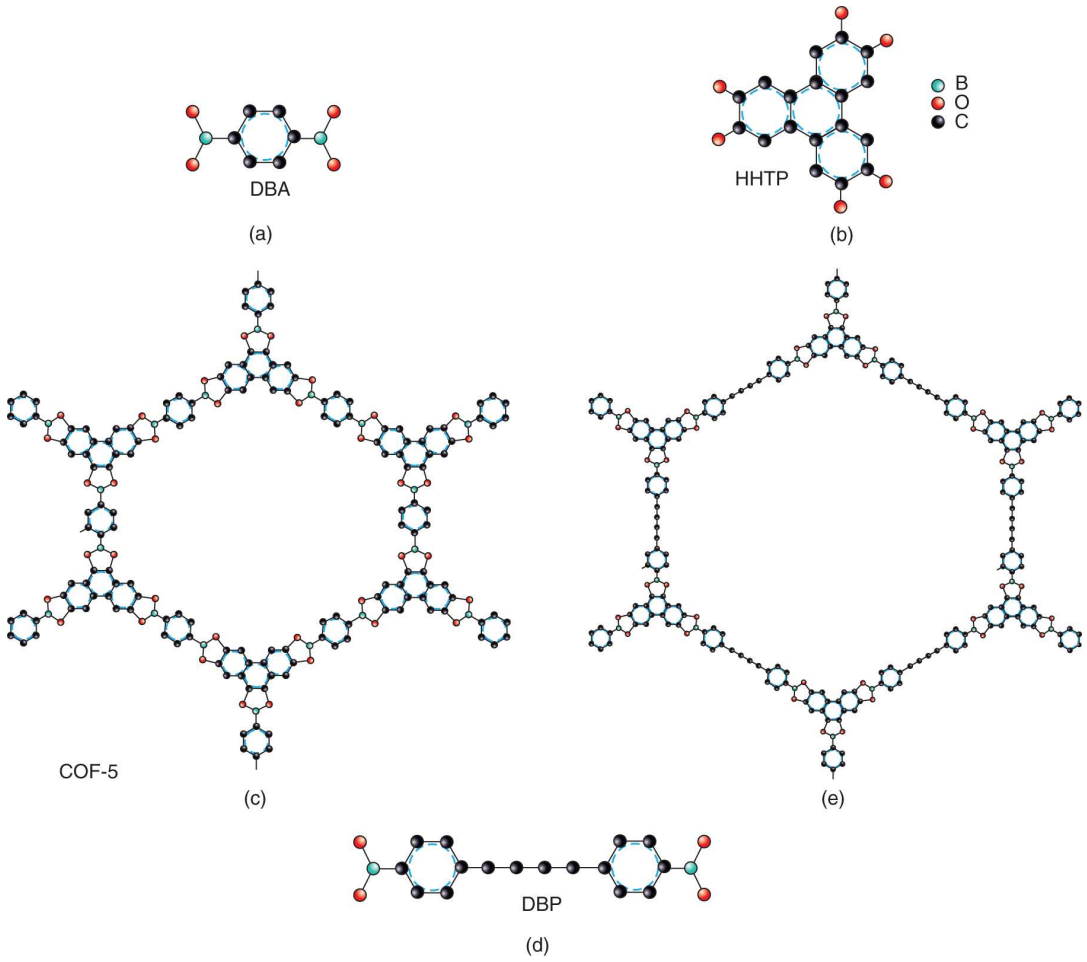


Figure 4.45 Covalent organic frameworks: (a) DBA; (b) HHTP; (c) MOF-5; (d) DBP; and (e) HHTP-DPB COF. H atoms are omitted for clarity.

The production of a glassy solid is usually achieved by a rapid quenching of a melt. This same technique has been used to prepare glass-like solids by rapidly quenching MOFs that are in the liquid state, rather than by slow cooling, which results in crystalline materials. These can be represented with reference to a group of MOFs known as zeolitic imidazolate frameworks. These are derived from a tetrahedral unit analogous to the Zn_4O unit described earlier. Here, the metal in these nodes is usually Zn or Co. The imidazole ($C_3N_2H_3$) units act as linking groups (Figure 4.46). In the slow cooled crystalline phases, these tetrahedral units are regular and link to form a crystalline framework. In the rapidly cooled liquid, they are distorted and link together in a typically glassy network. Annealing of these glasses can restore the crystalline state.

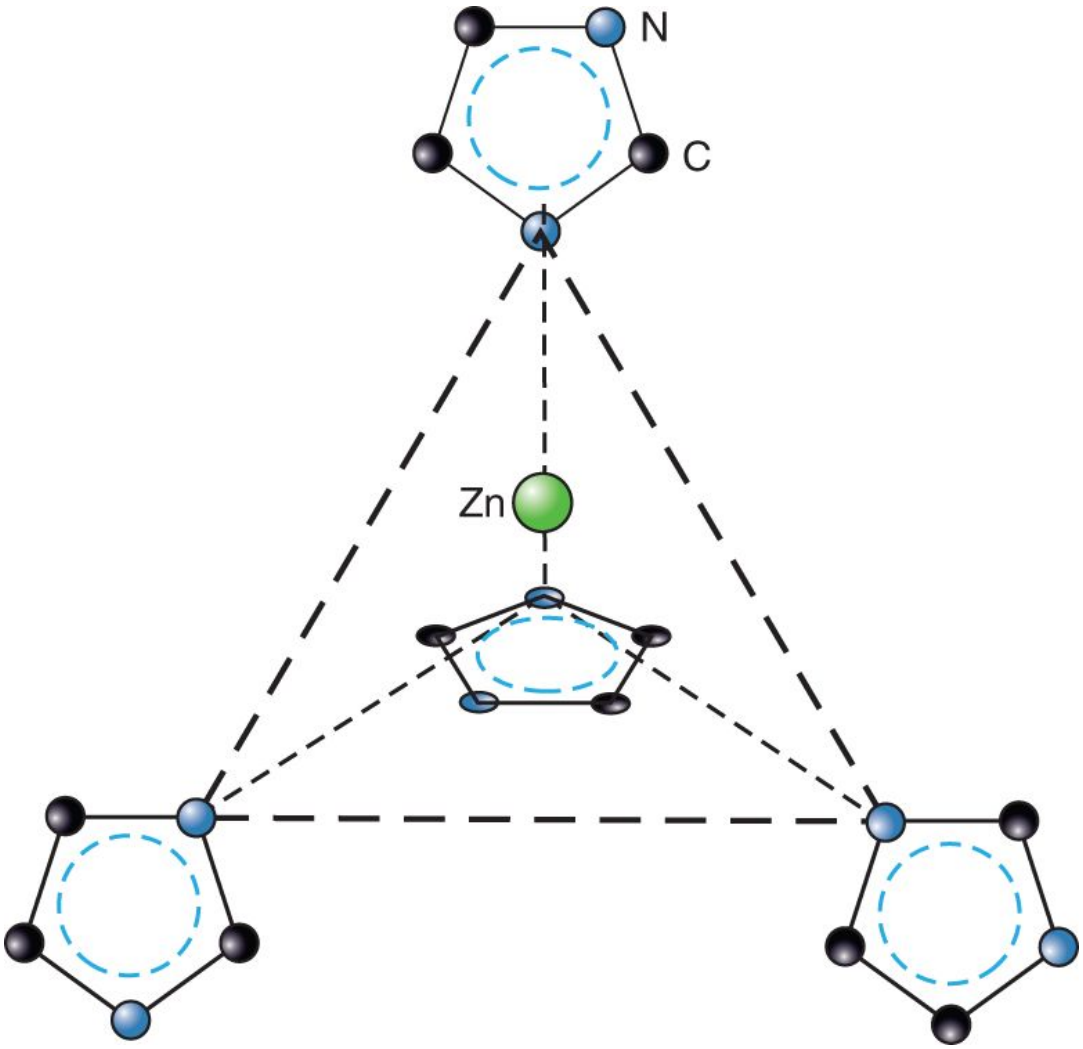


Figure 4.46 The basic tetrahedral unit found in zeolitic imidazolate framework structures and glasses. H atoms are omitted for clarity.

4.4.7 The Principal Properties of Polymers

The properties of ordinary polymers depend both upon the details of the carbon chain of the polymer molecule and upon the way in which these chains fit together. Polymers consist mainly of carbon and hydrogen, and one of the principal properties of this group of materials is that they are low-density solids. This is enhanced by the fact that the polymer chains often do not pack together very closely, although the crystalline regions in polymers increase overall density.

Many properties of polymers depend upon the functional groups that occur on the chains. Nylon fibres gain strength from >NH…O=C< hydrogen bonding. This also allows them to absorb water and adds to the comfortable feel of these materials when used in clothing. The –NH– group in the polymers can pick up

charge from the water to give $-NH_2^+-$. This is noticeable as electrostatic charge build-up on carpets and clothing.

Other properties of polymers can be varied widely using changes in processing and additives. The introduction of a small number of cross-links into a thermoplastic elastomer can change it from a sticky substance to a useful rubber. Similarly, although polymers are inherently insulators, doping can turn them into electronic or ionic conductors of electricity, allowing these materials to be used in batteries.

The behaviour of polymers under stress is very variable and depends upon the polymer structure and microstructure. Thermoplastic polymers are brittle at low temperatures and easily deformed and plastic at higher temperatures. This is because the polymer chains are not linked to each other. In the brittle state, cracks can easily pass right through the solid, between the chains. At higher temperatures, the molecules can easily slip past each other. Crystalline regions in thermoplastics oppose these tendencies and add appreciably to the strength of the solid. Cross-linking, characteristic of thermosetting polymers, prevents molecular movement and results in solids that combine good mechanical strength with chemical stability.

One of the principal properties shown by polymers relates to this aspect of chemical stability. Polymers do not degrade rapidly and are an increasing pollution problem in the world's oceans. The ease of degradation of a polymer rapidly decreases as the degree of cross-linking of the polymer chains increases. The heart of the problem is that the carbon–carbon backbone of polymers is extremely resistant to chemical attack as the bonds are so strong. The same is true of the carbon–hydrogen bonds that make up much of the rest of the molecules. In order to make a polymer more degradable, weak links must be introduced into the chain. To some extent, double bonds are more susceptible to attack than single bonds, and elastomers with only one double bond are less likely to be attacked and degraded than those rubbers with two double bonds available. Degradability can be enhanced by the introduction of deliberate points of attack into the polymer. For example, inclusion of oxygen atoms, or hydroxyl or acid groups, allow for water penetration and aid bacterial attack.

The porosity of MOFs and COFs adds another dimension to the properties that can be exploited. The polarity, size, shape, and chemical composition of pore surfaces can be modified, allowing these solids to be used for gas storage, gas separation, and catalysis.

As with normal plastics, MOFs are basically electronic insulators and a flat band model or molecular orbital model of electronic behaviour can be used. However, the modification of linker molecules is able to greatly influence these electronic properties so that porous semiconductors, of use in thermoelectrics, and ionic conducting frameworks, of use as battery materials, can be synthesised. Molecular inclusion in the pores gives the possibility of many more properties.

Thus, for photoluminescence, light emission in an MOF can arise from the metal cations, the linker molecules, or the incorporation of luminescent molecules within the pore. The perceived light emission from such adaptable materials can very over a wide range of wavelengths and include ‘white’ light.

4.5 Composite Materials

4.5.1 Reinforced Materials

Composites are solids made up of more than one material, designed to have enhanced properties compared with the separate materials themselves. Composites are the norm in nature. For example, wood is composed of strong flexible cellulose plus stiff lignin; bone consists of strong soft collagen (a protein) plus apatite (a brittle hard ceramic). Artificial composites are increasingly widely used for advanced engineering applications, from aircraft, to high-technology leisure items such as skis and sails. One of the earliest applications of a composite was the Macintosh raincoat. The fabric for this garment consisted of a sandwich of natural rubber between two sheets of a woven natural polymer, cotton.

Man-made composites fall into three broad classes, depending upon whether the main part of the composite, the *matrix*, is a polymer, a metal, or a ceramic. Often, but not always, composites combine materials from two classes, as in glass fibre reinforced plastics. However, the most widely used composite material, concrete, is a ceramic–ceramic composite. The most important classes of artificial composites are described in the following text. The mechanical properties of composites are outlined further in Section 8.6.

Fibre-reinforced plastics tend to use thermosetting resins, especially polyester and epoxy resins, as the matrix material. Polyester resins are relatively inexpensive, but tend to shrink during curing and tend to absorb water. Epoxy resins are more expensive, but do not shrink on curing and are fairly resistant to water penetration. In principle, any highly cross-linked polymer would make a potential matrix. The resins are reinforced by filling with fibreglass, carbon fibre or strong polymer fibres such Kevlar, an aramid fibre.

Fibres alone tend to be brittle, and although they have good tensile strength, they cannot sustain compression readily. The purpose of the matrix, therefore, is to hold the fibres together in the desired orientation. The purpose of the fibres is to add strength. The resultant strength depends upon the type of fibre utilised and geometrical factors. These include the amount of fibre added and the length and orientation of the fibres (Figure 4.47a,b), as well as the bonding between the matrix and the fibre inserts. Composites are strong in the direction of alignment, and weaker normal to this direction. To overcome this, the orientation of successive layers of fibres is often changed to form a *laminate* (Figure 4.47c,d).

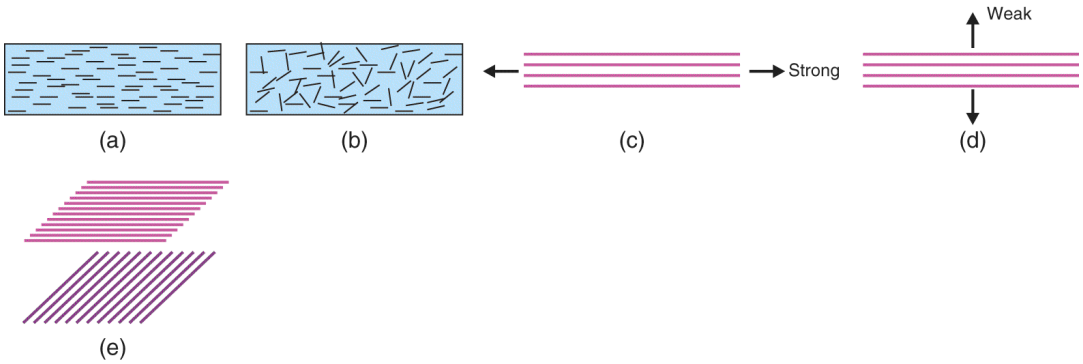


Figure 4.47 Fibre-reinforced composites: (a) aligned fibres; (b) random orientation fibres; (c) fibres tend to be strong along the alignment and weak normal to this direction; and (d) laminates are designed to offset this.

Metal–matrix composites are similar to their plastic counterparts. Metals are frequently reinforced with continuous fibres to give improvement in strength. The fibres used are ceramic, for example, silicon carbide or alumina, or metallic, such as boron or tungsten. However, as fibres of these materials are difficult to fabricate, unless superior performance is vital, it is easier to make composites using small particles of a hard material such as alumina or silicon carbide. These ceramics are usually mixed with the molten alloy that is then formed into intended shapes. Among the most widespread of metal-particle composites are the *cemented carbide* materials used as cutting tools for machining steel. The first of these was made from a matrix of the metal cobalt and contained hard tungsten carbide particles as the hard additive. Interestingly, this was the first combination tried by the producer, and although many other metal–metal carbide combinations were investigated later, the cobalt–tungsten carbide composite remained the best for most purposes.

Ceramic–matrix composites are utilised to overcome the inherent brittleness of ceramics. The reinforcement consists of fibres or particles. The materials used include silicon carbide and alumina. The toughening comes about because the fibres or particles deflect or bridge cracks in the matrix. Naturally occurring ceramic–ceramic composites include granite and marble.

Although glass in composite materials is mostly associated with the strengthening component in the form of fibres, laminated glass is a widely used composite. This material consists of a thin plastic film sandwiched between two or more sheets of glass. The purpose of the polymer is to prevent the glass splintering on impact. It is widely used as bullet-proof glass.

4.5.2 Cement and Concrete

Concrete is a composite material, made from *cement paste* and *aggregate* (a coarse stony material). Cement paste is simply a mixture of cement and water alone, while if the aggregate is sand the mixture of cement, water, and sand is called *mortar*. Early forms of cement, which are mainly composed of calcium

hydroxide, partly transformed to calcium carbonate, predate Roman times. The material was made from limestone (impure calcium carbonate), which was heated or *burned* to give *quicklime* or *burnt lime* (calcium oxide).

$$CaCO_3(s) \rightarrow CaO(s) + CO_2(g)$$

Quicklime reacts with water to release considerable heat, a process called *slaking*, to give *slaked lime*, calcium hydroxide.

$$CaO(s) + H_2O(l) \rightarrow Ca(OH)_2(s)$$

The slaked lime was mixed to a paste with water and used to cement sand or stone. The paste slowly reacted with carbon dioxide in the air to give calcium carbonate again.

$$Ca(OH)_2(s) + CO_2(g) \rightarrow CaCO_3(s) + H_2O(l)$$

The Romans improved the process by adding volcanic ash to the limestone, to produce more durable cement, which is still intact to this day.

Portland cement appears to have first been formulated by Joseph Aspdin of Leeds, and was patented by him in 1824. It was so named because it resembled expensive Portland Stone (at least to the eye of the inventor). A number of manufacturers worked on improving the material but the substance now generally regarded as Portland cement was developed by J. Aspdin's son William in the 1840s. It is made from about 80% limestone and about 20% clay. It was widely adopted because it possessed superior qualities to the older quicklime-based materials, including the especially important property of being able to harden in damp conditions. This latter property was especially valuable at a time when tunnel construction, for, among other projects, the London Underground system, was widespread.

To make cement powder, the raw materials are ground with water to form a *slurry*. This is heated in a kiln at gradually increasing temperatures to initially drive off water, then to decompose the calcium carbonate,

$$CaCO_3(s) \rightarrow CaO(s) + CO_2(g)$$

As the temperature increases, other reactions take place, and the reaction products partly melt and sinter, to produce *clinker*. In the final stage of manufacture, the cooled clinker is ground, and about 2–5% gypsum ($CaSO_4 \cdot 2H_2O$), is added to produce cement powder.

Portland cement powder contains five major constituents. These are complex minerals and are known by their chemical names, mineral names, and by a shorthand notation (Table 4.6). There are also traces of other impurities in

ordinary cement powder, which may have a large effect on the final strength and durability of the concrete.

Table 4.6 Typical Portland cement constituents.

Chemical name	Mineral name	Chemical formula	Notation	Composition (wt%)
Tricalcium silicate	Alite	Ca_3SiO_5	C_3S	40–65
Dicalcium silicate	Belite	Ca_2SiO_4	C_2S	10–20
Tricalcium aluminate		$Ca_3Al_2O_6$	C_3A	10
Tetracalcium aluminoferrite		$Ca_4Al_2Fe_2O_{10}$	C_4AF	10
Calcium sulphate dihydrate	Gypsum	$CaSO_4.2H_2O$	CSH_2	2–5

Cement hardens when the constituents react with water to produce an interlocking array of hydrated crystals. The main reactions are rather indeterminate, because of the variable quantities of water and cement powder involved, and because the reactions that take place are extremely complex. Broadly speaking, these are:

Reaction of alite and water:

$$\begin{aligned} \text{Alite} + \text{water} &\rightarrow \text{calcium hydroxide} \\ &\quad + \text{calcium silicate hydrate} \\ Ca_3SiO_4 + H_2O &\rightarrow Ca(OH)_2 \\ &\quad + (CaO)_x(SiO_2)_y \cdot aq. \end{aligned}$$

This chemical equation is not balanced, because in practice, x varies between the approximate limits of 1.8–2.2, y varies around a mean value of 1.0 and $aq.$ means that water is also combined in the material in an indeterminate amount. The idealised composition of the calcium silicate hydrate phase is $Ca_2SiO_4 \cdot aq.$

The reaction is rapid, and continues for up to approximately 20 days.

Considerable heat is evolved – of the order of 500 J g^{-1} of powder – and care must be taken to remove this heat when forming large masses of concrete into structures. The reaction rapidly yields a high strength product (Figure 4.48).

Reaction of belite and water

$$\text{Belite} + \text{water} \rightarrow \text{calcium hydroxide} + \text{calcium silicate hydrate}$$

$$Ca_2SiO_5 + H_2O \rightarrow Ca(OH)_2 + (CaO)_x(SiO_2)_y \cdot aq.$$

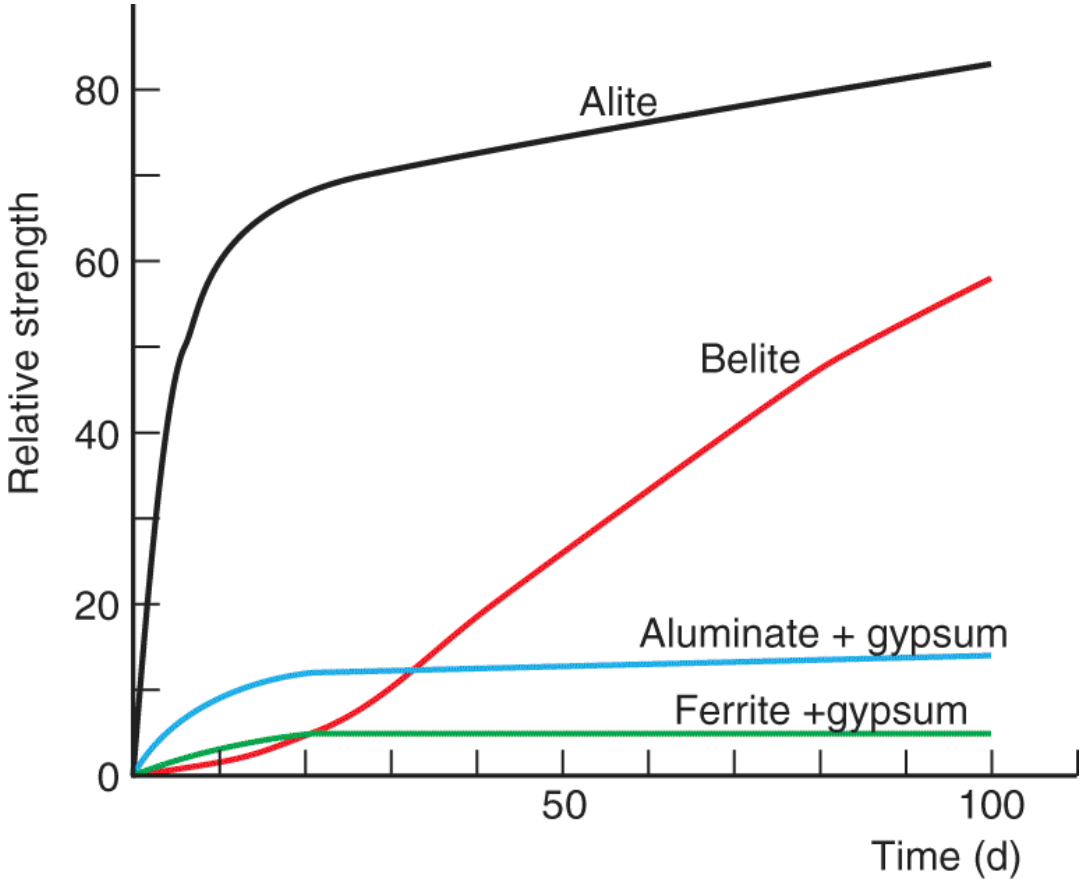


Figure 4.48 The approximate relative strengths of the components of Portland cement after hydration, as a function of time elapsed.

The reaction, like that of alite, is imprecise, and the chemical equation is not balanced. Belite reacts slowly, taking about one year to harden, and is responsible for the long-term strength of concrete (Figure 4.48). The heat liberated, approximately 250 J g^{-1} of powder, is not as great as that liberated in the reaction of alite, and because the reaction is slower, does not have such immediate consequences during construction.

Reaction of aluminate and water

$$\text{Aluminate} + \text{water} \rightarrow \text{calcium aluminate hydrate}$$

$$Ca_3Al_2O_6 + 6H_2O \rightarrow Ca_3Al_2O_6 \cdot 6H_2O$$

This is a very rapid reaction, and is completed in minutes. It is also a very energetic reaction, and about 900 J g^{-1} is released during the hydration. The product is of very low strength and is attacked by sulphate, a common impurity in many soils and rocks.

Reaction of ferrite and water

Ferrite and water react very slowly. Moderate amounts of heat are evolved, about 300 J g^{-1}. The product is not attacked by sulphate and is mainly beneficial to the strength of the cement. However, the main importance of ferrite is cosmetic, as it influences the colour of the product.

Reaction of gypsum with aluminate and water

$$\text{Gypsum} + \text{aluminate} + \text{water} \rightarrow \text{etringite}$$

$$Ca_3Al_2O_6 + 3CaSO_4 \cdot 2H_2O + 30H_2O \rightarrow Ca_6Al_2S_3O_{18} \cdot 32H_2O$$

Gypsum is important as it reacts with aluminate to give *ettringite* (calcium aluminium sulphate hydrate), $Ca_6Al_2S_3O_{18} \cdot 32H_2O$ [also written $Ca_6Al_2(SO_4)_3(OH)_{12} \cdot 26H_2O$], on the surface of the aluminate grains. This slows the reaction of aluminate with water and allows the wet cement paste to be worked for longer.

The hydration of cement involves a number of exothermic reactions, which liberate a great deal of heat, and when building substantial concrete structures this must be removed to prevent cracking or other deterioration. The evolution of heat takes place over a period, and the rate of heat evolution is as important as the total amount of heat given out. Several empirical relationships between the composition of the cement, the heat of hydration H, and the time elapsed have been developed. These take the typical form:

$$H = A\,x_{C3S} + B\,x_{C2S} + C\,x_{C3A} + D\,x_{C4AF}$$

where x is the weight fraction of the appropriate constituent, and A, B, C, and D are empirical constants that vary over time, reflecting the changes in the composition of the cement as it hardens. The heat of hydration is measured in J g^{-1} of cement. For example, the heat of hydration after three days, $H(3\text{ days})$, and after one year, $H(1\text{ year})$, is:

$$\begin{aligned} H(3\text{ days}) &= 240\,x_{\mathrm{C3S}} + 50\,x_{\mathrm{C2S}} \\ &\quad + 880\,x_{\mathrm{C3A}} + 290\,x_{\mathrm{C4AF}} \\ H(1\text{ year}) &= 490\,x_{\mathrm{C3S}} + 225\,x_{\mathrm{C2S}} \\ &\quad + 1160\,x_{\mathrm{C3A}} + 375\,x_{\mathrm{C4AF}} \end{aligned}$$

The microstructures that form as the cement paste reacts with water are important in controlling the final strength of the concrete. Initially water reacts to give a silicate gel. This material is amorphous and produces a glutinous coating on the powder particles, holding them together and causing a certain amount of swelling to occur. The gel slowly crystallises, to give a mass of interpenetrating needles and plates. Gypsum reacts slowly to form hexagonal needle-like prisms of ettringite, which further interlock the mass. In addition, the material contains free water, at least in the early stages of reaction, and pores. The details of these reactions still remain to be completely worked out for all the constituents present.

The microstructure of concrete is further complicated by the presence of the aggregate. Although this is often supposed to be inert chemically, it can react with the other constituents and with water, particularly if this contains acidic or alkaline impurities.

Reference

Saitoh, K., Tanaka, M., Tsai, A.P., and Rossouw, C.J. (2004). ALCHEMI studies on quasicrystals. *JEOL News***39** (1): 20.

Further Reading

Metals:

Smallman, R.E. and Ngan, A.H.W. (2007). *Physical Metallurgy and Advanced Materials*, 7e. Oxford: Elsevier.

Bulk Metallic Glasses:

Inoue, A., Kong, F.L., Zhu, S.L., and Greer, A.L. (2020). Multicomponent bulk metallic glasses with elevated-temperature resistance. *Mater. Res. Soc. Bull.***44**: 867.

Pan, J., Yu. P. Ivanov, Zhou, W.H. et al. (2020). Strain-hardening and suppression of shear-banding in rejuvenated bulk metallic glass. *Nature***578**: 559.

Suryanarayana, C. and Inoue, A. (2017). *Bulk Metallic Glasses*, 2e. Boca Raton, FL: CRC Press.

Ceramics and Glass:

Carter, C.B. and Norton, M.G. (2007). *Ceramic Materials: Science and Engineering*. New York: Springer Science+Business Media.

Huang, P.Y., Kurasch, S., Alden, J.S. et al. (2013). Imaging atomic rearrangements in two-dimensional silica glass: watching silica's dance. *Science***342**: 224.

Zheng, Q. and Mauro, J.C. (2017). Viscosity of glass-forming systems. *J. Am. Ceram. Soc.***100**: 6–25. and the references therein.

Zeolites:

Li, H., Qiu, C., Ren, S. et al. (2020). Na^+-gated water-conducting nanochannels for boosting CO_2 conversion to liquid fuels. *Science***367**: 667.

IZA Structure Commision. A data base of zeolite structures is at: www.iza-structure.org/databases.

Polymers:

Bates, F.S., Hillmyer, M.A., Lodge, T.P. et al. (2012). Multiblock polymers: panacea or Pandora's box? *Science***36**: 434–460.

Ebewele, R.O. (2000). *Polymer Science and Technology*. Boca Raton, FL: Chapman and Hall, CRC Press.

Young, R. and Lovell, P. (2011). *Introduction to Polymers*. Boca Raton, FL: CRC Press.

Metal-organic Frameworks:

Cui, Y., Zhang, J., He, H., and Qian, G. (2018). Photonic functional metal-organic frameworks. *Chem. Soc. Rev.***47**: 5740.

Dineă, M. and Leonard, F. (2016). Metal-organic frameworks for electronics and photonics. *MRS Bull.***41**: 854. and references therein.

Furukawa, H., Cordova, K.E., O'Keeffe, M., and Yaghi, O.M. (2013). The chemistry and applications of metal-organic frameworks. *Science***341**: 974.

Covalent Organic Frameworks:

Alahakoon, S.B., Diwakara, S.D., Thompson, C.M., and Smaldone, R.A. (2020). Supramolecular design in 2D covalent organic frameworks. *Chem. Soc. Rev.***49**: 1344.

Chen, X., Geng, K., Liu, R. et al. (2020). Covalent organic frameworks: chemical approaches to designer structures and built-in functions. *Angew. Chem. Int. Ed.***59**: 5050.

Gu, Y., Zhao, J., and Johnson, J.A. (2020). Polymer networks: from plastics and gels to porous frameworks. *Angew. Chem. Int. Ed.***59**: 5022.

Guan, X., Chen, F., Fang, Q., and Qiu, S. (2020). Design and applications of three-dimensional covalent organic frameworks. *Chem. Soc. Rev.***49**: 1357.

Composites:

Barbero, E.J. (2011). *Introduction to Composite Materials Design*, "2nd edition". Boca Raton, FL: Taylor and Francis, CRC Press.

Problems and Exercises

Quick Quiz

4.1 In the face-centred cubic structure the atoms are in contact along:

a. A cube body diagonal.

b. A cube face diagonal.

c. A cube edge.

4.2 Ordered alloy structures such as Cu_3Au are formed by:

a. Rapidly cooling the alloy.

b. Rapidly heating the alloy.

c. Annealing the alloy.

4.3 Steel is an alloy of iron and carbon in which the carbon occupies

a. Substitutional sites.

b. Interstitial sites.

c. Substitutional and interstitial sites.

4.4 In each unit cell of the face-centred cubic (A1) structure there are:

a. 4 octahedral sites.

b. 8 octahedral sites.

c. 12 octahedral sites.

4.5 In each unit cell of the body-centred cubic (A2) structure there are:

a. 4 octahedral sites.

b. 6 octahedral sites.

c. 8 octahedral sites.

4.6 The tetrahedral sites in interstitial alloys can be occupied by:

a. Hydrogen.

b. Nitrogen.

c. Both hydrogen and nitrogen.

4.7 The inclusion of semimetals in a mixture:

a. Hinders the formation of metallic glasses.

b. Aids the formation of metallic glasses.

c. Has no effect upon the formation of metallic glasses.

4.8 Isomorphous replacement in silicate ceramics creates:

a. Substitutional defects.

b. Interstitial defects.

c. No defects.

4.9 Silicates are stable because of strong bonds between:

a. Oxygen and oxygen.

b. Metal atoms and silicon.

c. Silicon and oxygen.

4.10 Ionic silicates contain isolated:

a. Silicate groups.

b. Silicate chains.

c. Silicate sheets.

4.11 Clays are silicates containing:

a. Chains of $[SiO_4]$ + hydroxyl.

b. Networks of $[SiO_4]$ + hydroxyl.

c. Sheets of $[SiO_4]$ + hydroxyl.

4.12 Ruby is a gemstone consisting of aluminium oxide containing small amounts of:

a. Titanium impurity.

b. Chromium impurity.

c. Titanium and iron impurities.

4.13 The formation of a glass during ceramic production is called:

a. Vitrification.

b. Sintering.

c. Glass–ceramic formation.

4.14 A refractory ceramic is one that is:

a. Difficult to process.

b. Particularly hard.

c. Able to withstand high temperatures.

4.15 Pyrex glass is also known as:

a. Flint glass.

b. Borosilicate glass.

c. Soda-lime glass.

4.16 The glass transition temperature marks the point at which:

a. A glass transforms from a solid to a viscous liquid.

b. The glass can be moulded and blown.

c. The glass becomes stable.

4.17 In glass technology, small atoms such as P and B are known as:

a. Network modifiers.

b. Network formers.

c. Intermediates.

4.18 A glass ceramic is:

a. A glass processed at high temperatures.

b. A transparent ceramic.

c. A ceramic containing both crystals and glass.

4.19 Polystyrene is an example of:

a. An addition polymer.

b. A condensation polymer.

c. An elastomer.

4.20 Nylon 6,6 is an example of:

a. An addition polymer.

b. A condensation polymer.

c. An elastomer.

4.21 Nylons are:

a. Polyamides.

b. Polyesters.

c. Polycarbonates.

4.22 The molar mass of a polymer:

a. Is a fixed number.

b. Varies between narrow limits.

c. Varies between wide limits.

4.23 The molar mass of a monomer:

a. Is a fixed number.

b. Varies between narrow limits.

c. Varies between wide limits.

4.24 A lamella in a polymer is:

a. A spherulite.

b. A chain.

c. A crystalline region.

4.25 A polymer in which the side groups lie on alternate sides of the polymer chain backbone is:

a. Isotactic.

b. Syndiotactic.

c. Atactic.

4.26 The growth of polymer chains by the joining of existing chains is called:

a. Chain growth.

b. Step growth.

c. Link growth.

4.27 The ‘snap back’ property of elastomers is due to:

a. Hydrogen bonding between chains.

b. A few strong cross-links between chains.

c. Large numbers of cross-links between chains.

4.28 Ceramic matrix composites are designed to overcome:

a. The weight of ceramics.

b. The brittle nature of ceramics.

c. The inertness of ceramics.

4.29 The initial hardening of Portland cement is attributed to:

a. Alite.

b. Belite.

c. Gypsum.

4.30 The main source of heat when Portland cement hardens is due to the reaction of:

a. Tricalcium silicate.

b. Dicalcium silicate.

c. Tricalcium aluminate.

4.31 The long-term hardening of Portland cement is attributed to the presence of:

a. Tricalcium silicate.

b. Dicalcium silicate.

c. Tricalcium aluminate.

4.32 Metal-organic framework (MOF) structures are:

a. Amorphous.

b. Glass-like.

c. Crystalline.

4.33 Covalent-organic framework (COF) structures are produced via polymeric:

a. Condensation reactions.

b. Addition reactions.

c. Substitution reactions.

Calculations and Questions

4.1 The radius of a gold atom is 0.144 nm. Gold adopts the fcc (A1) structure. Estimate the unit cell parameter of gold crystals.

4.2 The radius of a lead atom is 0.175 nm. Lead adopts the fcc (A1) structure. Estimate the unit cell parameter of lead crystals.

4.3 The radius of a palladium atom is 0.138 nm. Palladium adopts the fcc (A1) structure. Estimate the unit cell parameter of palladium crystals.

4.4 The radius of an iridium atom is 0.136 nm. Iridium adopts the fcc (A1) structure. Estimate the unit cell parameter of iridium crystals.

4.5 Nickel adopts the fcc (A1) structure, with a lattice parameter of 0.3524 nm. Estimate the metallic radius of nickel in this structure.

4.6 Rhodium adopts the fcc (A1) structure, with a lattice parameter of 0.3803 nm. Estimate the metallic radius of rhodium in this structure.

4.7 The radius of a barium atom is 0.224 nm. Barium adopts the bcc (A2) structure. Estimate the unit cell parameter of barium crystals.

4.8 The radius of a niobium atom is 0.147 nm. Niobium adopts the bcc (A2) structure. Estimate the unit cell parameter of niobium crystals.

4.9 Vanadium adopts the bcc (A2) structure, with a lattice parameter of 0.3024 nm. Estimate the metallic radius of vanadium in this structure.

4.10 Potassium adopts the bcc (A2) structure, with a lattice parameter of 0.5321 nm. Estimate the metallic radius of potassium in this structure.

4.11 At room temperature, iron adopts the bcc (A2) structure with a lattice parameter of 0.28665 nm.

a. Estimate the metallic radius of iron in this structure.

b. Ignoring thermal expansion, estimate the lattice parameter of the fcc (A1) allotrope that exists above 912 °C.

4.12 At room temperature, calcium adopts the fcc (A1) structure with a lattice parameter of 0.5588 nm.

a. Estimate the metallic radius of calcium in this structure.

b. Ignoring thermal expansion, estimate the lattice parameter of the A2 allotrope that exists above 445 °C.

4.13 At room temperature, strontium adopts the fcc (A1) structure. The metallic radius of strontium in this structure is 0.215 nm.

a. Estimate the lattice parameter of the fcc structure of strontium.

b. Ignoring thermal expansion, estimate the lattice parameter of the bcc (A2) allotrope that exists above 527 °C.

4.14 The metallic radius of magnesium, which adopts the hcp (A3) structure, is 0.160 nm. Estimate the value of the lattice parameters a and c (ideal).

4.15 The metallic radius of rhenium, which adopts the hcp (A3) structure, is 0.138 nm. Estimate the value of the lattice parameters a and c (ideal).

4.16 The a lattice parameter of titanium, which adopts the hcp (A3) structure, is 0.2951 nm. Estimate the radius of titanium in this structure and the ideal value of the parameter c.

4.17 The a lattice parameter of beryllium, which adopts the hcp (A3) structure, is 0.2286 nm. Estimate the radius of beryllium in this structure and the ideal value of the parameter c.

4.18 At room temperature, hafnium adopts the hcp (A3) structure, with lattice parameters a = 0.3195 nm, c = 0.5051 nm.

a. Estimate the metallic radius of hafnium in this structure.

b. Ignoring thermal expansion, estimate the lattice parameter of the bcc (A2) allotrope that exists above 1742 °C.

4.19 At room temperature, yttrium adopts the hcp (A3) structure, with lattice parameters a = 0.3648 nm, c = 0.5732 nm.

a. Estimate the metallic radius of yttrium in this structure.

b. Ignoring thermal expansion, estimate the lattice parameter of the bcc (A2) allotrope that exists above 1481 °C.

4.20 The rapidly cooled form of the alloy CuAu has the fcc (A1) structure, in which the metals atoms are distributed at random over the available sites. The unit cell parameter is 0.436 nm. Estimate the density of the alloy.

4.21 Cartridge brass is a composition in the phase range of the alloy α-brass, which is made up with 30 wt% zinc and 70 wt% copper. The alloy has the fcc (A1) structure, in which the metal atoms are disordered over the available sites. The density is 8470 kg m^{-3}. Estimate the lattice parameter of the alloy.

4.22 Which of the following metals would be expected to form the most extensive substitutional solid solution with nickel, fcc (A1) structure, metallic radius 0.1246 nm, electronegativity 1.8?

a. Cobalt, hcp (A3) structure, metallic radius 0.125 nm, electronegativity 1.7.

b. Chromium, bcc (A2) structure, metallic radius 0.128 nm, electronegativity 1.4.

c. Platinum, fcc (A1) structure, metallic radius 0.139 nm, electronegativity 2.1.

d. Silver, fcc (A1) structure, metallic radius 0.145 nm, electronegativity 1.8.

4.23 Which of the following metals would be expected to form the most extensive substitutional solid solution with copper, fcc (A1) structure, metallic radius 0.1278 nm, electronegativity 1.8?

a. Aluminium, fcc (A1) structure, metallic radius 0.143 nm, electronegativity 1.5.

b. Palladium, fcc (A1) structure, metallic radius 0.138 nm, electronegativity 2.0.

c. Vanadium, bcc (A2) structure, metallic radius 0.135 nm, electronegativity 1.4.

d. Titanium, hcp (A3) structure, metallic radius 0.146 nm, electronegativity 1.6.

4.24 The viscosity of a soda-lime glass is given in the table. Estimate the glass transition temperature.

Viscosity (dPa s)	Temperature (°C)
5×10^{14}	450

5×10^7	700
1×10^4	1050
1×10^2	1450

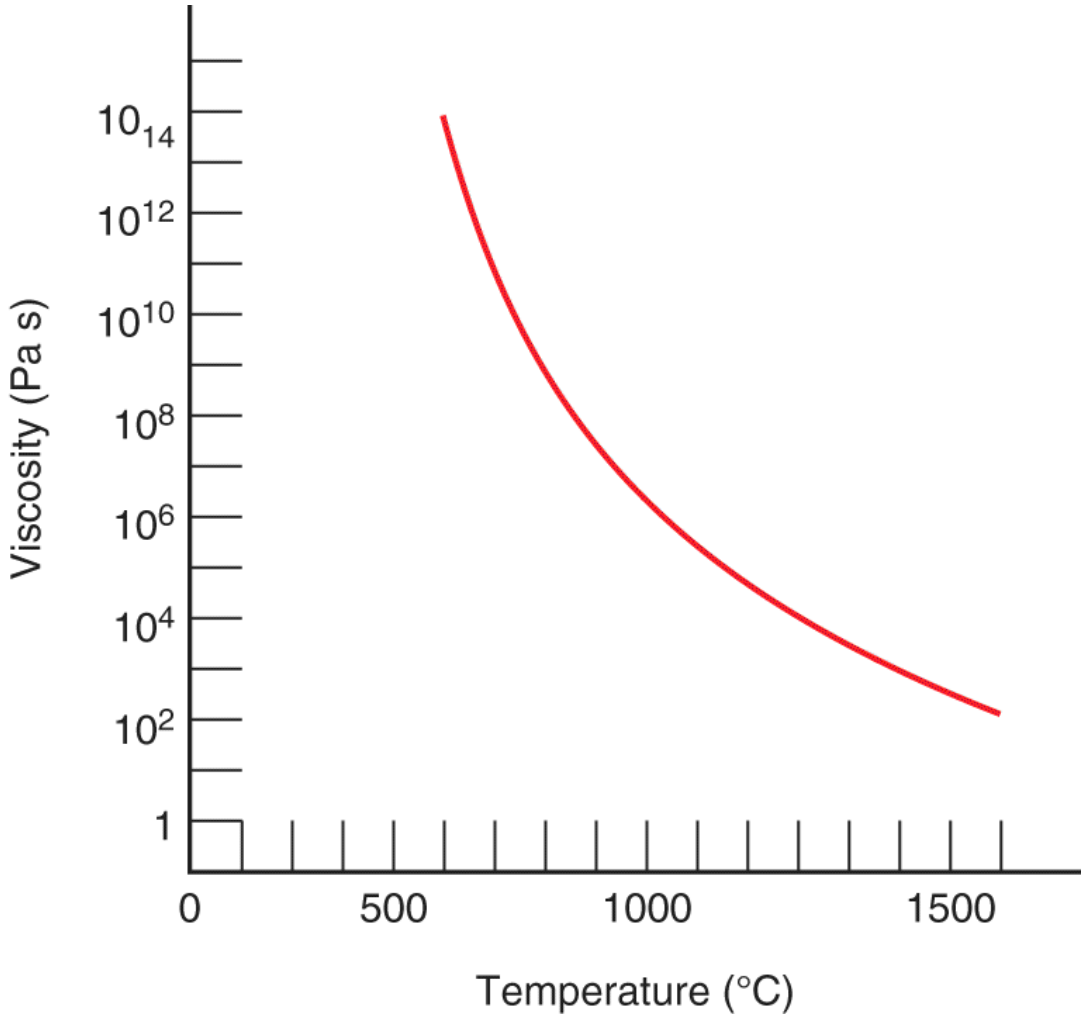
4.25 Calculate the viscosity at 940° and 1400 °C of a high silica glass for which, η_0 is 3.5×10^{-5} Pa s, and the activation energy is 382 kJ mol^{-1}.

4.26 The viscosity parameters for a clear float glass are: softening point, 720 °C, annealing point, 535 °C, strain point, 504 °C. Estimate the activation energy for the viscosity.

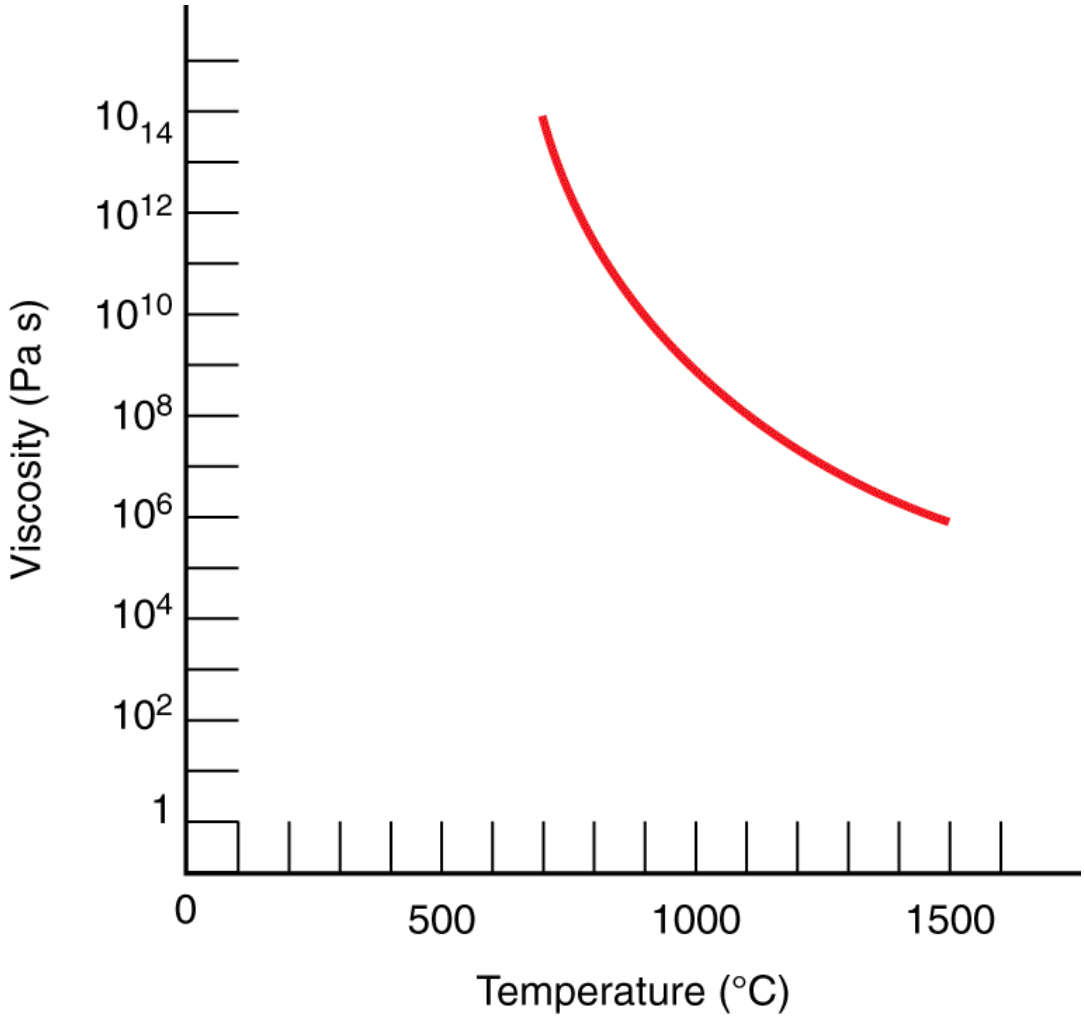
4.27 The viscosity parameters for a glass are: softening point, 677 °C, annealing point, 532 °C, strain point, 493 °C. Estimate the activation energy for the viscosity.

4.28 The viscosity parameters for a borosilicate glass are: softening point, 794 °C, annealing point, 574 °C, strain point, 530 °C. Estimate the activation energy for viscosity.

4.29 The viscosity of a borosilicate glass is shown in the figure. Estimate the activation energy of the viscosity and comment on the form of the Arrhenius plot.



4.30 The viscosity of a high silica glass is shown in the figure. Estimate the activation energy of the viscosity and comment on the form of the Arrhenius plot.



4.31 Write the reaction equation for the reaction of two monomer molecules of styrene to produce a dimer.

a. What weight of monomer is needed to produce 100 kg of polymer?

b. How many monomer molecules is this?

c. The number average molar mass of the polymer is 250 000 g mol^{-1}. What is the degree of polymerisation?

4.32 Write the reaction equation for the reaction of two monomer molecules of methyl methacrylate to produce a dimer.

a. What weight of monomer is needed to produce 100 kg of polymer?

b. How many monomer molecules is this?

c. The number average molar mass of the polymer is 200 000 g mol^{-1}. What is the degree of polymerisation?

4.33 Nitrile rubber is a copolymer of butadiene (CH_2=CH–CH=CH_2), and acrylonitrile (propene nitrile) (CH_2=CH–CN).

a. Write the reaction equation between these two molecules to produce a dimer.

b. What masses of the reactants are needed to give 100 kg of polymer?

c. How many molecules of each reactant is this?

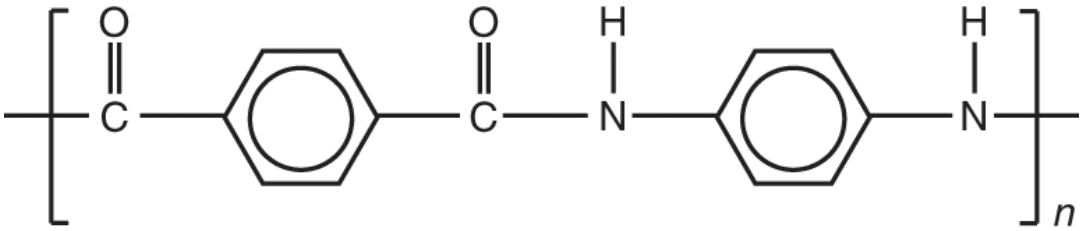
4.34 34 PET is a polymer of terephthalic acid and ethylene glycol (see Figure 4.26).

a. Write the hypothetical reaction equation between these two molecules to produce a dimer.

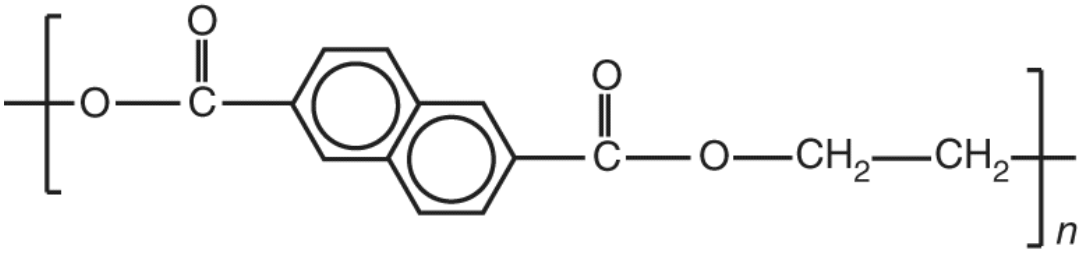
b. What masses of the reactants are needed to give 100 kg of polymer?

c. How many molecules of each reactant is this?

4.35 The structural formula of the aramid polymer Kevlar, used in bulletproof vests, is drawn below. Write the formulae of the two monomers used.



4.36 The structural formula of the polycarbonate PEN [poly(ethylene naphthalate)] used in recyclable jars and bottles, is drawn in the following. Write the formulae of the two monomers used.



4.37 Compute:

The number-average molar mass

The degree of polymerisation for polypropylene from the following data.

Molar mass range (g mol^{-1})	Mean molar mass in range (g mol^{-1})	Fraction of molecules in range

5 000–10 000	7 500	0.01
10 000–15 000	12 500	0.09
15 000–20 000	17 500	0.17
20 000–25 000	22 500	0.18
25 000–30 000	27 500	0.20
30 000–35 000	32 500	0.17
35 000–40 000	37 500	0.09
40 000–45 000	42 500	0.06
45 000–50 000	47 500	0.03

4.38 Compute

The weight-average molar mass

The degree of polymerisation for polypropylene from the following data.

Molar mass range ($g\ mol^{-1}$)	Mean molar mass in range ($g\ mol^{-1}$)	Weight fraction of molecules in range
5 000–10 000	7 500	0.01
10 000–15 000	12 500	0.07
15 000–20 000	17 500	0.16
20 000–25 000	22 500	0.20
25 000–30 000	27 500	0.23
30 000–35 000	32 500	0.18
35 000–40 000	37 500	0.08
40 000–45 000	42 500	0.05
45 000–50 000	47 500	0.02

4.39 The density of polyethylene crystals is 998 $kg\ m^{-3}$, and the unit cell has dimensions: a = 0.741 nm, b = 0.494 nm, c = 0.255 nm.

a. How many CH_2 units are there in a unit cell?

b. How many monomer (CH_2–CH_2) units are there?

4.40 The density of amorphous polythene is approximately 810 $kg\ m^{-3}$. Estimate the crystallinity of low-density polyethylene, density 920 $kg\ m^{-3}$, medium density polyethylene, density 933 $kg\ m^{-3}$, and high-density polyethylene, density 950 $kg\ m^{-3}$. (The density of crystalline polyethylene is given in Question 4.39.)

4.41 41 Calculate the heats of hydration after three days and one year for the Portland cement compositions (wt%):

a. 50% C_3S; 25% C_2S; 12% C_3A; 8% C_4AF

b. 45% C_3S; 30% C_2S; 7% C_3A; 12 C_4AF

c. 60% C_3S; 15% C_2S; 10% C_3A; 8% C_4AF

Notes

1 The figures in this section chapter are descriptive. In real structures the silicon-oxygen tetrahedra are often slightly distorted, and crystal structure studies should be consulted when precise detail is required.

2 The cgs unit for viscosity, Poise (P) is commonly found in the literature. The conversion of SI Pa s is 1 Pa s = 10 P.

Part II
Reactions and Transformations

5
Diffusion and Ionic Conductivity

- What is steady state diffusion?
- What are diffusion coefficients?
- What is a correlation factor?

In this chapter, attention is focussed upon atomic and ionic transport through the solid matrix rather than gas percolation through porous materials. Diffusion is mostly driven by a concentration gradient and is quantified by measuring the concentration of the diffusing atoms at different distances from the release point after a given time has elapsed. The results are expressed in terms of a *diffusion coefficient*, *D*. Ionic conductivity is, in principle, similar to diffusion, except that an electric field gradient is imposed upon the material that influences movement of the charged species. The relationship between ionic conductivity and diffusion is discussed later in this chapter.

Movement through the body of a solid is called *volume*, *lattice*, or *bulk* diffusion. In amorphous or glassy solids and in cubic crystals, diffusion is the same in all directions and the material is described as *isotropic*. In porous solids such as zeolites, metal-organic, and covalent organic framework structures, diffusion is *anisotropic*, with transport along pores quicker than the diffusion perpendicular to them, to such an extent that a factor of as much as 10^4 can separate the two quantities. Moreover, in polycrystalline aggregates such as a ceramic body, atoms can diffuse along surfaces and grain boundaries or along dislocations. As the regular crystal geometry is disrupted in these regions, atom movement is often much faster than for volume diffusion. Diffusion by way of these latter pathways is collectively referred to as *short-circuit* diffusion.

5.1 Self-Diffusion and Tracer Diffusion

When atoms in a pure crystal diffuse under no concentration gradient or another driving force, the process is called *self-diffusion*. In such a case, the atomic movements are random, with motion in one direction just as likely as another (Figure 5.1). The relevant diffusion coefficient is called the *self-diffusion coefficient* and is given the symbol D_S.

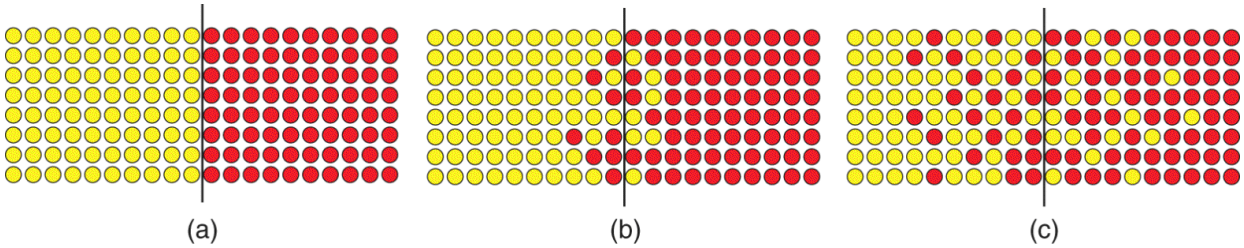


Figure 5.1 Self-diffusion of atoms in a solid: (a) initial state; (b) after heating for a shorter time; (c) after heating for a longer time.

It is by no means easy to measure the self-diffusion coefficient because it is not possible to keep track of the movements of one atom in a crystal composed of many identical atoms. However, it is possible to measure something that is a very good approximation to the self-diffusion coefficient, if some of the atoms can be uniquely labelled and their movement tracked. In this case, the diffusion coefficient that is measured is called the *tracer diffusion coefficient*, written D^*. The experiment can be repeated with impurity tracer atoms, *A*, to yield the *tracer impurity diffusion coefficient*, D^*_A.

Tracer diffusion coefficients are often measured by tracking the movement of radioactive atoms. For example, to determine the tracer diffusion coefficient of magnesium atoms in MgO, a thin layer of radioactive Mg, comprising the *tracer atoms*, is evaporated onto the surface of a carefully polished single crystal of MgO. This layer is oxidised to MgO by exposing the surface to oxygen, after which another carefully polished single-crystal slice of MgO is placed on top to form a *diffusion couple* (Figure 5.2). The crystal sandwich is heated for a known time and temperature. The whole slab is then carefully sliced parallel to the original interface containing the radioactive MgO layer and the radioactivity of each slice, which is a measure of the

concentration of radioactive Mg in each section, is determined. A graph of concentration of the radioactive component is then plotted against the distance from the interface to give a *diffusion profile*, or *concentration profile* (Figure 5.3). Note that the concentration of the tracer at any point changes over time.

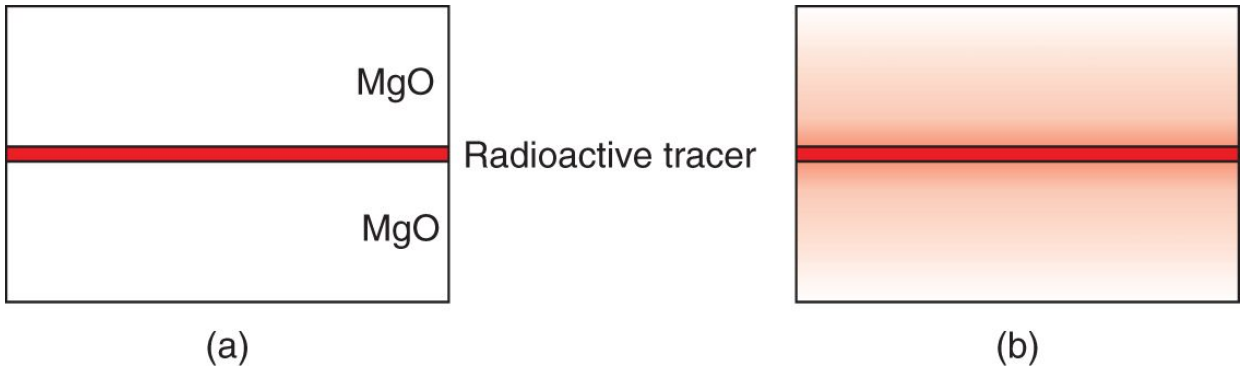


Figure 5.2 Diffusion couple formed by two MgO crystals separated by a film of MgO containing radioactive Mg; (a) initially; (b) after heating.

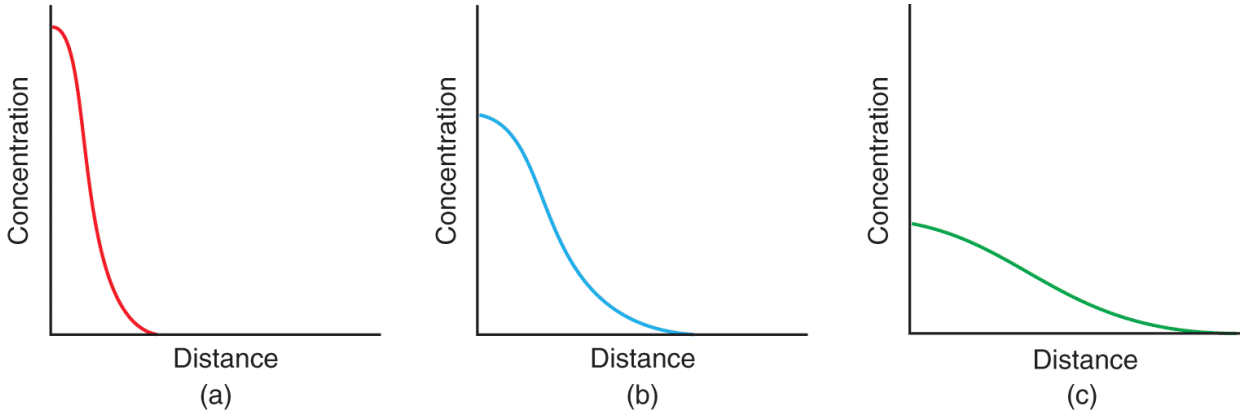


Figure 5.3 Diffusion profiles: (a), (b), (c), gradually increasing heating times.

The tracer diffusion coefficient is obtained from such profiles via the *diffusion equation*, also called *Fick's second law*:

$$dc_x/dt = D^*(d^2c_x/dx^2)$$

where c_x is the concentration of the diffusing radioactive ions at a distance x from the original interface after time t has elapsed and D^* is the tracer diffusion coefficient. When D^* is a constant the equation

can be solved analytically to give an expression for c in terms of x. For the experimental arrangement in Figure 5.2, the solution is:

$$c_x = [c_0/2(\pi D^* t)^{1/2}] \exp [-x^2/4D^* t]$$

where c_0 is the initial concentration on the surface. A value for the tracer diffusion coefficient is obtained by taking logarithms of both sides this equation, to give:

$$\ln c_x = \ln [c_0/2(\pi D^* t)^{1/2}] - [x^2/4D^* t]$$

This has the form:

$$\ln c_x = \text{constant} - [x^2/4D^* t]$$

and a plot of ln c_x vs. x^2 will have a gradient of $[-1/4D^*t]$ (Figure 5.4). A measurement of the gradient gives a value for the tracer diffusion coefficient at the temperature at which the diffusion couple was heated. To obtain the diffusion coefficient over a variety of temperatures, the experiments must be repeated.

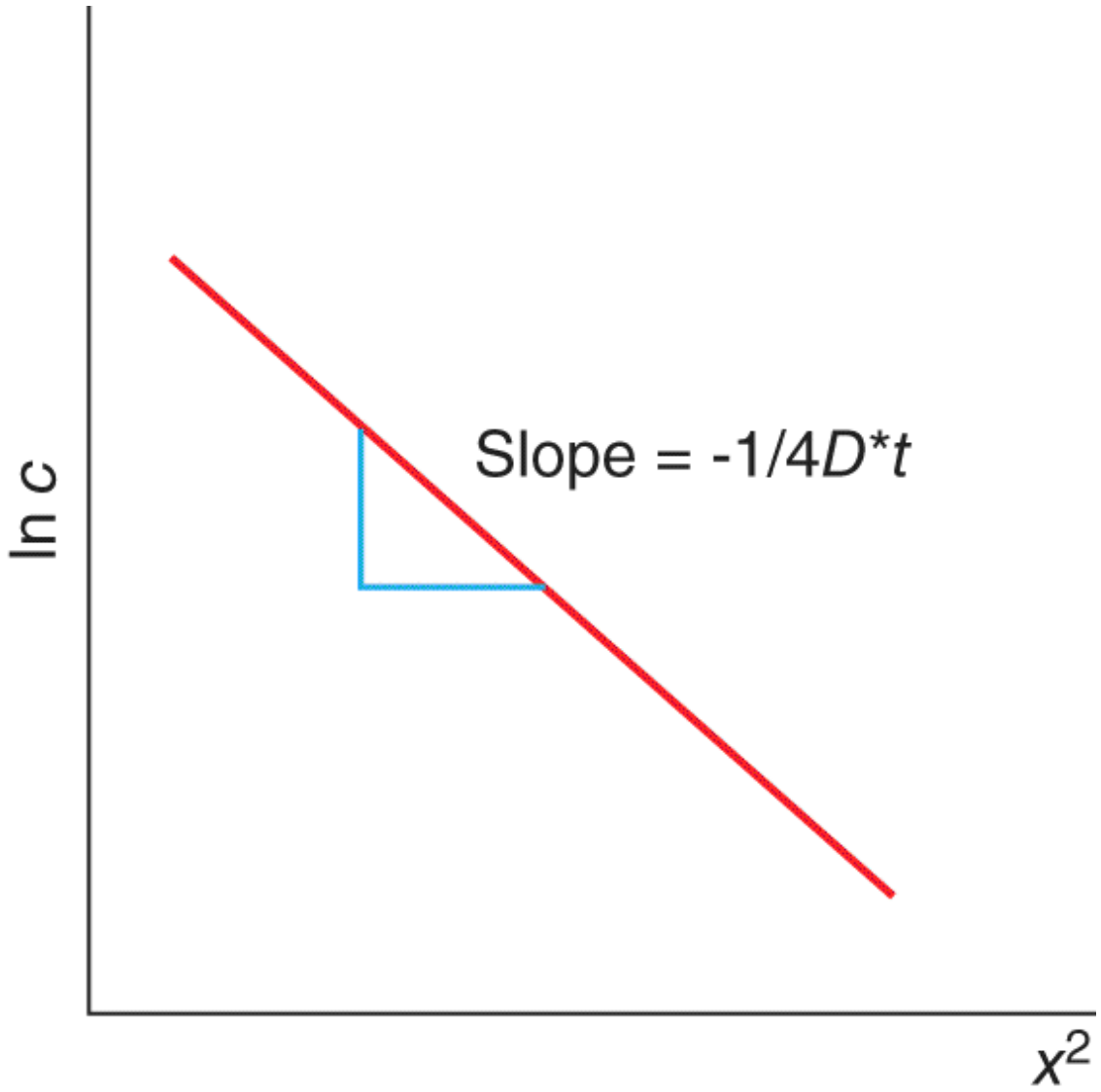


Figure 5.4 The straight-line graph of ln c vs. x^2 from a diffusion experiment.

In this experiment, there is a concentration gradient, because the concentration of the radioactive isotopes in the coating will be different to the concentration of radioactive isotopes, if any, in the original crystal pieces. However, if the layer of tracer atoms is very thin, the concentration gradient will be small and will rapidly become smaller as diffusion takes place and in these circumstances D^*, the tracer diffusion coefficient will be very similar to the self-diffusion coefficient, D_S.

5.2 Non-steady-state and Steady-State Diffusion

The normal state of affairs during a diffusion experiment is one in which the concentration at any point in the solid changes over time, as in the example described earlier. This situation is called *non-steady-state diffusion*, and diffusion coefficients are found by solving the diffusion equation. Provided the diffusion coefficient, D, is not dependent upon composition and position, analytical solutions can be found (Figure 5.5 and Table 5.1).

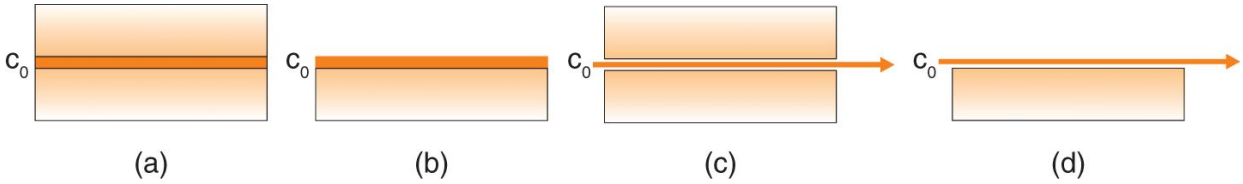


Figure 5.5 Geometries for non-steady-state diffusion: (a), (b), the concentration of diffusant is unreplenished; (c), (d) the concentration of diffusant is maintained at a constant value c_0.

Table 5.1 Solutions of the diffusion equation.

Experimental arrangement	**Solution**[a]
Initial concentration, c_0, unreplenished	
Thin film planar sandwich	$c_x = [c_0/2(\pi Dt)^{1/2}] \exp [-x^2/4Dt]$
Open planar thin film	$c_x = [c_0/(\pi Dt)^{1/2}] \exp [-x^2/4Dt]$
Small spherical precipitate	$c_r = [c_0/8(\pi Dt)^{3/2}] \exp [-r^2/4Dt]$
Initial concentration, c_0, maintained constant	
Open plate	$(c_x - c_0)/(c_s - c_0) = 1 - \text{erf}\,(x/2(Dt)^{1/2})$
Sandwich plate	$(c_x - c_0)/(c_s - c_0) = \frac{1}{2} [1 - \text{erf}\,((x/2(Dt)^{1/2})]$

[a] Derived assuming diffusion occurs in infinite solids with constant D.

In some experimental arrangements (Figure 5.5a,b), the initial concentration of the tracer is fixed and the amount remaining as the

diffusion progresses will diminish over the course of the experiment. In other cases, the initial concentration at the surface is maintained as a constant throughout the experiment, as when, for example, molecules diffuse into a solid from a flowing gas or liquid and the supply is constantly replenished (Figure 5.5c,d). The solution for diffusion in these cases involve the *error function*, in the form $\text{erf}\,(x/2\,(Dt)^{1/2})$. The error function cannot generally be evaluated analytically, but numerical values are readily calculated via many computer routines. An abbreviated list is given in Table 5.2.

Table 5.2 Values of the error function.[a]

z	Erf (z)	z	Erf (z)	Z	Erf (z)	Z	Erf (z)
0	0	0.40	0.4284	0.85	0.7707	1.60	0.9763
0.025	0.0282	0.45	0.4755	0.90	0.7970	1.70	0.9838
0.05	0.0564	0.50	0.5205	0.95	0.8209	1.80	0.9891
0.10	0.1125	0.55	0.5633	1.00	0.8427	1.90	0.9928
0.15	0.1680	0.60	0.6039	1.10	0.8802	2.00	0.9953
0.20	0.2227	0.65	0.6420	1.20	0.9103	2.20	0.9981
0.25	0.2763	0.70	0.6778	1.30	0.9340	2.40	0.9993
0.30	0.3286	0.75	0.7112	1.40	0.9523	2.60	0.9998
0.35	0.3794	0.80	0.7421	1.50	0.9661	2.80	0.9999

[a] erf (z) is equal to 0 when z is equal to 0 and equal to 1 when z is equal to 2.8. erf ($-z$) = [$-$erf (z)].

Steady-state diffusion differs from non-steady state diffusion in that the concentration of the diffusing atoms at any point, x, and hence the concentration gradient at x, in the solid, remains constant (Figure 5.6). Steady-state diffusion can occur when a gas permeates through a thin metal foil or a framework structure such as a zeolite, MOF or COF. Hydrogen gas, for example, can be purified by allowing it to diffuse through a palladium thimble.

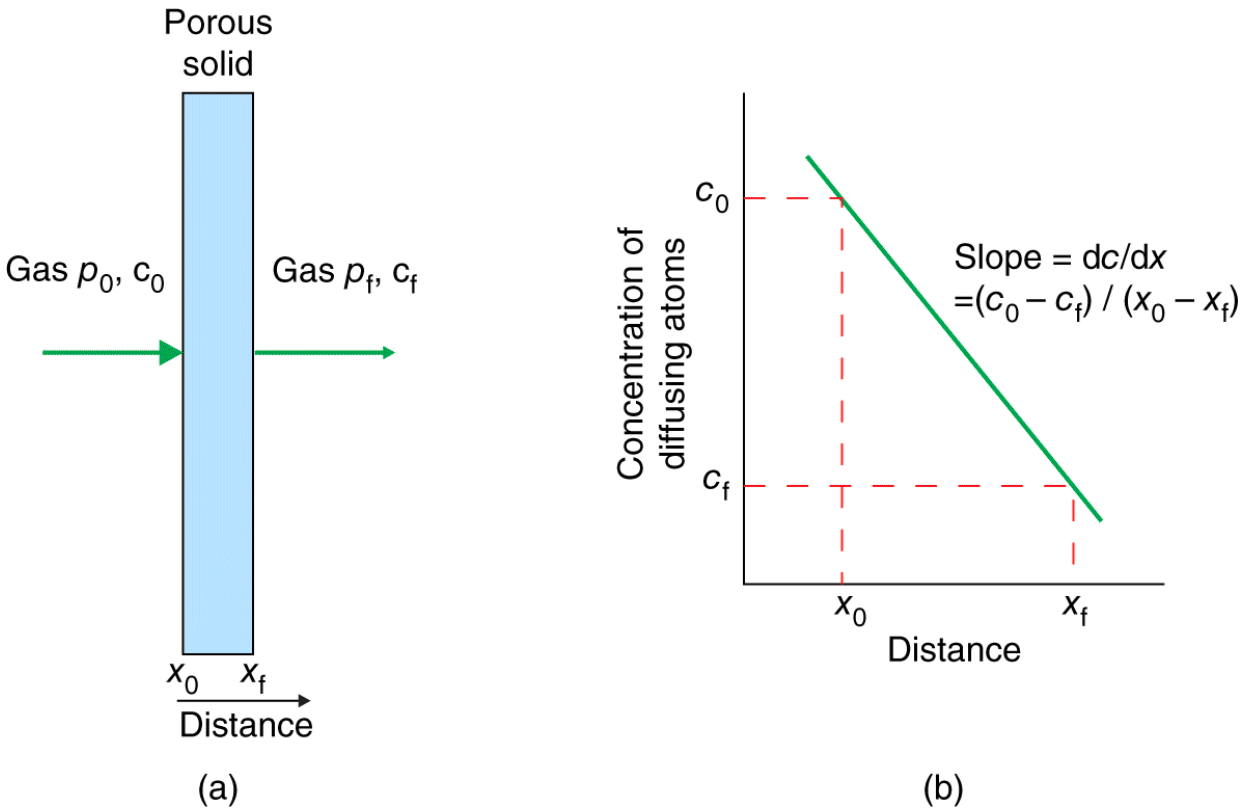


Figure 5.6 Steady-state diffusion: (a) gas diffusing through a porous solid; (b) the resultant diffusion profile.

Under steady-state conditions, the diffusion coefficient is obtained using *Fick's first law*. This is written as:

$$J_i = -D_i \, (dc_i/dx)$$

where D_i is the diffusion coefficient of atom type i, c_i is the concentration of these atoms, and x is the position in the solid. J_i is called the *flux* of atoms of type i, that is, the net flow of these atoms through the solid. It is measured in atoms (or a related unit such as grams or moles), $m^{-2}\ s^{-1}$. When the steady state has been reached, the diffusion coefficient across the foil, D, will be given by:

$$D = J\, l/(c_1 - c_2)$$

where the concentrations on each side of the foil are c_1 and c_2 and the foil thickness is l.

5.3 Temperature Variation of Diffusion Coefficient

Diffusion coefficients are usually found to vary considerably with temperature. This can often be expressed in terms of the *Arrhenius equation*:

$$D = D_0 \exp(-E/RT)$$

where R is the gas constant, T is the temperature (K), and D_0 is a constant term referred to as the *pre-exponential factor* or, less often, as the *frequency factor*. The term E is called the *activation energy* of diffusion. Taking logarithms of both sides of this equation gives:

$$\ln D = \ln D_0 - E/RT$$

The activation energy can be determined from the gradient of a plot of $\ln D$ vs. $1/T$ (Figure 5.7). Such graphs are known as *Arrhenius plots*. Diffusion coefficients are usually expressed in terms of D_0 and E (Table 5.3).

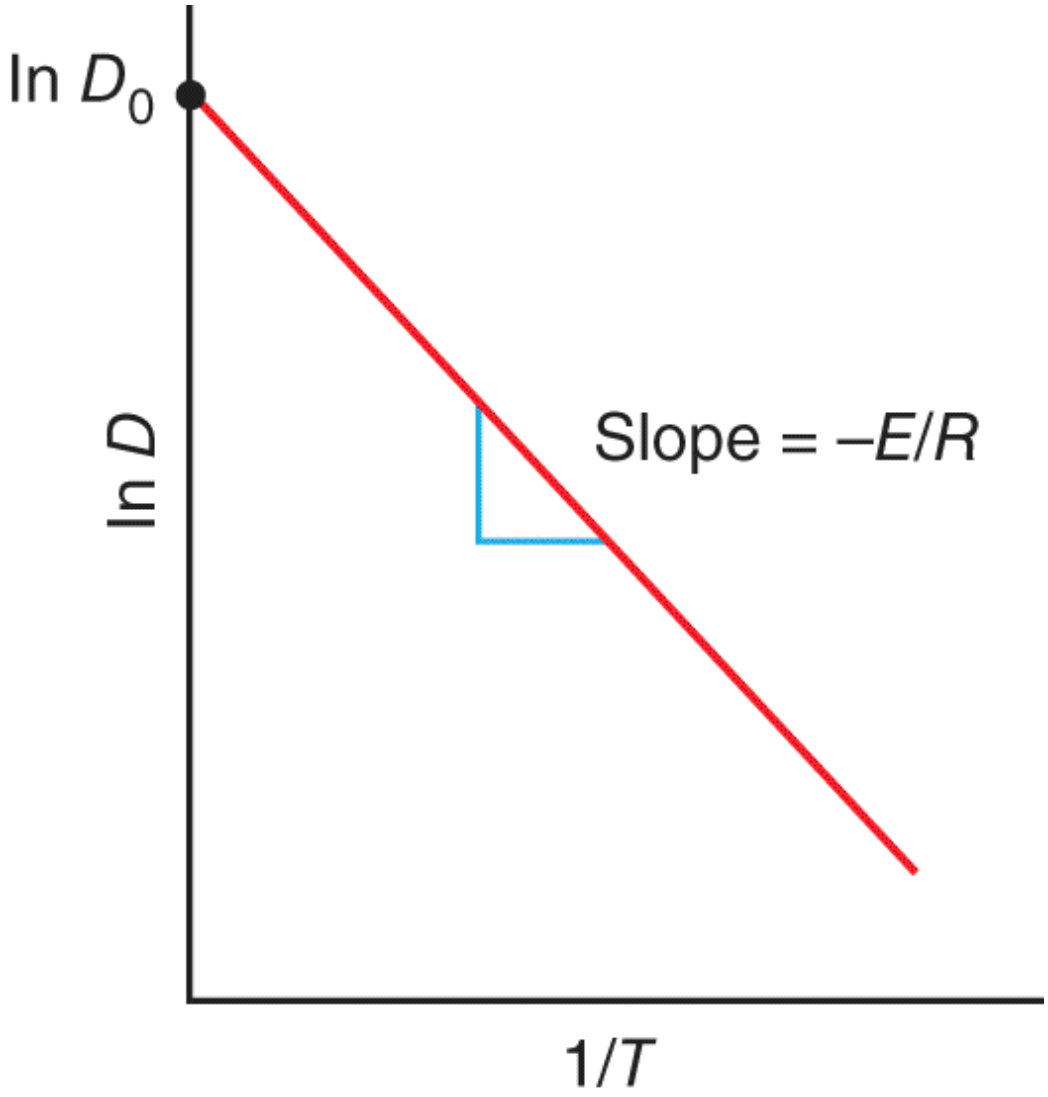


Figure 5.7 Arrhenius plot of ln D vs. $1/T$.

Table 5.3 Some representative values for self-diffusion coefficients.[a]

Atom	Matrix[b]	D_0 ($m^2\ s^{-1}$)[c]	E ($kJ\ mol^{-1}$)
Metals and semiconductors			
Cu	Cu (fcc)	2.0×10^{-5}	200
γ-Fe	Fe (fcc)	2.0×10^{-5}	270
α-Fe	Fe (bcc)	2.0×10^{-4}	240
Na	Na (bcc)	2.5×10^{-5}	45
Si	Si (diamond)	0.5×10^{-1}	455
Ge	Ge (diamond)	9.5×10^{-4}	290
Compounds			
Na^+	NaCl (NaCl)	8.5×10^{-8}	190
Cl^-	NaCl (NaCl)	0.5×10^{-4}	245
K^+	KCl (NaCl)	0.5×10^{-4}	255
Cl^-	KCl (NaCl)	1.5×10^{-6}	230
Mg^{2+}	MgO (NaCl)	2.5×10^{-9}	330
O^{2-}	MgO (NaCl)	4.5×10^{-13}	345
Ni^{2+}	NiO (NaCl)	5.0×10^{-10}	255
O^{2-}	NiO (NaCl)	6.0×10^{-12}	240
Pb^{2+}	PbS (NaCl)	8.5×10^{-13}	145
S^{2-}	PbS (NaCl)	7.0×10^{-13}	135
Ga	GaAs (sphalerite)	2.0×10^{-10}	400
As	GaAs (sphalerite)	7.0×10^{-5}	310
Zn^{2+}	ZnS (sphalerite)	3.0×10^{-12}	145
S^{2-}	ZnS (sphalerite)	2.0×10^{-4}	305
Cd^{2+}	CdS (sphalerite)	3.5×10^{-8}	195
S^{2-}	CdS (sphalerite)	1.5×10^{-10}	200

Atom	Matrix[b]	D_0 (m^2 s^{-1})[c]	E (kJ mol^{-1})
Glasses			
Na	Soda-lime	1.0×10^{-6}	85
Si	Silica	3.0×10^{-2}	580
Na	Soda	1.0×10^{-6}	80

[a] Literature values for self-diffusion coefficients vary widely, indicating the difficulty of making reliable measurements. The values here are meant to be representative only.

[b] The structure type is shown in parenthesis.

[c] The values of diffusion coefficients in the literature are mostly given in cm^2 s^{-1} To convert the values given here to cm^2 s^{-1}, multiply by 10^4.

5.4 The Effect of Impurities

Although Arrhenius plots for the majority of materials resemble Figure 5.7, plots obtained from very pure materials often consist of two straight-line parts with differing slopes (Figure 5.8). The region corresponding to diffusion at lower temperatures has lower activation energy than the high temperature region. The point where the two straight lines intersect is called a *knee*. If a number of different crystals of the same compound are studied, it is found that the position of the knee varies from one crystal to another, and depends upon the impurity content. The part of the plot sensitive to impurity content is called the *impurity* or *extrinsic* region. The high temperature part of the plot is unaffected by the impurities present and is called the *intrinsic* region.

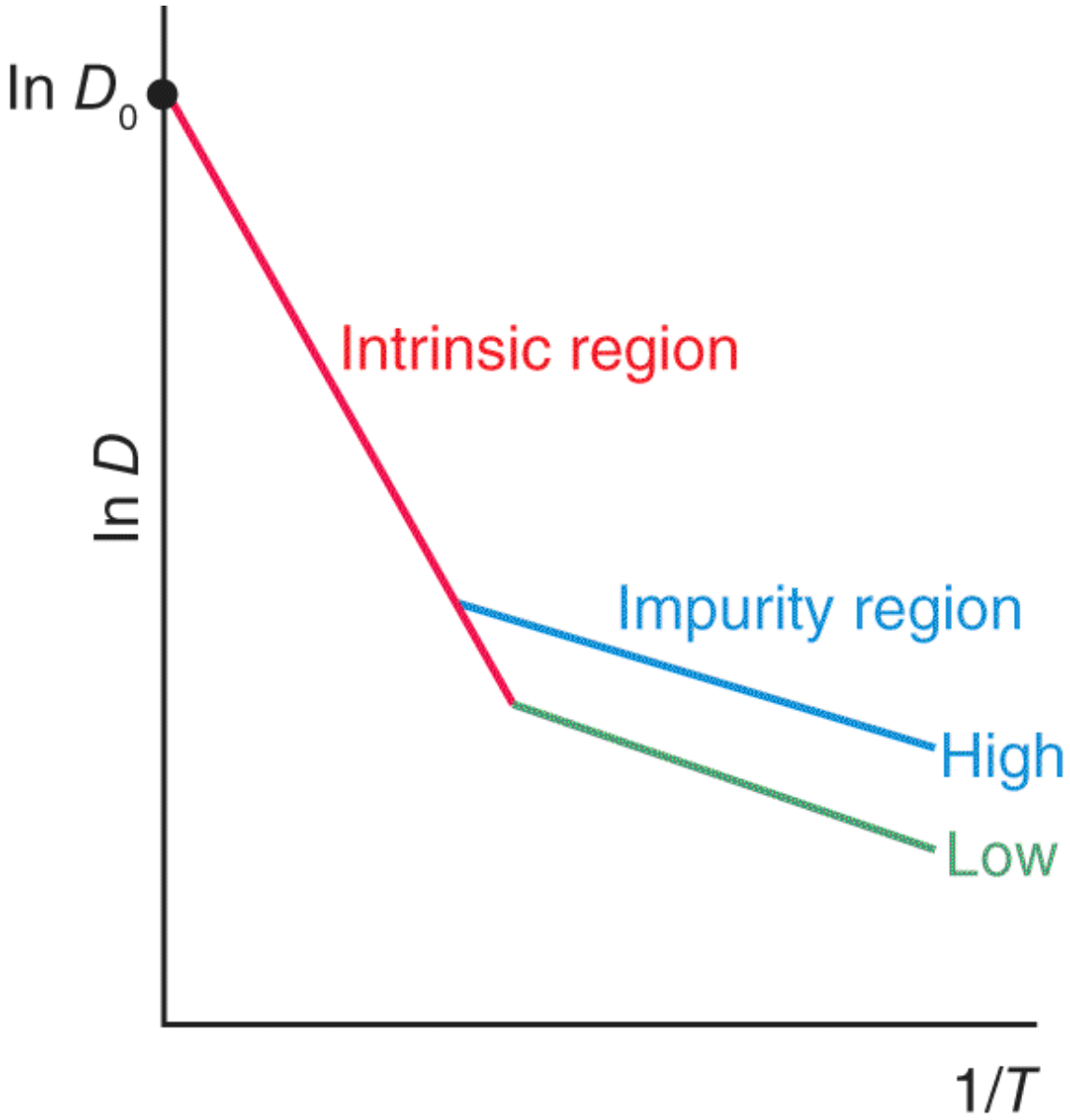


Figure 5.8 A form of the Arrhenius equation found for crystals with lower (low) and higher (high) concentrations of impurities.

5.5 Random Walk Diffusion

The simplest model for the diffusion of an atom is to assume that it moves by a series of random jumps, due to the fact that it is being continually jostled by thermal energy. Diffusion in terms of this model was formulated by Einstein in 1905 to explain Brownian motion, the apparent agitation of minute particles suspended in a liquid at normal temperatures. The path followed is called a *random* (or *drunkard's*) *walk* (Figure 5.9). It is, at first sight, surprising that any diffusion will take place under these circumstances, because, intuitively, the distance that an atom will move via random jumps in one direction would be balanced by jumps in the opposite direction, so that the overall displacement would be expected to average out to zero. Nevertheless, diffusion does occur. For example, Figure 5.10 shows the distance moved by a particle over each step of a random walk of 100 steps with no pause between each jump.

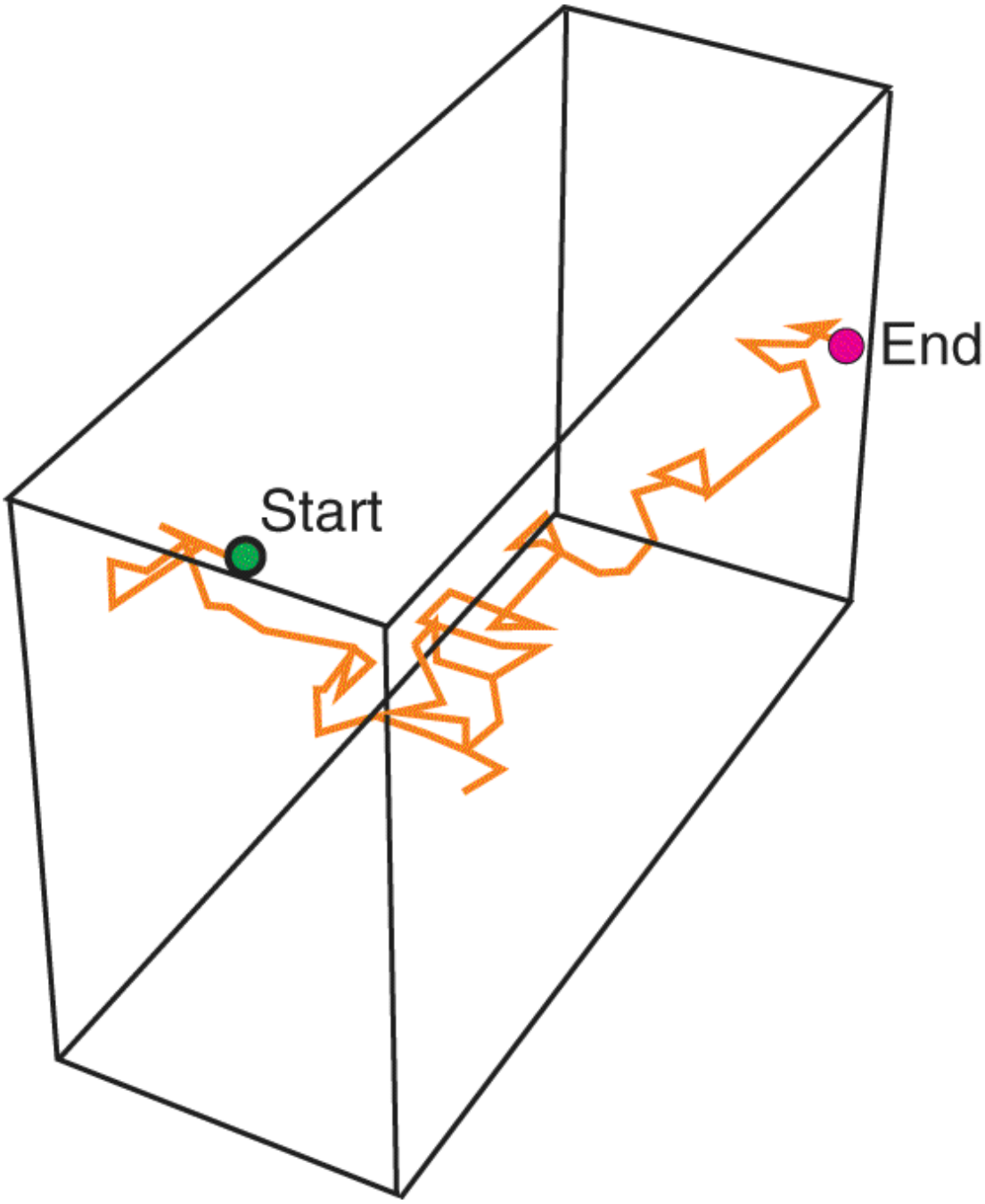


Figure 5.9 Computed random walk in three-dimensions.

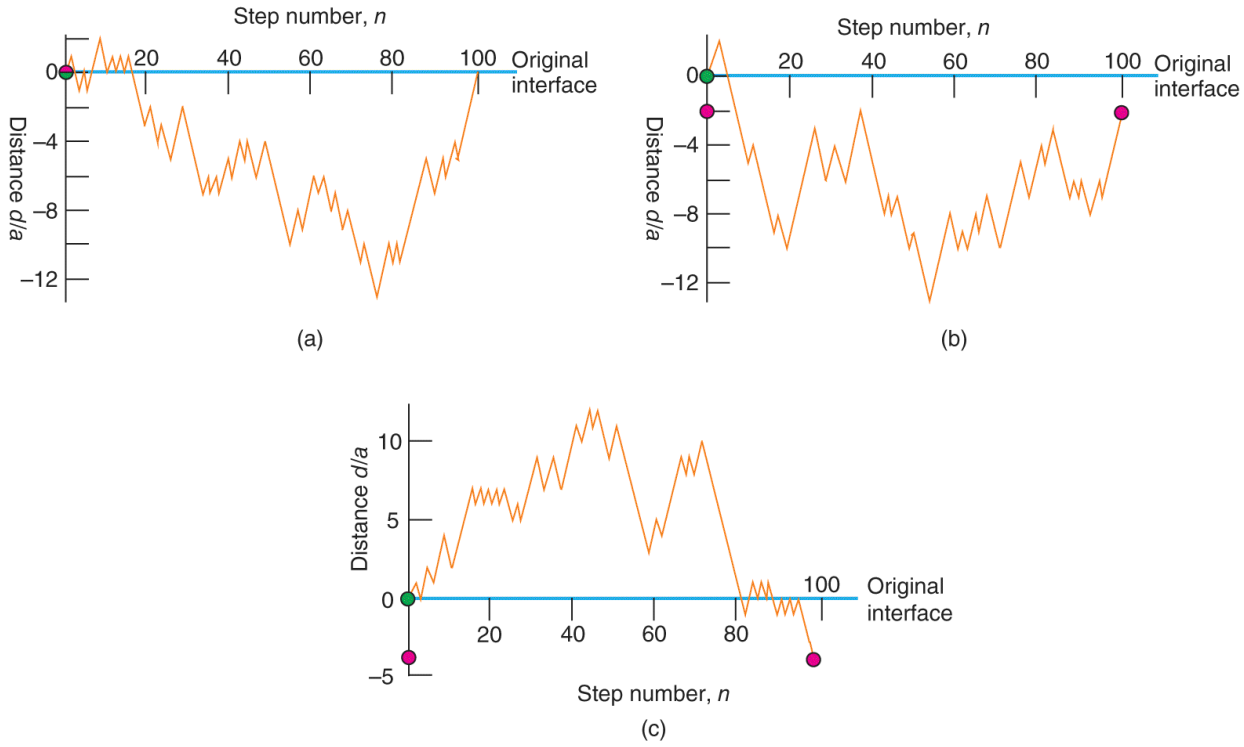


Figure 5.10 One-dimensional random walks of 100 steps of length a. The total distance moved is: (a) 0; (b) $-2a$; (c) $-4a$.

Suppose a planar layer of tracer atoms is the starting point and suppose that an atom diffuses from the interface by a random walk in a direction perpendicular to the interface, in what is effectively one-dimensional diffusion, again, with no pauses between each jump. The net displacement of a diffusing atom after n jumps will be the algebraic sum of the individual jumps. If x_i is the distance moved along the x-axis in the ith jump, the distance moved after a total of n jumps, x, will simply be the sum of all the individual steps, i.e.:

$$x = x_1 + x_2 + x_3 \ldots \ldots \ldots = \sum x_i$$

If the sites are separated by a distance a, each individual value of x_i can be $+a$ or $-a$. As the jumps take place with an equal probability in both directions, after n jumps the total displacement may have any value between zero and na. The probability of arriving at any site ma from the origin is approximated by a Gaussian (normal) distribution (Figure 5.11).

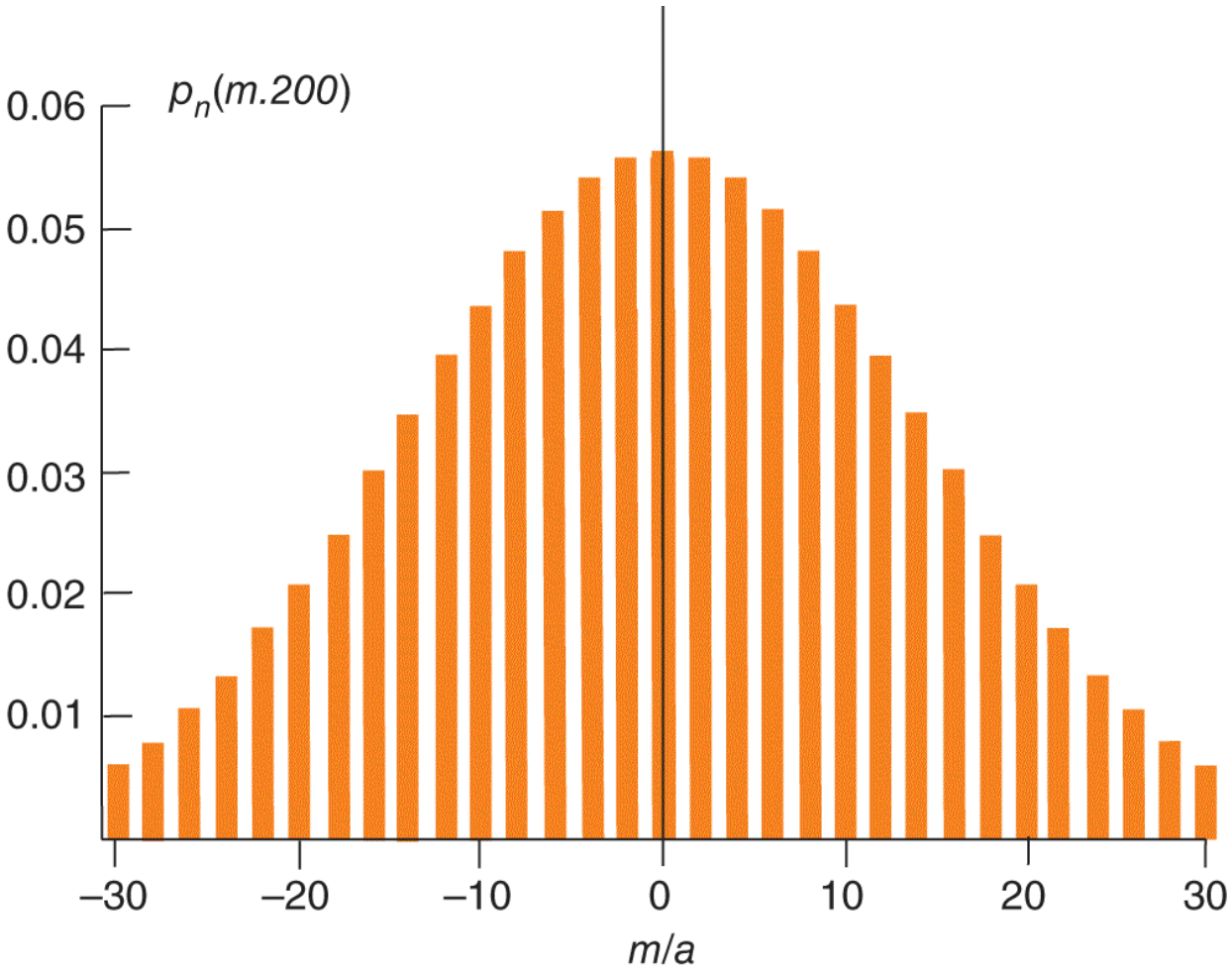


Figure 5.11 The probability of a diffusing atom arriving at a point *ma* after 200 random steps.

To relate this to the observed diffusion coefficient of an atom, the jump distances are squared so as to eliminate negative quantities.

$$
\begin{aligned}
x^2 &= (x_1 + x_2 + x_3 \ldots x_n)\,(x_1 + x_2 + x_3 \ldots x_n) \\
&= (x_1x_1 + x_1x_2 + x_1x_3 \ldots x_1x_n) \\
&\quad + (x_2x_1 + x_2x_2 + x_2x_3 \ldots x_2x_n) \\
&\quad + \\
&\quad + x_nx_1 + x_nx_2 + x_nx_3 \ldots x_nx_n)
\end{aligned}
$$

This can be written:

$$
\begin{aligned}
x^2 &= \sum {x_i}^2 + 2 \sum x_i x_{i+1} \\
&\quad + 2 \sum x_i x_{i+2} + \ldots \ldots \\
&= \sum {x_i}^2 + 2 \sum \sum x_i x_{i+j}
\end{aligned}
$$

In the limit of a large number of jumps, knowing that each jump may be either positive or negative, the double sum terms average to zero. The equation therefore reduces to the manageable form:

$$\langle x^2 \rangle = \sum {x_i}^2$$

where $\langle x^2 \rangle$ is the *mean square displacement*. As each jump, x_i, can be equal to $+a$ or $-a$,

$$
\begin{aligned}
\langle x^2 \rangle &= {x_1}^2 + {x_2}^2 + {x_3}^2 \cdots + {x_n}^2 \\
&= a^2 + a^2 + a^2 \cdots + a^2 \\
&= na^2
\end{aligned}
$$

The *frequency* with which an atom jumps from one site to another along the x direction is defined as Γ jumps per second, so that the total number of jumps, n, will be given by Γ multiplied by the time, t, over which the diffusion experiment has lasted, that is:

$$n = \Gamma t$$

Hence:

$$\langle x^2 \rangle = n\, a^2 = \Gamma t\, a^2$$

To relate this equation to a diffusion coefficient for the random walk, it must agree with Fick's first law:

$$D_r = \tfrac{1}{2}\, \Gamma\, a^2$$

so that:

$$\langle x^2 \rangle = 2 D_r\, t$$

where D_r is the random-walk diffusion coefficient. This relationship is known as the *Einstein* (or *Einstein–Smoluchowski*) diffusion equation. The diffusion constant D_r is often equivalent to the self-diffusion coefficient, D_S, or the tracer diffusion coefficient, D^*. For one-dimensional diffusion, the probability that any particular atom will be found in the region between the starting point of the diffusion and a distance of $\pm\sqrt{\langle x^2 \rangle} = \pm\sqrt{(2\, D_r\, t)}$ on either side of it is approximately 68%.

Random walk statistics can be used to give an approximate estimate of the extent to which diffusion is significant. This distance, the *penetration depth*, x_p, is the distance at which an appreciable change in the concentration of the diffusing species be said to have occurred after a diffusion time t. This is chosen as equivalent to root mean square displacement of the diffusing atoms, $\pm\sqrt{\langle x^2 \rangle}$, so that the penetration depth can be expressed as:

$$x_p \approx \sqrt{(2\, D^* t)} \qquad (5.1)$$

where D^* is the tracer diffusion coefficient and t is the time of the diffusion process. This choice is somewhat arbitrary, and sometimes the penetration depth is defined as:

$$x_p \approx 2\sqrt{(2\, D^* t)}$$

or as the diffusion length or diffusion distance, L, where:

$$L = \sqrt{(D^* t)}$$

The factor of 2 in Eq. (5.1) arises from the one-dimensional nature of the random walk, and hence is a result of the geometry of the diffusion process. In the case of random-walk diffusion on a two-dimensional surface:

$$\langle x^2 \rangle = 4D_r t$$

while for random-walk diffusion in a three-dimensional crystal:

$$\langle x^2 \rangle = 6D_r t$$

The random walk mechanism is applicable to diffusion of particles in a liquid or molecules in a gas, but in a crystal the vast majority of the possible atom positions are already occupied and random jumps in any direction are clearly prohibited and it will not, in general, be possible for a diffusing atom to make a series of jumps in any direction without any pauses between each step. In addition, ions diffusing in an electric field will show a preferred direction of movement. An accurate model of atomic diffusion in a crystal should ideally take this factor into account.

A one-dimensional random walk including a preferred direction (a bias) and a possible waiting time between steps is modelled by supposing that the diffusing atom moves a distance $\pm a$ with a probability p (+) or q (−) or stays at the same place (waits) with a probability $(1-p-q)$:

$$p\,(x, t) = [1/\sqrt{(4\pi Dt)}]\ \exp[-(x - vt)^2/4Dt]$$

This is now *drift diffusion* or *advection diffusion*. The drift shifts the centre of the Gaussian distribution from $x = 0$ to $x = vt$. The drift velocity is given by:

$$v = a\,(p - q)/\tau$$

where a is the step length. Note, if $p = q$, irrespective of staying time, the drift velocity $v = 0$.

The diffusion coefficient is given by:

$$D = (p + q)\,(a^2/2\tau)$$

The mean location becomes:

$$\langle x^2 \rangle = vt$$

and the mean squared displacement is:

$$\langle x^2 \rangle = u^2t^2 + 2Dt$$

Note that the mean squared displacement drift is proportional to t^2 not t.

5.6 Diffusion in Solids

In a crystal, because the vast majority of atom positions are occupied, and random jumps in any direction are clearly prohibited, diffusion must then also be considered in terms of defects present, which will allow diffusion to occur. Diffusion along dislocations and grain boundaries, called *short-circuit* diffusion, will allow rapid transport of atoms across a solid, while bulk diffusion requires the presence of point defects throughout the crystal.

If vacancies are present, atoms or ions can jump from a normal site into a neighbouring vacancy and so gradually move through the crystal (Figure 5.12). Movement of a diffusing atom into a vacant site corresponds to movement of the vacancy in the other direction. This process is therefore frequently referred to as *vacancy diffusion*. In practice, it is often very convenient, in problems where vacancy diffusion occurs, to ignore atom movement and to focus attention upon the diffusion of the vacancies as if they were particles. In the case of interstitials, two diffusion mechanisms can be envisaged. An interstitial can jump to a neighbouring interstitial position. This is called *interstitial diffusion* and is the mechanism by which tool steels are hardened by incorporation of nitrogen or carbon. Alternatively, an interstitial can jump to a filled site and displace the occupant into a neighbouring interstitial site. This knock-on process is called *interstitialcy diffusion*.

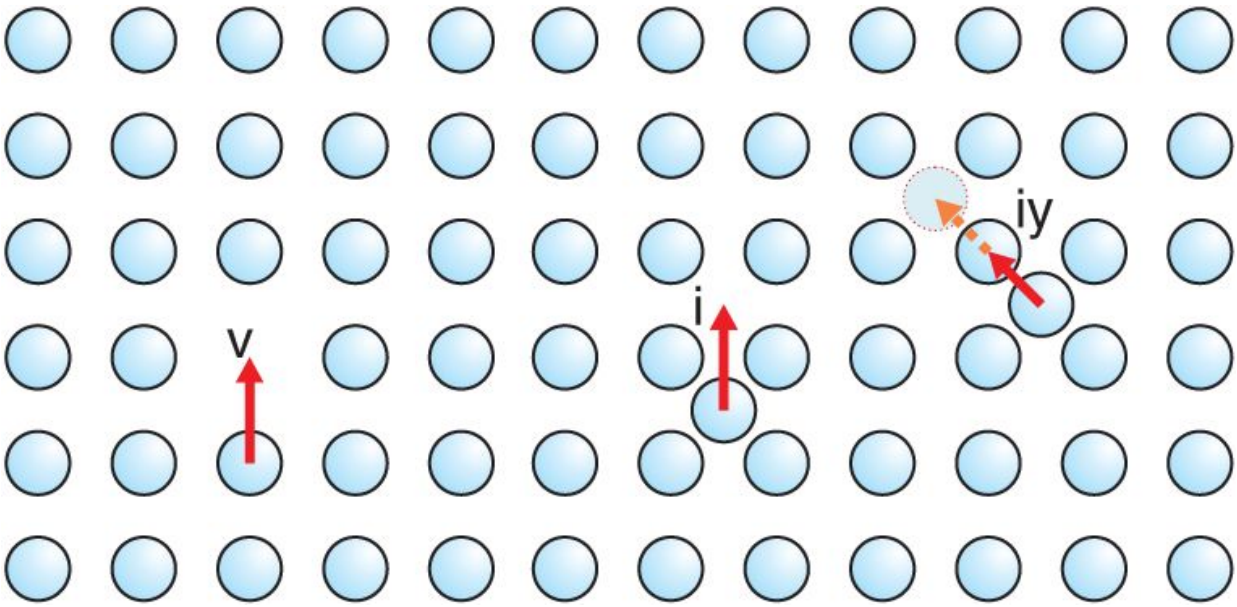


Figure 5.12 Self-diffusion mechanisms: v, vacancy; i, interstitial; iy, interstitialcy.

Substitutional impurity defects can move by way of three mechanisms (Figure 5.13). As well as vacancy diffusion, an impurity can swap places with a neighbouring normal atom, *exchange diffusion*, while in *ring diffusion* co-operation between several atoms is needed to make the exchange. These processes have been found to take place during the doping of semiconductor crystals. Interstitial impurities can move by interstitial and interstitialcy jumps similar to those described earlier. When Schottky defects are present in a crystal, vacancies are found on both the cation and anion sublattices, allowing both cation and anion diffusions to occur (Figure 5.14a). In the case of Frenkel defects, interstitials and vacancies occur, allowing interstitial, interstitialcy, and vacancy diffusion to take place in the same crystal (Figure 5.14b).

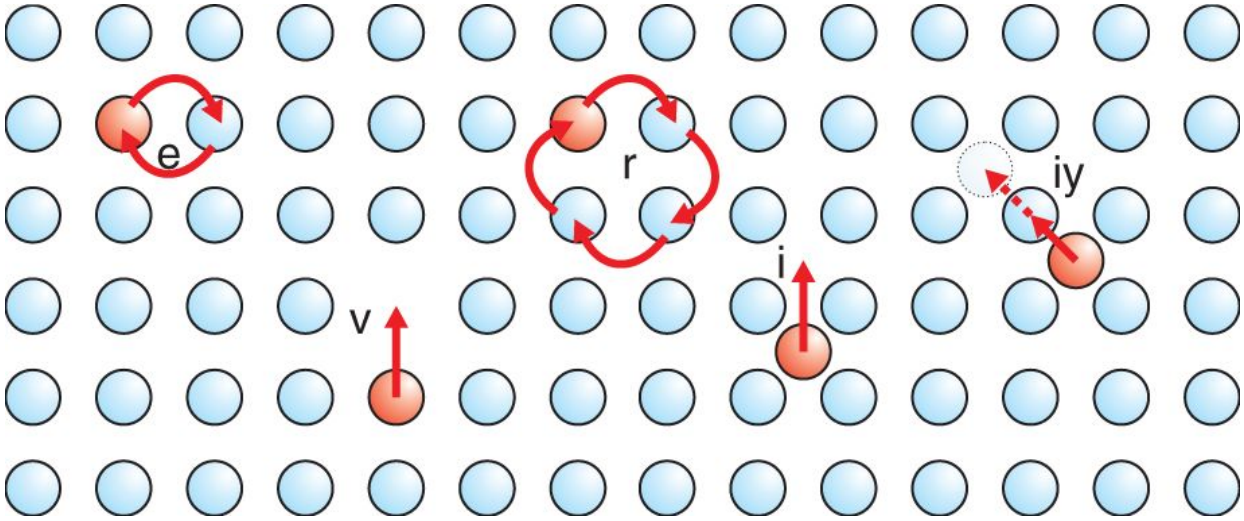


Figure 5.13 Impurity diffusion mechanisms; v, vacancy; e, exchange; r, ring; i, interstitial; iy, interstitialcy.

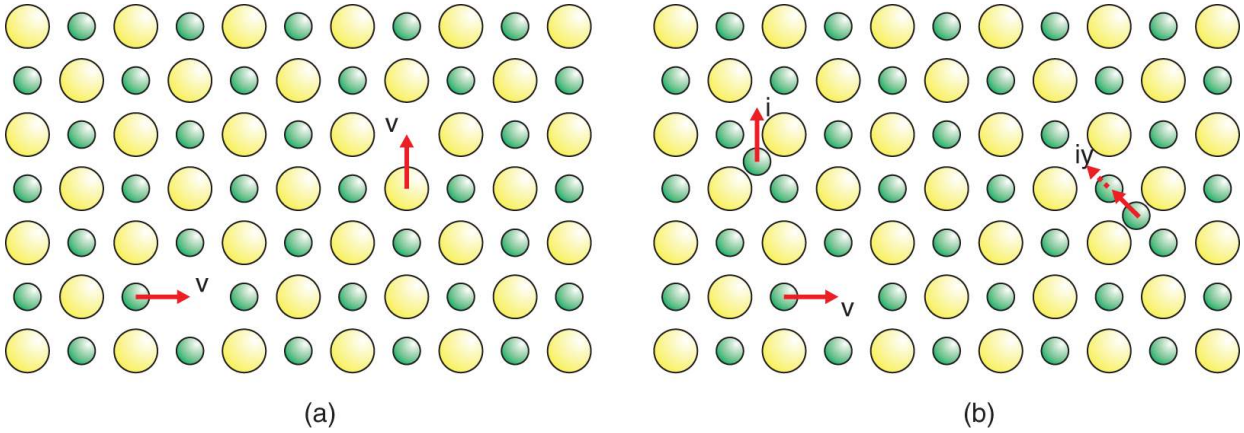


Figure 5.14 Diffusion mechanisms: (a) a crystal containing Schottky defects; (b) a crystal containing Frenkel defects.

It is of considerable practical importance to have some idea of how far an atom or ion will diffuse into a solid during a diffusion experiment. For example, the electronic properties of integrated circuits are created by the careful diffusion of selected dopants into single crystals of very pure silicon. Similarly, metallic components are hardened by diffusing carbon or nitrogen from the surface into the bulk. Vacancy diffusion and interstitial diffusion are quite similar to random-walk diffusion. For example, a vacancy can jump to any neighbouring position in a crystal regardless of whether it is occupied by an atom or not, in a random way. If the diffusion process follows these

mechanisms, the penetration depth, defined earlier, is a good indicator of the spatial extent of diffusion.

5.7 Self-Diffusion in One Dimension

During self-diffusion, each time an atom moves it will have to overcome an energy barrier. This is because the migrating species have to leave normally occupied positions that are, by definition, the most stable positions for atoms in the crystal, to pass through less stable positions normally unoccupied. Often atoms may be required to squeeze through a bottleneck in order to move at all (Figure 5.15a). The simplest model of this process is given by one-dimensional diffusion in which the moving atoms have to overcome a single barrier of height E (Figure 5.15b).

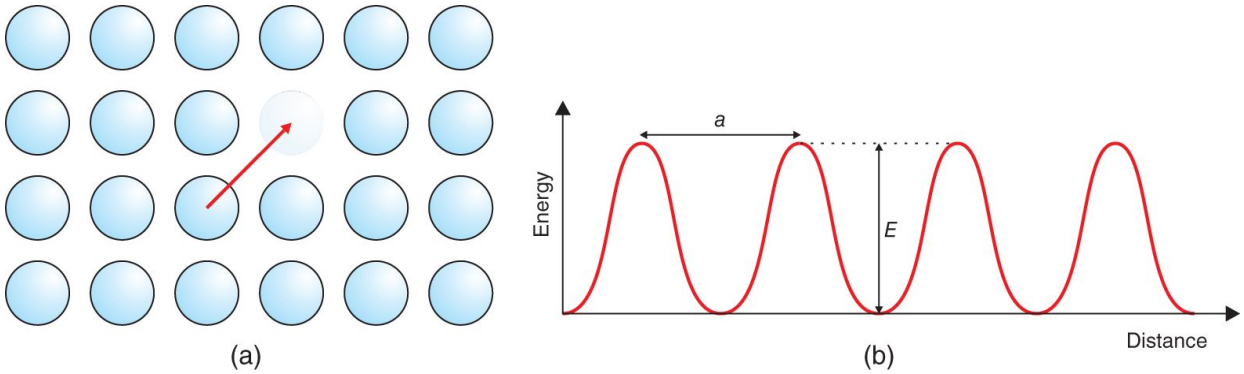


Figure 5.15 (a) Atom diffusion from one stable position to another; (b) The energy barrier E, encountered in a series of jumps of distance a.

The larger the magnitude of E, the less chance there is that the atom has the necessary energy to make a successful jump. The probability p that a single atom will move from one position of minimum energy to an adjacent position is given by Maxwell–Boltzmann statistics:

$$p = \exp(-E/k_B T)$$

where k_B is the Boltzmann constant and T the temperature (K). This equation indicates that if E is very small, the probability that the atom will clear the barrier approaches 1.0; if E is equal to $k_B T$ the probability for a successful jump is about one third, and if E increases above $k_B T$ the probability that the atom will jump the barrier rapidly becomes negligible.

The atoms in a crystal are vibrating continually with a frequency, ν, which is usually taken to have a value of about 10^{13} Hz at room temperature. It is reasonable to suppose that the number of attempts at a jump, sometimes called the *attempt frequency*, will be equal to the frequency with which the atom is vibrating. The number of successful jumps that an atom will make per second, Γ, will be equal to the attempt frequency multiplied by the probability of a successful move, i.e.:

$$\Gamma = \nu \exp(-E/k_B T) \qquad (5.2)$$

This flow of atoms, say along the x-direction, is related to the diffusion coefficient by Fick's first law:

$$J = -D\, dc/dx$$

To relate these equations, imagine two adjacent planes in a crystal, 1 and 2, separated by the atomic jump distance, a, and containing n_1 and n_2 diffusing atoms per unit area, respectively (Figure 5.16). If Γ_{12} is the frequency with which an atom moves from planes (1) to (2), then the numbers of atoms moving from planes (1) to (2) per second is j_{12}, where:

$$j_{12} = n_1 \Gamma_{12}$$

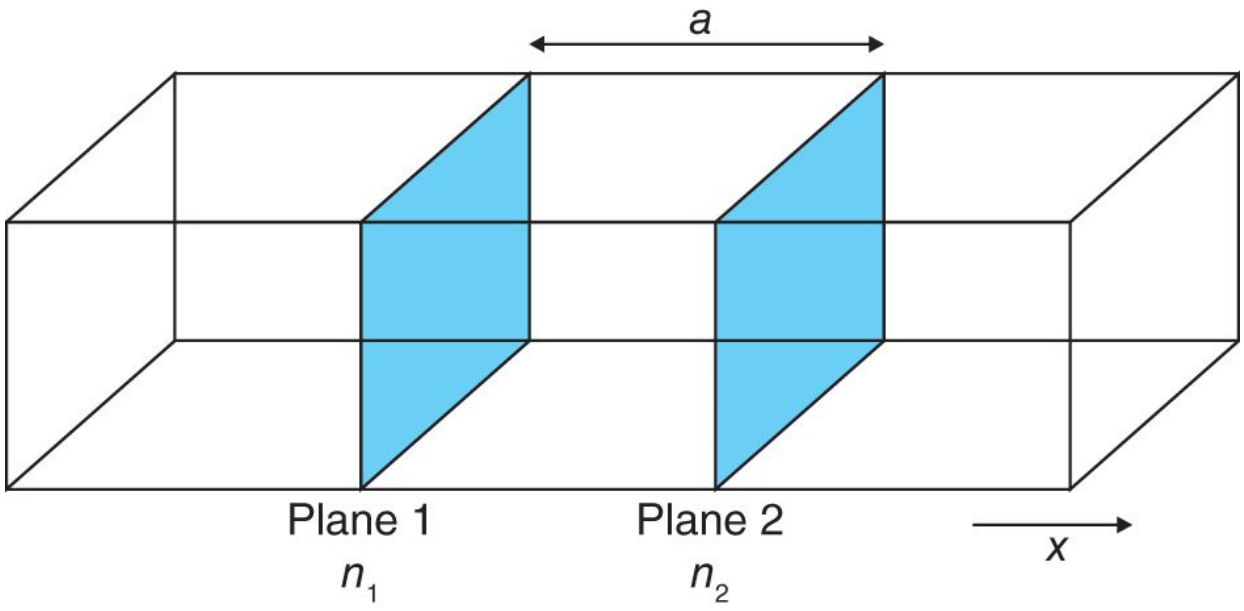


Figure 5.16 Two adjacent planes in a crystal, containing n_1 and n_2 diffusing atoms separated by the jump distance a.

Similarly, the number moving from plane (2) to plane (1) is j_{21} where:

$$j_{21} = n_2\Gamma_{21}$$

The net movement, the flux, between the planes, J, is given by:

$$J = j_{12} - j_{21} = (n_1\Gamma_{12} - n_2\Gamma_{21})$$

If the process is random, the jump frequency is independent of direction and we can set Γ_{12} equal to Γ_{21}. Moreover, half of the jumps, on average, will be in one direction and half will be in the opposite direction, hence:

$$\Gamma_{12} = \Gamma_{21} = \tfrac{1}{2}\Gamma$$

where Γ represents the overall jump frequency of the diffusing atoms. Thus:

$$J = \tfrac{1}{2}(n_1 - n_2)\Gamma$$

The number of mobile atoms on plane (1) is n_1 per unit area, so that the concentration per unit volume at plane (1), c_1, is n_1/a. Similarly, the concentration per unit volume at plane (2), c_2, is n_2/a. Thus:

$$(n_1 - n_2) = a(c_1 - c_2)$$

Hence:

$$J = \tfrac{1}{2}\, a(c_1 - c_2)\, \Gamma$$

The concentration gradient, $\mathrm{d}c/\mathrm{d}x$, is given by the change in concentration between planes (1) and (2) divided by the distance between planes (1) and (2), that is:

$$-\mathrm{d}c/\mathrm{d}x = (c_1 - c_2)/a$$

where the – sign is introduced as the concentration falls on moving from plane (1) to plane (2). Hence,

$$(c_1 - c_2) = -a\, \mathrm{d}c/\mathrm{d}x$$

and:

$$J = -\tfrac{1}{2}\, \Gamma a^2\, \mathrm{d}c/\mathrm{d}x$$

A comparison with Fick's first law shows that:

$$D = \tfrac{1}{2}\, \Gamma a^2$$

Substituting the expression for the jump frequency, Γ, in terms of the barrier height to be negotiated, E, yields:

$$D = \tfrac{1}{2}\, a^2\, \nu\, \exp(-E/k_{\mathrm{B}}T)$$

or in molar quantities:

$$D = \tfrac{1}{2}\, a^2\, \nu\, \exp(-E/RT)$$

5.8 Self-Diffusion in Crystals

In real crystals it is necessary to take some account of the three-dimensional nature of the self-diffusion process. An easy way of doing this is to add a *geometrical factor*, g, into equation for D so that it becomes:

$$D = g\, a^2\, \nu\, \exp(-E/RT)$$

In the one-dimensional case, the factor ½ is a geometrical term to account for the fact that an atom jump can be in one of two directions. In a cubic structure diffusion can occur along six equivalent directions and a value of g of ⅙ is appropriate.

In the foregoing discussion, every possible atom jump is allowed. This may not be true in real crystals. Eq. (5.2) ignores this, and should contain a term, p_J, that expresses the probability that the jump is possible from this structural point of view:

$$\Gamma = p_J\, \nu\, \exp(-E/RT)$$

The diffusion coefficient is then given by:

$$D = p_J\, g\, a^2\, \nu\, \exp(-E/RT) \tag{5.3}$$

For example, in the face-centred cubic structure, each metal atom is surrounded by 12 nearest neighbours. If, on average throughout the crystal, two of these sites are empty, the probability of a successful jump will be $p_J = {}^2/_{12} = {}^1/_6$. In a similar way, the diffusion of an impurity atom in a crystal, say, K in NaCl, involves factors such as the relative size of the impurity compared with the host atoms. In the case of ionic movement, the charge on the diffusing species will also play a part. These uncertainties can also be expressed by the inclusion of a p_J, term.

5.9 The Arrhenius Equation and Point Defects

A comparison of Eq. (5.3) with the Arrhenius equation (Section 5.3) shows that the pre-exponential factor D_0 is equivalent to:

$$D_0 = p_J\, g\, a^2\, \nu$$

The term p_J, which is the probability that a jump can take place for structural reasons, is closely related to the number of defects present. In most ordinary solids, the value of p_J is fixed by the impurity content. Any variation in D_0 from one sample of a material to another is accounted for by the variation of the impurity content. However, the value of p_J does not affect the energy of migration, E, so that Arrhenius plots for such crystals will consist of a series of parallel lines (Figure 5.17).

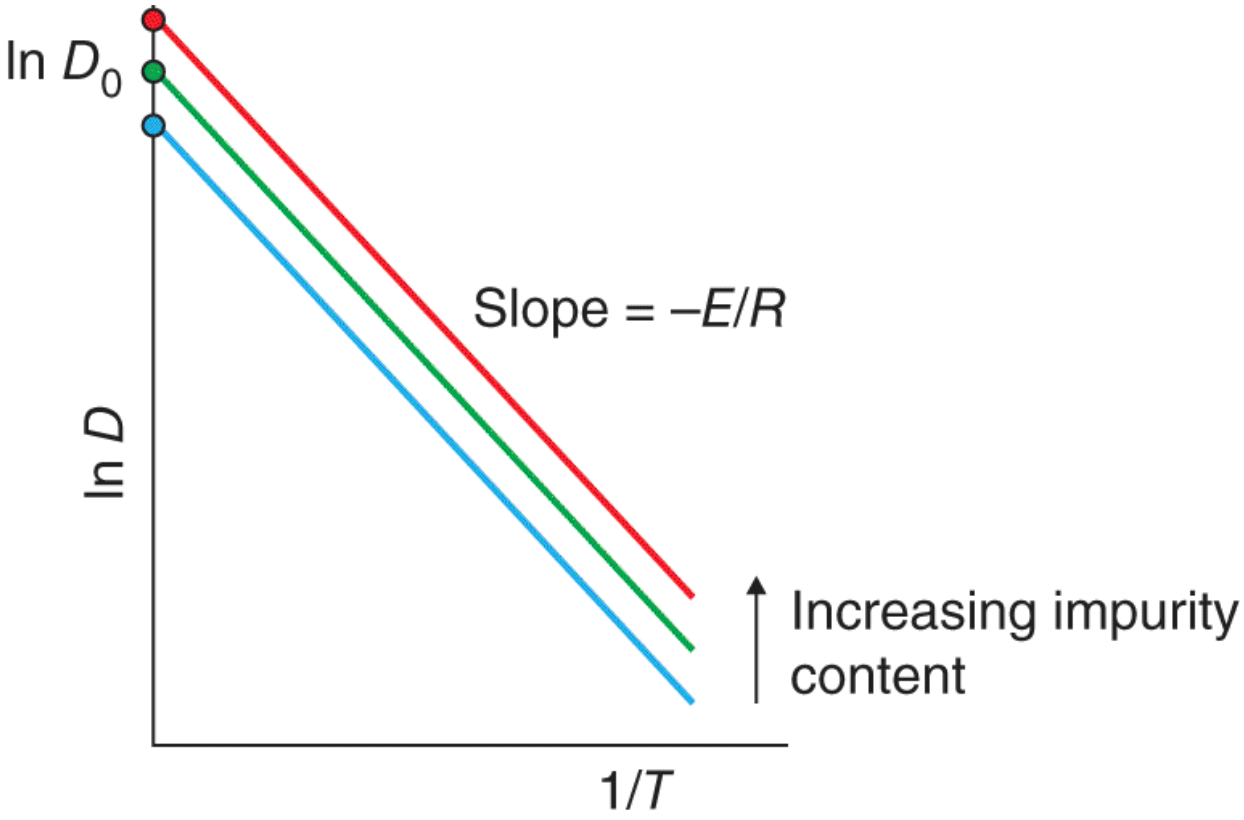


Figure 5.17 The variation of Arrhenius plots with impurity content. D, diffusion coefficient; T, temperature (K).

In very pure crystals, another feature becomes important. In this case, the number of intrinsic defects may be greater than the number of defects due to impurities, especially at high temperatures. Under these circumstances, the value of p_J will be influenced by the intrinsic defect population, and can contribute to the observed value of E.

To illustrate this, suppose that cation vacancy diffusion is the predominant migration mechanism, and the crystal, of formula MX, contains Schottky defects as the major type of intrinsic defects. The probability of a jump taking place is equal to the fraction of cation vacancies in the crystal, given by:

$$n_S = N \exp(-\Delta H_S/2\,RT)$$

At high temperatures the appropriate form of Eq. (5.3) becomes:

$$D = g\,\nu\,a^2 \exp(-E_v/RT) \exp(-\Delta H_S/2RT)$$

where E_v represents the height of the energy barrier to be overcome in vacancy diffusion and ΔH_S is the enthalpy of formation of Schottky defects.

Similarly, at high temperatures, diffusion in a crystal of formula MX by interstitials will reflect the population of Frenkel defects present, given by:

$$n_F = \sqrt{(NN^*)} \exp(-\Delta H_F/RT)$$

In these circumstances, the probability factor, p_J, will be proportional to the number of Frenkel defects present, and it is possible to write Eq. (5.3) as:

$$D = g\,\nu\,a^2 \exp(-E_i/RT) \exp(-\Delta H_F/2RT)$$

where E_i represents the potential barrier to be surmounted by an interstitial atom and ΔH_F is the enthalpy of formation of Frenkel defects.

Both of these equations retain the form:

$$D = D_o \exp(-E/RT)$$

However, the activation energy, E, will consist of the energy required for migration, E_i or E_v, plus the energy of defect formation. For Schottky defects

$$E = E_v + \Delta H_s/2$$

and for Frenkel defects:

$$E = E_i + \Delta H_F/2$$

This means that an Arrhenius plot will have a steeper slope at high temperatures, where the point defect equilibria are significant, than at low temperatures, where the impurity content dominates. The plot will show a knee between the high and low temperature regimes (Figure 5.8). A comparison of the two slopes will allow an estimate of both the energy barrier to migration and the relevant defect formation energy to be made. Some values found in this way are listed in Table 5.4.

Table 5.4 Approximate enthalpy values for the formation and movement of defects in alkali halide crystals.

Material	ΔH_s (kJ mol^{-1})	E_v (cation) (kJ mol^{-1})	E_v (anion) (kJ mol^{-1})
Schottky defects			
NaCl	192	84	109
NaBr	163	84	113
KCl	230	75	172
KBr	192	64	46
Frenkel defects			
AgCl	155	13	36
AgBr	117	11	23

In the foregoing discussion, it has been supposed that the height of the potential barrier will be the same at all temperatures. This is

probably not so. In addition, as the temperature increases the crystal will expand and E would be expected to decrease. Moreover, some of the other constant terms in the preceding equations, such as vibration frequency, may vary with temperature. Arrhenius plots reveal this by being slightly curved.

5.10 Correlation Factors for Self-Diffusion

So far, each step taken by a diffusing atom has been supposed to be unrelated to the preceding one. However, in some circumstances a given jump direction may be *correlated* with (may depend upon) the direction of the previous jump. Suppose the diffusion of a *vacancy* in a close packed structure, such as that of a typical metal, is the subject of investigation (Figure 5.18a). Clearly the vacancy can jump to any nearest neighbouring site. In general, there is no preference, so that the jump is entirely random. The same is true of each succeeding situation. Thus, the vacancy can always follow a truly random path. However, the diffusion of a *tracer* atom by the mechanism of vacancy diffusion is different. A tracer can only move if it is next to a vacancy, and in this case, the tracer can only jump to the vacancy (Figure 5.18b). The possibility of any other jump is excluded. Similarly, when the tracer has made the jump then it is equally clear that the most likely jump for the tracer is back to the vacancy (Figure 5.18c). The tracer can only jump to a new position after the vacancy has diffused to an alternative neighbouring position.

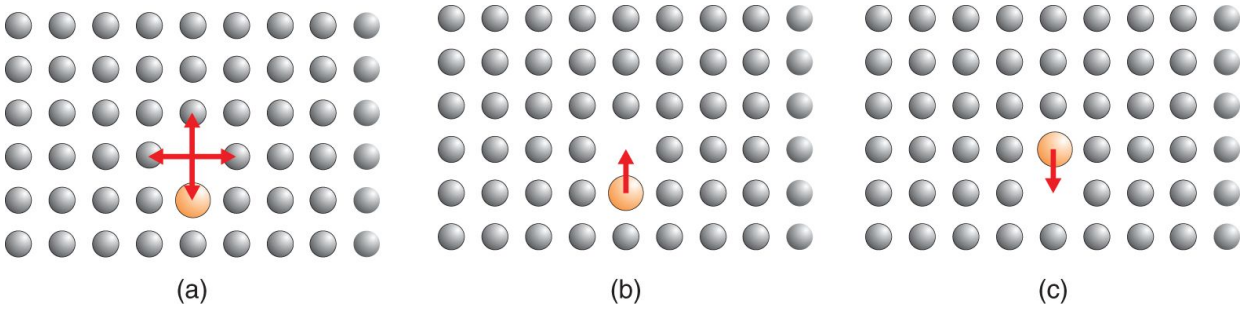


Figure 5.18 Correlated motion during vacancy diffusion: (a) a vacancy can jump to any neighbouring position; the diffusion of a tracer is correlated so that a jump into a vacancy (b) is most likely to be followed by a reverse jump (c).

The vacancy will follow a random walk diffusion route, while the diffusion of the tracer by a vacancy diffusion mechanism will be constrained. When these processes are considered over many jumps, the mean square displacement of the tracer will be less than that of the vacancy, even though both have taken the same number of jumps. Therefore, it is expected that the observed diffusion coefficient of the tracer will be *less* than that of the vacancy. In these circumstances,

the random-walk diffusion equations need to be modified for the tracer. This is done by ascribing a different probability to each of the various jumps that the tracer may make. The result is that the random walk diffusion expression must be multiplied by a *correlation factor*, *f*, which takes the diffusion mechanism into account. For the truly random diffusion of a vacancy, a correlation factor of 1.0 is appropriate.

In the case of *interstitial* diffusion, in which we have only a few diffusing interstitial atoms and many available empty interstitial sites, random walk equations would be fairly accurate, and a correlation factor close to 1.0 would be expected. This will be so whether the interstitial is a native atom or a tracer atom. When tracer diffusion by an *interstitialcy* mechanism is considered this will not be true and the situation is analogous to that of tracer atom vacancy diffusion. An initial jump can be to any atom position in the structure, but the most likely next jump is a return jump. Once again, the diffusion of the interstitial is different from that of a completely random walk, and once again a correlation factor, *f*, is needed to compare the two situations.

The correlation factor, for any mechanism, is given by the ratio of the values of the mean square displacement of the atom (often the tracer) moving in a correlated motion to that of the atom (or vacancy) moving by a random walk process. If the number of jumps considered is large, the correlation factor, *f*, can be written as:

$$f = \langle x^2 \rangle_c / \langle x^2 \rangle_r = D^* / D_r$$

that is: $D^* = f D_r$where $\langle x^2 \rangle_c$ represents the mean square displacement of a correlated walk by the diffusing atom and $\langle x^2 \rangle_r$ the mean square displacement for a truly random diffusion process with the same number of jumps, and D_r and D^* are the random walk and tracer diffusion coefficients. The correlation factor will then be a function of the diffusion mechanism and the crystal structure matrix in which the diffusion occurs (Table 5.5).

Table 5.5 Correlation factors for self-diffusion.

Crystal structure	**Correlation factor (*f*)**
Vacancy diffusion	
Diamond	0.50
Tungsten (bcc, A2)	0.73
Copper (fcc, A1)	0.78
Magnesium (hcp, A3), all axes	0.78
Corundum (α-Al_2O_3) (cations) || **a**	0.50
Corundum (α-Al_2O_3) (cations) || **c**	0.65
Colinear interstitialcy diffusion	
Diamond	0.73
Sodium chloride (NaCl) (cations or anions)	0.33
CsCl (cations or anions)	0.33
Fluorite (CaF_2) cations	0.4
Fluorite (CaF_2) anions	0.74

Source: All data from Compaan and Haven (1956).

5.11 Ionic Conductivity

Ionic conductivity in solids refers to the passage of ions across a solid under the influence of an externally applied electric field. The ionic conductivity of a material, due to the movement of cations and anions, is given by:

$$\sigma = c_a Z_a e \mu_a + c_c Z_c e \mu_c$$

where σ is the conductivity, c_a, c_c represent the concentrations of mobile anions and cations, $Z_a e$, $Z_c e$ are their charges, and μ_a, μ_c are their mobilities, defined as the drift velocity of the ions under unit applied field. The fraction of the conductivity that can be apportioned to each ion is called its *transport number*, defined by:

$$\sigma_c = t_c \sigma$$

$$\sigma_a = t_a \sigma$$

where σ_c, σ_a are the conductivities of the cations and anions, and t_c, t_a are the transport numbers for cations and anions, respectively. As can be seen from these relationships:

$$\sigma = \sigma (t_c + t_a)$$

$$t_c + t_a = 1$$

During ionic conductivity, ions jump from one stable site to another separated by a distance *a*. Hence, the process can be described by equations similar to those for diffusion. The movement of the ions, however, is not random, but is influenced by the presence of an electric field, *V*, so that positive and negative ions move in opposite directions. The electric field changes the potential barrier encountered by a diffusing ion (Figure 5.15) by reducing it by ½ *ZeaV* in one direction and raising it by the same amount in the other (Figure 5.19). In crystals of low symmetry, and in directions at an angle to the applied field, similar barriers will appear, but of varying heights, so that some ionic conductivity directions may be favoured.

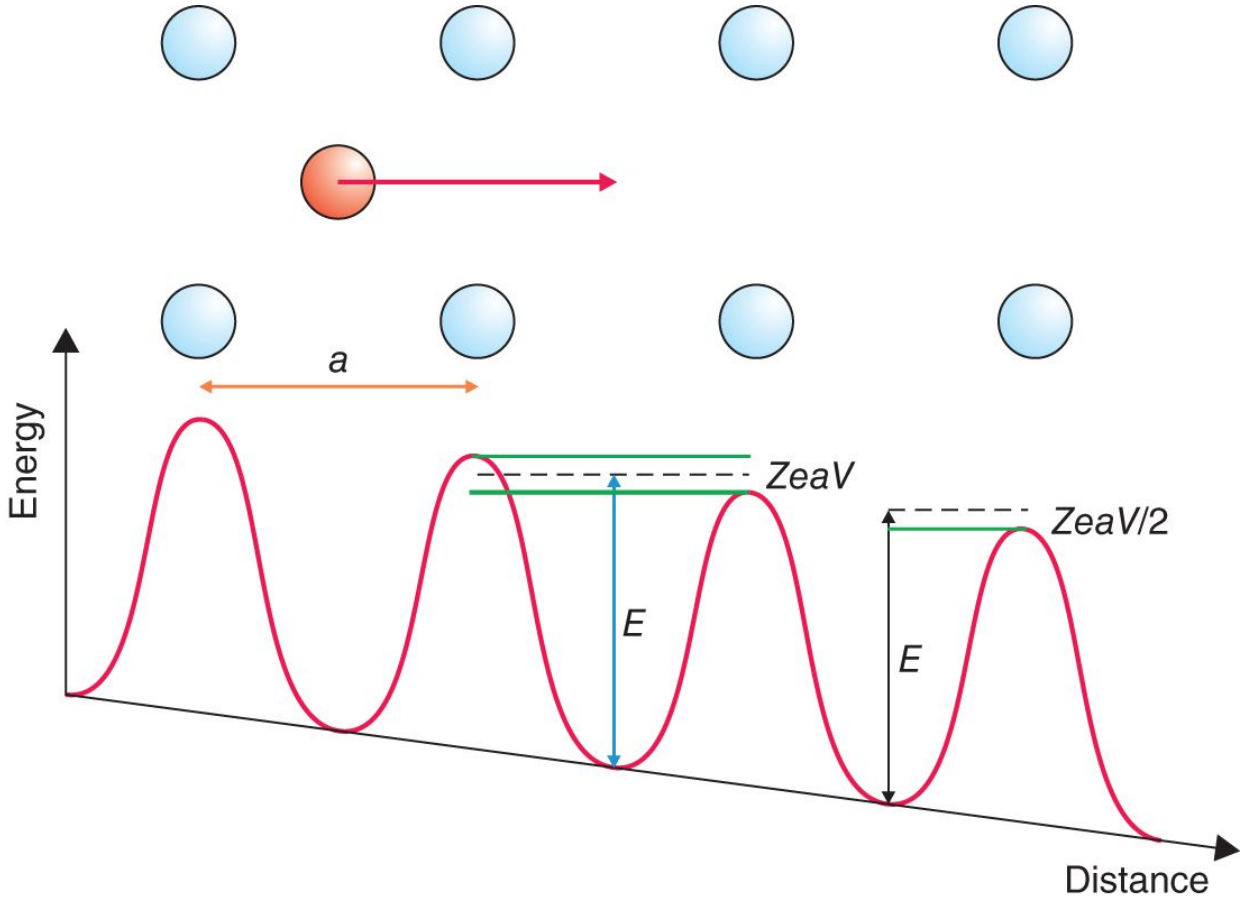


Figure 5.19 Energy barriers to be surmounted by an ion during ionic conductivity.

Following the methods set out for random diffusion, it is possible to calculate the relative number of jumps that an ion will make with and against the field and hence estimate the ionic conductivity. The number of jumps that an ion will make in the direction of the field per second is given by:

$$\Gamma_+ = \nu \exp[-(E - \tfrac{1}{2} ZeaV)/k_B T]$$

where the potential barrier is ($E - \frac{1}{2} ZeaV$) and ν is the vibration frequency of the solid. In a direction against the field, the number of successful jumps will be given by:

$$\Gamma_- = \nu \exp[-(E + \tfrac{1}{2} ZeaV)/k_B T]$$

where the potential barrier is (E + ½ $ZeaV$). The overall jump rate in the direction of the field is ($\Gamma_+ - \Gamma_-$), and as the net velocity of the ions in the direction of the field, v, is given by the net jump rate multiplied by the distance moved at each jump we can write:

$$\begin{aligned} v &= \nu\, a\{\exp[-(E - \tfrac{1}{2}ZeaV)/k_B T] \\ &\quad - \exp[-(E + \tfrac{1}{2}eaV)/k_B T]\} \\ &= \nu\, a\ \exp(-E/k_B T)\,[\exp(ZeaV/2k_B T) \\ &\quad - \exp(-ZeaV/2k_B T)] \end{aligned}$$

For low field strengths $ZeaV$ is much less than $k_B T$, and [exp($ZeaV$/2 $k_B T$) – exp (–$ZeaV$/2 $k_B T$)] may be replaced by $ZeaV/k_B T$.[1] Using this approximation:

$$v = (\nu Z\, a^2 eV/k_B T)\ \exp(-E/k_B T)$$

The mobility, μ, of the ion is defined as the velocity when the value of V is unity, so:

$$\mu = (\nu Z\, a^2 e/k_B T)\ \exp(-E/k_B T)$$

Often a geometrical factor g is included to take into account different crystallographic features to give:

$$\mu = (g\, \nu Z\, a^2 e/k_B T)\ \exp(-E/k_B T)$$

Substituting for μ in the equation:

$$\sigma = c\, Ze\, \mu$$

gives:

$$\sigma = (c\, g\, \nu\, a^2\, Z^2\, e^2/k_B T)\ \exp(-E/k_B T)$$

This equation takes the form:

$$\sigma T = \sigma_0 \exp(-E/k_B T)$$

where

$$\sigma_0 = (c\, g\, \nu\, a^2\, Z^2\, e^2/k_B)$$

When ion migration takes place via a vacancy diffusion mechanism:

$$\sigma_v T = \sigma_0 \exp(-E_v/k_B T)$$

and when it takes place by the migration of interstitials:

$$\sigma_i T = \sigma_0 \exp(-E_i/k_B T)$$

The value of E_v or E_i can be obtained from the Arrhenius-like plots of $\ln \sigma T$ vs. $1/T$ (Figure 5.20).

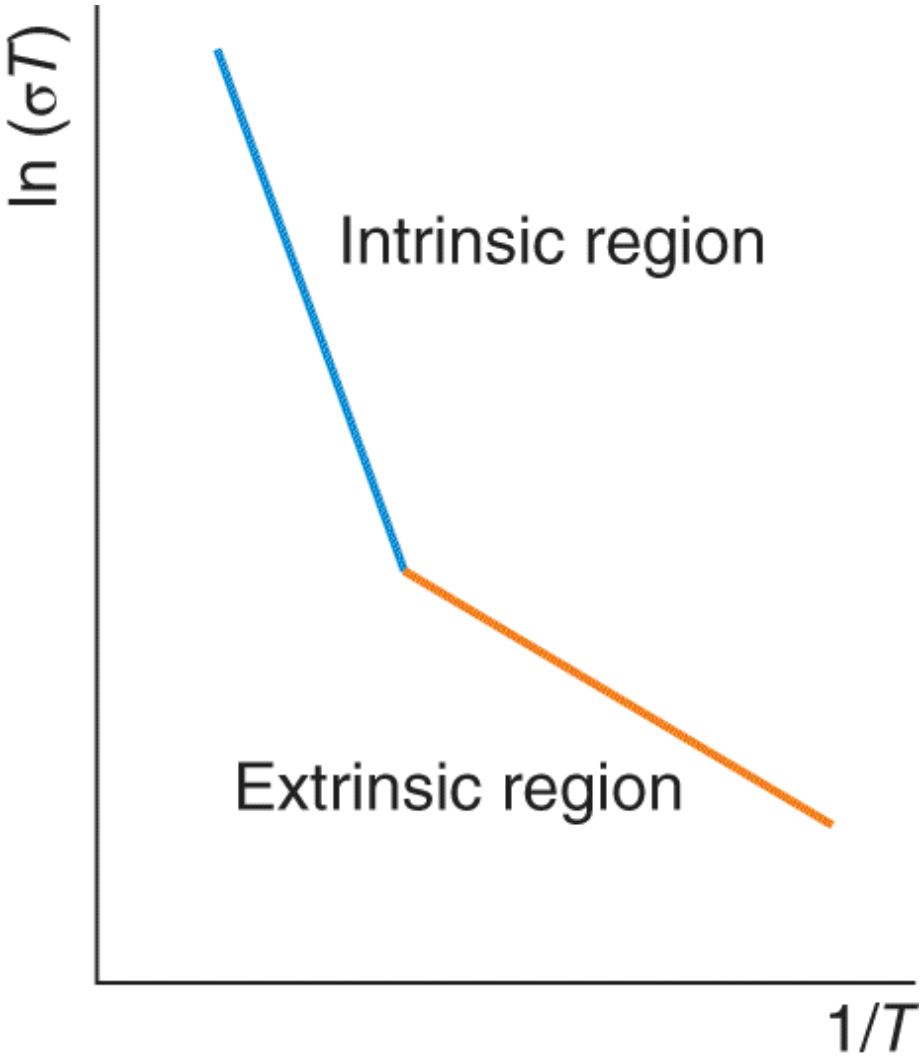


Figure 5.20 Arrhenius plot of ln (σT) vs. $1/T$ for ionic conductivity in an almost pure crystal.

In the case of very pure solids, it is necessary to take into account the number of intrinsic defects. The value of c must then reflect the type of intrinsic defect present. For example, should Schottky defects predominate, substitution for c gives:

$$\sigma_S = \sigma_0 \exp(-E_v/k_B T) \exp(-\Delta H_S/2k_B T)$$

Similarly, should Frenkel defects predominate:

$$\sigma_F = \sigma_0 \exp(-E_i/k_B T) \exp(\Delta H_F/2k_B T)$$

In these cases, Arrhenius plots of ln σT vs. 1/T will show a knee between the low temperature and high temperature regions (Figure 5.20). The high temperature value for E will be composed of two terms:

$$E_S = E_v + \Delta H_s/2$$

for Schottky defects, and:

$$E_F = E_i + \Delta H_F/2$$

for Frenkel defects.

5.12 The Relationship Between Ionic Conductivity and Diffusion Coefficient

If both ionic conductivity and ionic diffusion occur by the same random walk mechanism, a relationship between the self-diffusion coefficient and the ionic conductivity can be derived. In the simplest case, assume one-dimensional diffusion along x, and that both processes involve the same energy barrier, E, and jump distance, a. For this case the diffusion coefficient is:

$$D = g\,\nu\,a^2\ \exp\left(-E/k_B T\right)$$

The ionic conductivity of an ion is:

$$\sigma = \left(g\,c\,\nu\,a^2 Z^2 e^2/k_B T\right)\ \exp\left(-E/k_B T\right)$$

leading to the *Nernst–Einstein equation*:

$$D_\sigma = k_B T\,\sigma/c\,Z^2\,e^2$$

where Ze is the charge on the mobile ions, D_σ is the diffusion coefficient calculated from the conductivity, and c is the number of mobile ions per unit volume. Although this equation shows that it is possible to determine the diffusion coefficient from the easier measurement of ionic conductivity, D_σ is derived by assuming that the conductivity mechanism utilises a random walk mechanism.

The *Haven ratio*, H_R, is defined as:

$$H_R = D^*/D_\sigma$$

where D^* is the tracer diffusion coefficient. The correlation factor, f, is defined by the ratio of the tracer diffusion coefficient to the random walk diffusion coefficient (Section 5.10):

$$f = D^*/D_r$$

Taking D_σ to be equal to D_r (which may not always be correct) allows H_R to be equated with the correlation factor:

$$H_R = D^* / D_\sigma = f$$

When ionic conductivity is by way of interstitials, both conductivity and diffusion can occur by random motion, so that the correlation factor and H_R are both equal to 1. In general, the correlation factor for a diffusion mechanism will differ from 1, and in such a case D_σ can be described by the relationship:

$$D_\sigma = f\, k_B T\, \sigma / c\, Z^2\, e^2$$

where f is the correlation factor appropriate to the diffusion mechanism. For vacancy diffusion in a cubic structure,

$$D_\sigma = f_v\, k_B T\, \sigma / c\, Z^2\, e^2$$

where f_v is the correlation factor for vacancy self-diffusion, and:

$$D_\sigma = f_{iy}\, k_B T\, \sigma / c\, Z^2\, e^2$$

for interstitialcy diffusion, where f_{iy} is the appropriate interstitialcy correlation factor.

5.13 Superionic Conductors

Diffusion and ionic conductivity as described earlier are generally not sufficiently rapid for many applications. For example, the ionic conductivity of a typical ionic halide such as NaCl or KCl is of the order of $10^{-10} - 10^{-14}$ S m^{-1}, while solid-state batteries that operate at room temperature require an ionic conductivity of the order of $10^{2} - 10^{1}$ S m^{-1} or better to function. Solids that show such high ionic conductivities, called *superionic conductors*, were first discovered towards the end of the nineteenth century, but development of such materials accelerated during the latter half of the twentieth century, especially in response to need for high power, compact, and lightweight batteries. In order to achieve the necessary conductivities, the solid must have a high concentration of relevant ions, and these must also display high mobility. That is to say both the formation energy and migration energy of the mobile constituents must be low. Broadly speaking these properties are achieved as a direct result of the structures of the materials. Here four illustrative examples are given, to compare the way in which these attributes have been achieved:

(i) Disordered cation compounds, in which a population of cations is located in an array of neighbouring empty sites that are easily reached by a single jump;

(ii) Solids in which relatively unpopulated planes separate crystal slabs, typified by the β-aluminas;

(iii) Crystals with a high population of vacancies, allowing for easy atom movement, as in the stabilised zirconia materials;

(iv) Framework materials with sparsely populated tunnels, allowing easy migration along the tunnels, the NASICON related oxides.

5.13.1 Disordered Cation Compounds

The first disordered cation compounds to be studied, silver sulphide, Ag_2S, and lead fluoride, PbF_2, were investigated by Faraday in the nineteenth century, and it was he who discovered that at elevated

temperatures both showed conductivities similar to that of liquid electrolytes. The cuprous halides CuI and CuBr exhibit similar behaviour at temperatures of the order of 400 °C. The reason for the high conductivity of these and similar solids was not known until crystallographic studies were completed. The prototype material is silver iodide. When the compound is prepared at room temperature two crystallographic forms, β-AgI and γ-AgI, both occur, both with normal ionic conductivity. The phase possessing high ionic conductivity is the high temperature polymorph, α-AgI, stable above 147 °C. The structure of α-AgI consists of a body-centred cubic array of iodide (I^-) ions, with a unit cell edge of 0.5044 nm. There are two iodide (I^-) ions and (on average) two silver (Ag^+) in the unit cell. The silver ions occupy tetrahedral positions between the iodide ions in this array. There are four of these positions on each face of the cubic unit cell (Figure 5.21). In the low temperature polymorphs two specific tetrahedral positions are occupied by Ag^+, but in the high temperature form the Ag^+ ions jump continuously from one tetrahedral site to another forming a 'molten sublattice' of Ag^+ ions. The application of an electric field will bias the movement of the Ag^+ ions, resulting in the high ionic conductivity observed.

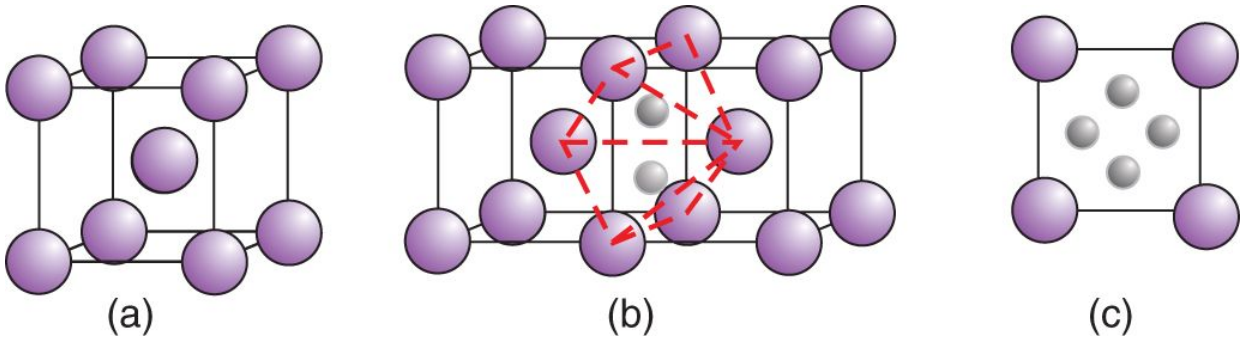


Figure 5.21 The structure of α-AgI: (a) body-centred cubic arrangement of I^- anions; (b) two (of four) tetrahedral sites on a cube face; (c) the four tetrahedral sites located on a cube face.

In the main, all of the disordered cation compounds behave similarly. At some higher temperature all or part of the metal atom array becomes mobile with cations able to freely move over some or all of the various allowed cation sites in the unit cell, while the anion array, or this and some of the large cations, remains relatively immobile. For

example, in the phase AgCuS the sulphur atoms form a slightly distorted hexagonal close-packed array. The Cu^+ ions are located in positions within this framework to form layers, while the Ag^+ ions lie between the sulphur – copper layers. These Ag^+ ions show a progressively greater anisotropic thermal motion as the temperature rises, until, above 93 °C, they are essentially completely mobile, leading to extremely high silver ion conductivity. The closely related compound Ag_3CuS_2 has a distorted body-centred cubic packing of sulphur atoms analogous to the body-centred cubic packing of I^- ions in α-AgI. The Ag^+ ions occupy distorted octahedral or tetrahedral sites and Cu^+ ions are in linear coordination at room temperature. The cation structure disorders at 117 °C, giving rise to high conductivity. In this material, although it seems that both Ag^+ and Cu^+ are mobile, the conductivity is dominated by Ag^+ movement.

5.13.2 β-Alumina Oxides

The phase β-alumina, with a nominal composition of $NaAl_{11}O_{17}$, is the prototype for a family of solids with compositions $A_2O{\cdot}nM_2O_3$, with n taking values of between 5 and 11, where A represents a large cation, typically Na, and M represents a trivalent ion, usually Al. These solids have a high ionic conductivity, with transport numbers for the A cations close to 1.0 over a wide range of temperatures.

The unit cell of β-alumina is hexagonal, with lattice parameters a = 0.595 nm, c = 2.249 nm. The dominant features of the idealised structure of β-alumina, with composition $NaAl_{11}O_{17}$, are layers called *spinel blocks* stacked perpendicular to the **c**-axis (Figure 5.22a). These blocks are composed of four oxygen layers in a cubic close packed arrangement with a thickness of approximately 0.66 nm. Within each slab Al^{3+} ions occupy octahedral and tetrahedral positions, so the structure resembles a thin slice of spinel, $MgAl_2O_4$. There are two such sheets in a unit cell. Adjacent sheets are linked by AlO_4 tetrahedra, which share a common apical oxygen atom. These oxygen atoms are often called *bridging* or *column oxygen*. The spinel

blocks are easily separated, and β-alumina cleaves readily into mica-like foils along these planes.

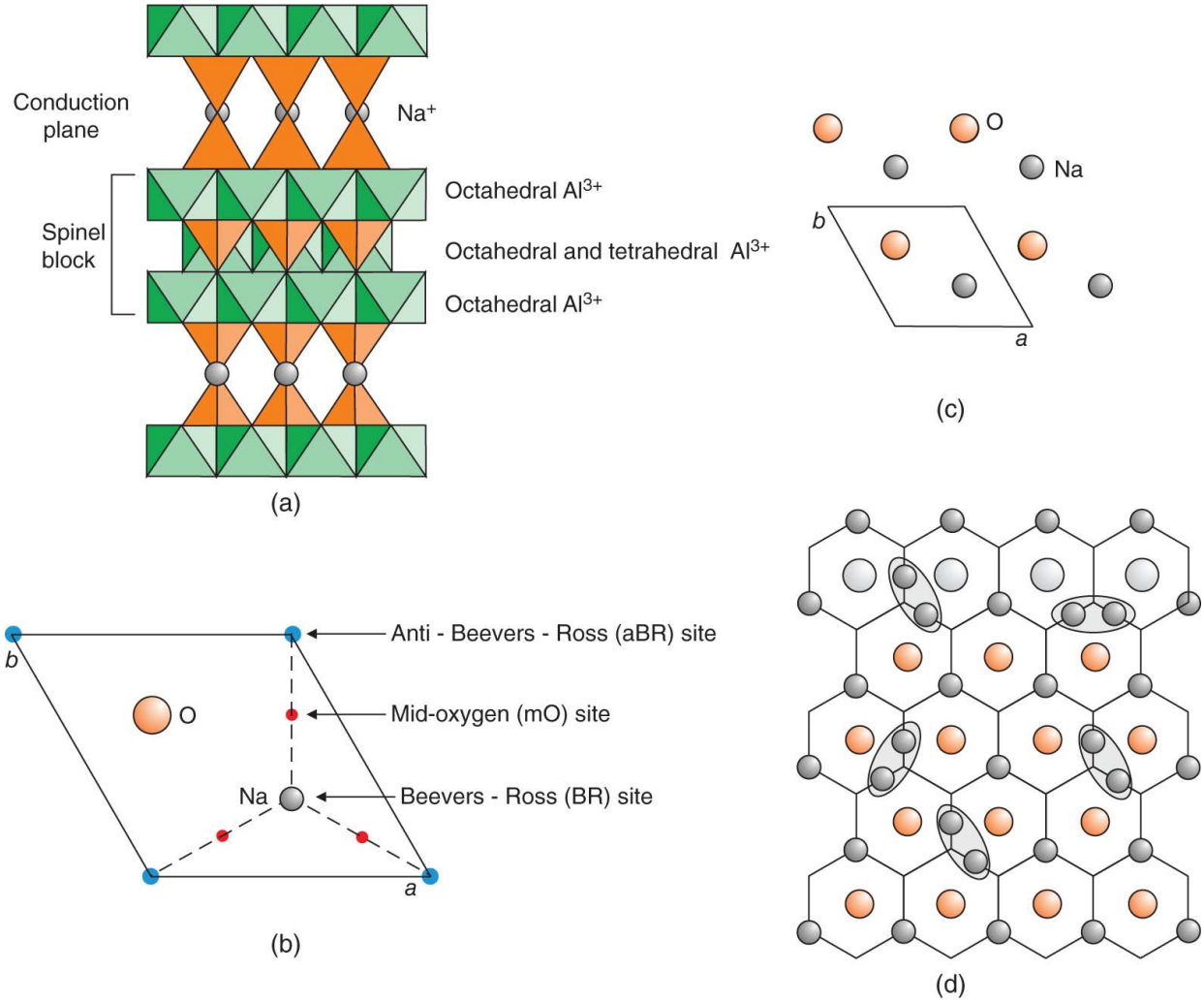


Figure 5.22 (a) Idealised β-alumina structure projected down [110] drawn as Al-centred polyhedra; (b) the conduction plane in one unit cell; (c) ideal disposition of atoms in the conduction plane; (d) possible pair-wise disposition of excess Na^+ in the conduction plane.

The Na^+ ions reside in the region between the spinel blocks, called the *conduction planes*, which also contain the apical oxygen atoms of the linking tetrahedra (Figure 5.22b,c). The conduction planes have a thickness of about 0.46 nm. The position occupied by Na^+, called the *Beevers–Ross* (BR) site, is diagonally opposite to the apical bridging oxygen atom of the tetrahedral groups linking the spinel blocks. Other important positions in the unit cell are the corners, called anti-

Beevers–Ross (aBR) sites and the positions mid-way between the BR and aBR sites, called mid-oxygen (mO) sites.

The idealised composition of β-alumina, $NaAl_{11}O_{17}$, is never found and an excess of sodium is necessary to stabilise the phase. The excess sodium is balanced by the incorporation of additional O^{2-} ions into the conduction planes, so that the material has a formula $Na_{1+x}Al_{11}O_{17+x/2}$ with x typically taking values between 0.15 and 0.3. The location of the extra sodium and oxygen varies with temperature and composition. Broadly speaking, the excess sodium forms pairs with an Na^+ ion on a BR site (Figure 5.22d). The pair then occupies two mO sites adjacent to a BR site so that the originally occupied BR site becomes vacant. The additional oxygen atoms (not shown in Figure 5.22d), also located in the conduction plane, tend to share sites with normal bridging oxygen atoms. The resulting pairs lie along lines joining the oxygen site to the adjacent corners of the unit cell, in positions rather similar to the mid-oxygen sites surrounding the sodium ions in BR sites.

The high Na^+ ion conductivity of β-alumina is attributed to the correlated diffusion of pairs of ions in the conduction plane. The sodium excess is accommodated by the displacement of pairs of ions onto mO sites (Figure 5.22d). A series of atom jumps will then allow the cations to reorient and diffuse through the crystal (Figure 5.23). Calculations suggest that this diffusion mechanism has a low activation energy, which would lead to high Na^+ ion conductivity. A similar, but not identical, mechanism can be described for related phases such as β″-alumina, described in the following text.

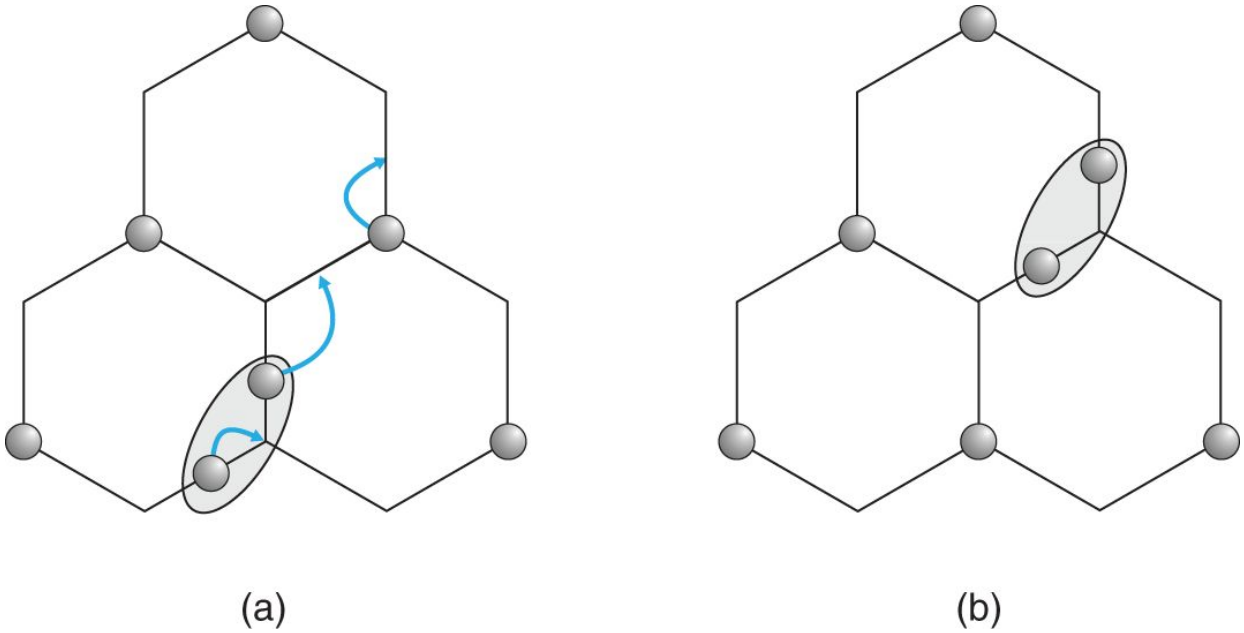


Figure 5.23 Possible correlated jumps to move an Na^+ pair in (a) to that in (b).

The spinel blocks in β-alumina are related by mirror planes that run through the conduction planes; that is, the orientation of one block relative to another is derived by a rotation of 180°. A second form of this compound, called β″-alumina, has similar spinel blocks. However, these are related to each other by a rotation of 120°, so that three spinel block layers are found in the unit cell, not two. The ideal composition of this phase is identical to that of β-alumina, but the unit cell is now rhombohedral. Referred to a hexagonal unit cell, the lattice parameters are a = 0.614 nm, c = 3.385 nm. The thickness of the spinel blocks and the conduction planes is similar in both structures.

The disposition of the atoms in the conduction plane itself is identical in to that in β-alumina. However, the local environment of the conduction planes is different in the two phases, because the oxygen ions on both sides of the conduction plane are superimposed in β-alumina and staggered in β″-alumina. This means that Beevers–Ross and anti-Beevers–Ross sites are not present. Instead the sodium sites, which have a tetrahedral geometry, are called Beevers–Ross type (BR – type). The mid-oxygen (mO) sites, located midway between the BR – type sites, also have a different geometry in the two phases.

The β″-alumina structure can be stabilised by the addition of a number of different monovalent or divalent cations, notably Li^+ or Mg^{2+}. The general formula for these phases is $Na_{1+x}Li_{x/2}Al_{11-x/2}O_{17}$ and $Na_{1+x}Mg_xAl_{11-x}O_{17}$ with x taking a value of approximately 0.67. Additional oxygen is not needed to maintain charge neutrality, which is taken care of by substitution of Li^+ and Mg^{2+} on Al^{3+} sites within the spinel blocks. There is almost unimpeded motion in the Na^+ layers as β″-alumina lacks interstitial oxygen ions in the conduction plane, and generally the β″-alumina phases are better ionic conductors than their β-alumina counterparts, with a conductivity of the same order of magnitude as in a strong solution of a sodium salt in water.

There are a number of analogous phases that can be fabricated using different A cations. The structures of these differ in small details from those described earlier. For example, in the silver analogue of β-alumina, the excess silver occupies aBR sites. When higher valence ions replace the sodium, the BR sites are generally occupied, but charge balance means that a considerable number of these sites are empty, creating a population of vacancies. Ideally half are unoccupied for A^{2+} cations and 2/3 are unoccupied for A^{3+} cations.

5.13.3 Stabilised Zirconia Oxides

A number of oxides with the fluorite structure are used in solid state electrochemical systems. They have formulae $AO_2{\cdot}xCaO$ or $AO_2{\cdot}xM_2O_3$, where A is typically Zr, Hf, and Th, and M is usually La, Sm, Y, Yb, or Sc. The technological importance of these materials lies in the fact that they are fast ion conductors for oxygen ions at moderate temperatures and are stable to high temperatures. This property is enhanced by the fact that there is negligible cation diffusion or electronic conductivity in these materials, which makes them ideal for use in a diverse variety of batteries and sensors.

The parent phase of this group is the oxide ZrO_2, which has the monoclinic baddeleyite structure at room temperature. As the temperature increases, the structure changes to a tetragonal phase at 1170 °C and then to the cubic CaF_2 (fluorite) structure at 2370 °C. The

high-temperature cubic form can, however, be stabilised to room temperature by reaction with the oxides noted earlier. In the fluorite structure the Ca^{2+} atoms form a face-centred cubic array, while the F^- anions form a cube within this array and occupy tetrahedrally coordinated positions (Figure 5.24). When calcia-stabilised zirconia, $ZrO_2{\cdot}xCaO$, is prepared, reaction with CaO forms a solid solution in which Ca^{2+} substitutes for Zr^{4+} and the charge discrepancy is balanced by a population of oxygen vacancies to form $Ca_xZr_{1-x}O_{2-x}$ (also see Section 3.6.4). Hence, an oxide containing 20 mol% CaO will have 20 mol% oxygen vacancies in the structure. Exactly the same situation holds when zirconia is reacted with the M_2O_3 oxides. In these cases, one oxygen vacancy will form for every two M^{3+} cations incorporated, leading to a generalised formula of $M^{3+}{}_xZr_{1-x}O_{2-x/2}$. For example, the reaction between Y_2O_3 and ZrO_2 forms yttria-stabilised zirconia, $Y^{3+}{}_xZr_{1-x}O_{2-x/2}$: Every two Y^{3+} substituents inserted into the crystal are balanced by one oxygen vacancy.

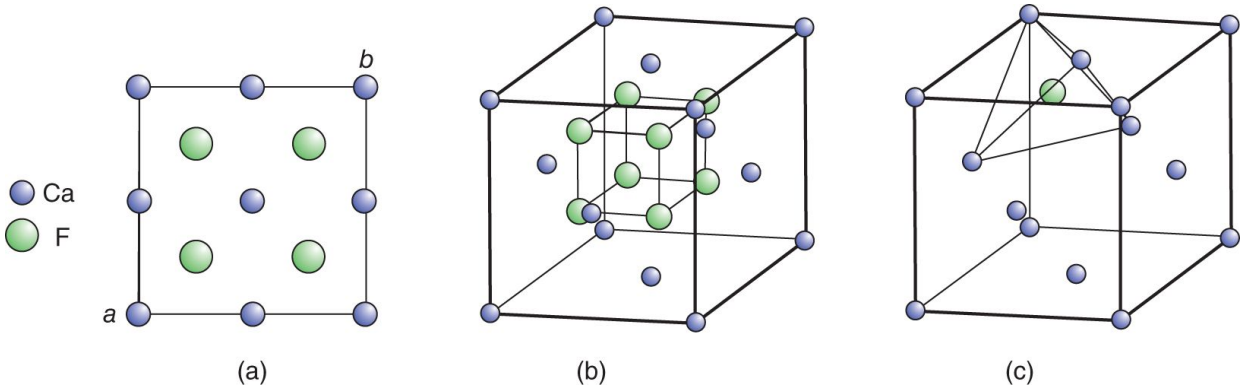


Figure 5.24 The fluorite structure: (a) projection of the structure down the **c**-axis; (b) perspective view if a unit cell; (c) the tetrahedral coordination around F.

The result of the high oxygen vacancy population is that the diffusion coefficient of oxygen ions is increased by many orders of magnitude compared to a normal oxide.

5.13.4 NASICON-Related Crystals

NASICON is an acronym for sodium (Na) superionic conductor and refers to a large number of solid phases, which have a framework structure permeated by tunnels running across planes within the material. The NASICON series is usually given the generic formula $Na_{1+x}Zr_2Si_xP_{3-x}O_{12}$, with $0 < x < 3$, typified by the phases $Na_2Zr_2(SiO_4)(PO_4)_2$ ($x = 1$), and $Na_3Zr_2(SiO_4)_2PO_4$ ($x = 2$). However, a large number of structurally related phases, given by the general formula $Na_xMM'(XO_4)_3$, where M and M′ are transition-metal atoms, including V, Ti, Fe, Nb, and so on, X is P or S, and with a composition range specified by $0 < x < 4$, have been synthesised. In addition, recent studies have focussed on analogous phases in which Li replaces Na, also known as LISICONS.

The first reported crystal structure of these phases was that of $NaZr_2(PO_4)_3$. This structure, with a hexagonal unit cell, $a = 0.8804$ nm, $c = 2.276$ nm, has a framework made up of ZrO_6 octahedra and PO_4 tetrahedra, which are linked by shared vertices so as to form tunnels running parallel to the **a**- and **b**-axes, within (0001) planes perpendicular to the **c**-axis. The Na^+ cations are located exclusively within these tunnels. The structures of the various NASICON and LISICON phases are similar, but more complex, as precise details vary with both composition and temperature. However, all are frameworks derived from a linkage of ZrO_6 octahedra and PO_4/SiO_4 tetrahedra joined by corner-sharing to form tunnels (Figure 5.25). The tunnels run in two directions, parallel to the **a**- and **b**-axes and on (0001) planes in hexagonal phases, as with $NaZr_2(PO_4)_3$. The Na^+ cations occupy the tunnel sites, the exact positions depending upon both the composition of the phase and the temperature.

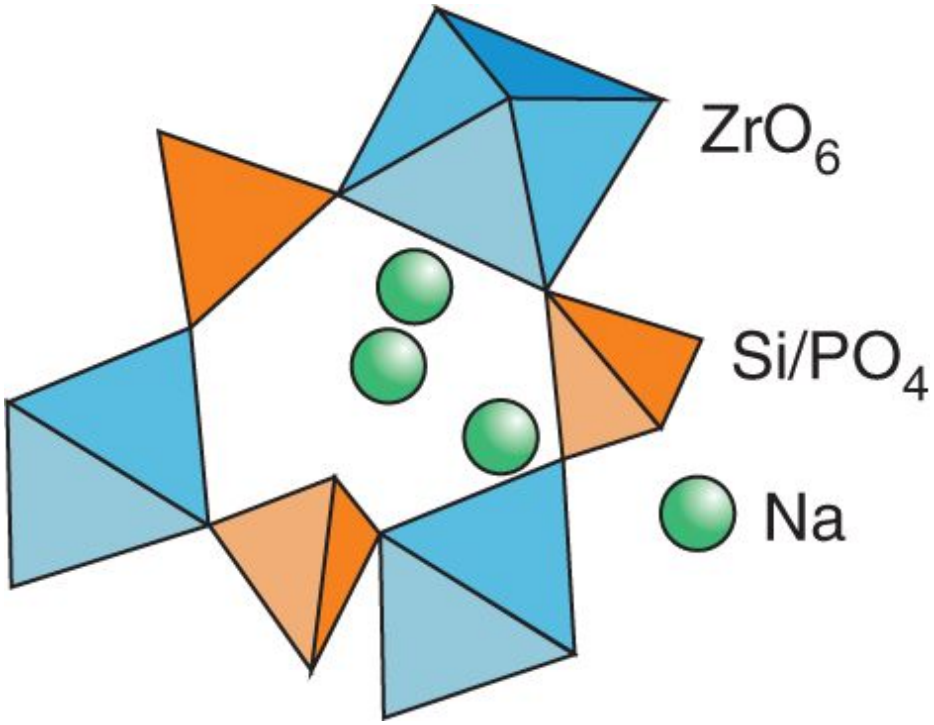


Figure 5.25 The tunnel framework in NASICON phases.

The high Na^+ (or Li^+) conductivity is due to the fact that these ions can migrate along the intersecting tunnel network with only minimal hindrance, which occurs at slight constrictions between polyhedra that make up the tunnel network. The phases with the highest Na^+ ion conductivity are close in composition to the $x = 2$ ($Na_3Zr_2(SiO_4)(PO_4)_2$ material, and these show ionic conductivity values close to that of the β-aluminas.

Reference

Compaan, K. and Haven, Y. (1956). Correlation factors for diffusion in solids. *Trans. Faraday Soc.* 52: 786–801; Correlation factors for diffusion in solids. Part 2. Indirect interstitial mechanism. 54, 1498–1508 (1958).

Further Reading

Open University Course Team (2009). *Random Walks and Diffusion*. Milton Keynes: Open University.

Tilley, R.J.D. (2008). *Defects in Solids*. Hoboken, NJ: John Wiley & Sons, Chapters 5, 6.

Yoon, M., Suh, K., Natarajan, S., and Kim, K. (2013). Proton-conduction in metal–organic frameworks and related modularly built porous solids. *Angew. Chem. Int. Ed.* 52: 2688.

Superionic Conductors: See Also References Therein:

Chen, S., Wu, C., Shen, L. et al. (2017). Challenges and perspectives for NASICON-type electrode materials for advanced sodium–ion batteries. *Adv. Mater.* 29: 1700431.

Chen, M., Hua, W., Xiao, J. et al. (2019). NASICON-type air-stable and all-climate cathode for sodium-ion batteries with low cost and high-power density. *Nat. Commun.* 10: 1480.

Ghidiu, M., Ruhl, J., Culver, S.P., and Zeier, W.G. (2019). Solution-based synthesis of lithium thiphosphate superionic conductors for solid state batteries: a chemistry perspective. *J. Mater. Chem.* 7: 17735.

Hou, M., Liang, F., Chen, K. et al. (2020). Challenges and perspectives of NASICON-type solid electrolytes for all-solid-state lithium batteries. *Nanotechnology* 31: 132003.

Ohkoshi, S.-i., Nakagawa, K., Imoto, K. et al. (2020). A switchable polar crystal that exhibits superionic conduction. *Nat. Chem.* 12: 338.

Zhao, Q., Stalin, S., Zhao, C.-Z., and Archer, L.A. (2020). Designing solid-state electrolytes for safe energy-dense batteries. *Nat. Rev. Mater.* 5: 229.

Problems and Exercises

Quick Quiz

5.1 Diffusion through a crystalline structure is called:

a. Tracer diffusion.

b. Volume diffusion.

c. Self-diffusion.

5.2 Self-diffusion is diffusion of:

a. Radioactive atoms in a crystal.

b. Impurity atoms in a crystal.

c. Native atoms in a pure crystal.

5.3 Tracer diffusion refers to the diffusion of:

a. Marked atoms.

b. Traces of impurities.

c. Single atoms.

5.4 A diffusion profile is a graph of:

a. Concentration vs. time.

b. Concentration vs. distance.

c. Distance vs. time.

5.5 In order to obtain the diffusion coefficient from a diffusion profile, use:

a. Fick's first law.

b. The Arrhenius equation.

c. The diffusion equation.

5.6 What graph would you plot to determine the tracer diffusion coefficient in a diffusion couple:

a. c vs. x.

b. ln c vs. x^2.

c. ln c vs. ln x.

5.7 Steady state diffusion is characterised by the fact that the concentration at:

a. Any point in the solid changes over time.

b. Any point in the solid is constant.

c. The surface of the solid changes over time.

5.8 In steady state diffusion, the diffusion coefficient is obtained via:

a. The diffusion equation.

b. The Arrhenius equation.

c. Fick's first law.

5.9 The variation of a diffusion coefficient with temperature is given by:

a. The diffusion equation.

b. Fick's first law.

c. The Arrhenius equation.

5.10 The activation energy for diffusion can be determined by a plot of:

a. ln D vs. $1/T$.

b. ln c vs. $1/T$.

c. ln D vs. T.

5.11 The part of an Arrhenius plot above a knee (i.e. the high temperature part), is called the:

a. Extrinsic region.

b. Intrinsic region.

c. Impurity region.

5.12 Vacancy diffusion refers to atoms moving into:

a. Vacant sites in the crystal structure.

b. Empty interstitial sites in the crystal structure.

c. Adjacent atom sites in the crystal structure.

5.13 Interstitialcy diffusion refers to atoms moving into:

a. Vacant sites in the crystal structure.

b. Empty interstitial sites in the crystal structure.

c. Adjacent atom sites and interstitial sites in the crystal structure.

5.14 The diffusion mechanism of exchange involves:

a. Two impurity atoms.

b. An impurity atom and a vacancy.

c. An impurity atom and a normal atom.

5.15 In crystals containing Frenkel defects:

a. One diffusion mechanism is possible.

b. Two diffusion mechanisms are possible.

c. Three diffusion mechanisms are possible.

5.16 The geometrical factor in diffusion varies according to:

a. The crystal structure:

b. The number of defects present.

c. The size of the diffusing atoms.

5.17 Correlation factors take into account:

a. Non-random diffusion.

b. Impurity content.

c. Frenkel and Schottky defects.

5.18 The high ionic conductivity in α-AgI is due to:

a. I^- ion migration.

b. Ag^+ ion migration.

c. Both Ag^+ and I^- ion migration.

5.19 In β-alumina superionic conductors the high ionic mobility is attributed to the correlated motion of pairs of Na^+ ions:

a. Between the spinel blocks.

b. Within the spinel blocks.

c. From one spinel block to a neighbouring block.

5.20 The superionic conductivity in calcia-stabilised zirconia is due to:

a. Ca^{2+} diffusion.

b. O^{2-} diffusion.

c. Zr^{4+} diffusion.

5.21 The high ionic conductivity in the NASICON phases is attributed to:

a. A framework structure.

b. A layer structure.

c. A disordered structure.

Calculations and Questions

5.1 Show that the units of the diffusion coefficient are $m^2\ s^{-1}$.

5.2 Show that the two equations are equivalent

a. $$\frac{c_x - c_0}{c_s - c_0} = 1 - \text{erf}\left(\frac{x}{2\sqrt{Dt}}\right)$$

b. $$\frac{c_s - c_x}{c_s - c_0} = \text{erf}\left(\frac{x}{2\sqrt{Dt}}\right)$$

5.3 Radioactive nickel-63 was coated onto a crystal of CoO and made into a diffusion couple. The sample was heated for 30 minutes at 953 °C. The radioactivity perpendicular to the surface

is given in the table. Calculate the impurity tracer diffusion coefficient of nickel-63 in CoO.

Activity (counts s^{-1})	**Distance (μm)**
80	6
50	10
20	14
6	18
5	20

5.4 Radioactive cobalt-60 was coated onto a crystal of CoO and made into a diffusion couple. The sample was heated for 30 minutes at 953 °C. The radioactivity perpendicular to the surface is given in the table. Calculate the tracer diffusion coefficient of cobalt-60 in CoO.

Activity (counts s^{-1})	**Distance (μm)**
110	10
70	20
39	30
23	40
9	50

5.5 Radioactive iron-59 was coated onto the (001) face of a single crystal of TiO_2 (rutile) (tetragonal) and made into a diffusion couple. The sample was heated for 300 seconds at 800 °C. The radioactivity perpendicular to the surface is given in the Table. Calculate the impurity tracer diffusion coefficient of iron-59 parallel to the **c**-axis in rutile.

Activity (counts s^{-1})	**Distance ($m \times 10^{-4}$)**
520	3
400	4
270	5
185	6
130	6.5

Activity (counts s^{-1})	Distance ($m \times 10^{-4}$)
90	7
53	8

5.6 Carbon-14 is diffused into pure α-iron from a gas atmosphere of CO + CO_2. The gas pressures are arranged to give a constant surface concentration of 0.75 wt% C. The diffusivity of ^{14}C into α-Fe is 9.5×10^{-11} m^2 s^{-1} at 827 °C. Calculate the concentration of ^{14}C 1 mm below the surface after a heating time of two hours.

5.7 Using the data in Question 5.6, how long would it take to make the carbon content 0.40 wt% at 1 mm below the surface?

5.8 An ingot of pure titanium metal is heated at 1000 °C in an atmosphere of ammonia, so that nitrogen atoms diffuse into the bulk. The diffusivity of nitrogen in β-Ti, the stable structure at 1000 °C, is 5.51×10^{-12} m^2 s^{-1} at this temperature. What is the thickness of the surface layer of titanium that contains a concentration of nitrogen atoms greater than 0.25 at.% after heating for one hour?

5.9 Zircalloy is a zirconium alloy used to clad nuclear fuel. How much oxygen will diffuse through each square metre of casing in a day under steady state diffusion conditions, at 1000 °C, if the following apply: diffusivity of oxygen in zircalloy at 1000 °C, 9.89×10^{-13} m^2 s^{-1}, concentration of oxygen on the inside of the container, 0.5 kg m^{-3}, oxygen concentration on the outside of the container, 0.01 kg m^{-3}, container thickness 1 cm.

5.10 Pure hydrogen is made by diffusion through a Pd-20%Ag alloy 'thimble'. What is the mass of hydrogen prepared per hour if the total area of the thimbles used is 10 m^2, the thickness of each is 5 mm, and the operating temperature is 500 °C? The equilibrium alloy in the surface of the thimble on the hydrogen rich side has a composition $PdH_{0.05}$, and on the further side, the hydrogen is swept away rapidly so that the surface is essentially pure metal. The diffusion coefficient of hydrogen in the alloy at

500 °C is 1.3×10^{-8} m^2 s^{-1}. Pd has the face-centred cubic (A1) structure, with lattice parameter a = 0.389 nm.

5.11 The radioactive tracer diffusion coefficient of silicon atoms in silicon single crystals is given in the table. Estimate the activation energy for diffusion.

T (°C)	D^* (m^2 s^{-1})
1150	8.82×10^{-19}
1200	3.40×10^{-18}
1250	1.20×10^{-17}
1300	3.90×10^{-17}
1350	1.18×10^{-16}
1400	3.35×10^{-16}

5.12 Using the data in Question 5.11, at what temperature (°C) will the penetration depth be 2 μm after 10 hours of heating?

5.13 The diffusion coefficient of carbon impurities in a silicon single crystal is given in the table. Estimate the activation energy for diffusion.

T (°C)	D^* (m^2 s^{-1})
900	1.0×10^{-17}
1000	3.0×10^{-16}
1100	4.0×10^{-15}
1200	5.0×10^{-14}
1300	9.0×10^{-13}
1400	5.0×10^{-12}

5.14 Using the data in Question 5.13, at what temperature (°C) will the penetration depth be 10^{-4} m after 20 hours of heating?

5.15 The diffusion coefficient of radioactive Co^{2+} tracers in a single crystal of cobalt oxide, CoO, is given in the table. Estimate the activation energy for diffusion.

T (°C)	D^* ($m^2\ s^{-1}$)
1000	1.0×10^{-13}
1100	3.5×10^{-13}
1200	9.0×10^{-13}
1300	2.0×10^{-12}
1400	4.0×10^{-12}
1500	8.0×10^{-12}
1600	1.5×10^{-11}

5.16 The diffusion coefficient of radioactive Cr^{3+} tracers in Cr_2O_3 is given in the table. Estimate the activation energy for diffusion.

T (°C)	D^* ($m^2\ s^{-1}$)
1050	1.0×10^{-15}
1100	4.6×10^{-15}
1200	1.05×10^{-14}
1300	6.2×10^{-14}
1400	2.7×10^{-13}
1500	6.5×10^{-13}

5.17 The impurity diffusion coefficient of Fe^{2+} impurities in a magnesium oxide single crystal is given in the table. Estimate the activation energy for diffusion.

T (°C)	D^* ($m^2\ s^{-1}$)
1150	2.0×10^{-14}
1200	3.2×10^{-14}
1250	5.0×10^{-14}
1300	7.5×10^{-14}
1350	1.0×10^{-13}

5.18 Calculate the diffusivity of ^{51}Cr in titanium metal at 1000 °C. $D_0 = 1 \times 10^{-7}$ m^2 s^{-1}, $E = 158$ kJ mol^{-1}.

5.19 Calculate the diffusivity of ^{51}Cr in a titanium –18 wt% Cr alloy at 1000 °C. $D_0 = 9 \times 10^{-2}$ m^2 s^{-1}, $E = 186$ kJ mol^{-1}.

5.20 Calculate the diffusivity of ^{55}Fe in forsterite, Mg_2SiO_4, at 1150 °C. $D_0 = 4.17 \times 10^{-10}$ m^2 s^{-1}, $E = 162.2$ kJ mol^{-1}.

5.21 Calculate the diffusivity of ^{18}O in Co_2SiO_4 at 1250 °C. $D_0 = 8.5 \times 10^{-3}$ m^2 s^{-1}, $E = 456$ kJ mol^{-1}.

5.22 Calculate the diffusivity of Li in quartz, SiO_2, parallel to the **c**-axis, at 500 °C. $D_0 = 6.9 \times 10^{-7}$ m^2 s^{-1}, $E = 85.7$ kJ mol^{-1}.

5.23 The diffusion coefficient of Ni^{2+} tracers in NiO is 1×10^{-15} m^2 s^{-1} at 1100 °C. Estimate the penetration depth of the radioactive Ni^{2+} ions into a crystal of NiO after heating for one hour at 1100 °C.

5.24 Ge is diffused into silica glass for fibre optic light guides. How long should a fibre of 0.1 mm diameter be annealed at 1000 °C to be sure that Ge has diffused into the centre of the fibre? The diffusion coefficient of Ge in SiO_2 glass is 1×10^{-11} m^2 s^{-1}.

5.25 What is the probability of a diffusing atom jumping from one site to another at 500° and 1000 °C if the activation energy for diffusion is 127 kJ mol^{-1}.

5.26 Estimate the ratio of the ionic conductivity to diffusion coefficient for a monovalent ion diffusing in an ionic solid. Take a typical value for the number of mobile diffusing ions as the number of vacancies present, approximately 10^{22} defects m^{-3}, and T as 1000 K.

5.27 The ionic conductivity of F^- ions in the superionic conductor $Pb_{0.9}In_{0.1}F_{2.1}$ at 423 K is 1×10^{-4} Ω^{-1} m^{-1}. The cubic unit cell (fluorite type) has a cell parameter 0.625 nm, and there

are on average, 0.4 mobile F^- anions per unit cell. Estimate the diffusion coefficient of F^- at 423 K.

5.28 The ionic conductivity of SrO (NaCl structure, a = 0.5160 nm), depends upon oxygen partial pressure. The value of the ionic conductivity is $2 \times 10^{-3}\ \Omega^{-1}\ m^{-1}$ at 900 °C under 0.1 atm O_2. Assuming the ionic conductivity is due to vacancy diffusion of Sr^{2+} ions estimate the diffusion coefficient of Sr^{2+} at 900 °C.

Note

1 To see just what sort of field strength this approximation corresponds to, take a value of $eaV/2kT$ equal to 1. Taking a temperature of 500 K, and a value of a of about 0.3×10^{-9} m yields a value for V of 2.87×10^4 V m^{-1}. Thus, the approximation is a reasonable for field strengths up to about 30 000 V m^{-1}.

6
Phase Transformations and Reactions

- What is sintering?
- What is a first-order phase transition?
- What is a martensitic transformation?

Many of the methods that are used to make a useful material are grounded in reactions that involve an alteration of chemical composition or a change in the microstructure/nanostructure of the solid. Phase diagrams can be used to determine which phases occur when the temperature, pressure, or composition of the system is specified and how they transform to other phases when the ambient conditions change significantly. Following the changes on a phase diagram also gives considerable insight into the microstructures that form during these changes. However, rapid changes in conditions, often met with in practice, may well result in the formation of metastable phases or microstructures that do not figure on any phase diagram, although these may be of great importance from the point of view of material properties.

This chapter is concerned with these sorts of transformations. It starts by examining conversions that involve little or no change of composition (Sections 6.1–6.5). This is followed by a discussion of a few of the consequences that composition change necessitates (Sections 6.6–6.8).

6.1 Sintering

6.1.1 Sintering and Reaction

Sintering is the process by which a compacted powder is converted into a solid body by heating below the melting point of the main constituents, so that the object essentially remains a solid throughout the process. There is (in principle) no composition change involved – all that happens is that the degree of agglomeration of the material alters. Sintering is widely used in the ceramics, powder metallurgy, and polymer industries to form small components such as valves, cutting tools, and sleeves.

In general, the procedure involves the careful preparation of powders with a suitable particle size and morphology, pressing these powders into the desired shape, the *green body*, followed by heating to sinter the powder particles together into a strong solid. Ideally, there should be little change of overall shape during sintering and the resulting object should be strong and pore free. Although the initial powder is frequently a single phase, small amounts of additives are often included to achieve strength and to eliminate porosity. This desirable result is sometimes obtained by an additive that melts below the sintering temperature to form a small amount of liquid between the grains of the major component; a process called *liquid-phase sintering*, widely used for the manufacture of cemented carbide tool tips, which consist of particles of a hard carbide such as tungsten carbide embedded in a matrix of softer metal, such as cobalt. Another variation is the technique of *reaction sintering*. This is exemplified by the production of silicon nitride objects. The desired shape is pressed from silicon powder and this is heated in an atmosphere of nitrogen gas. The silicon sinters and reacts simultaneously to form a compact silicon nitride part. A technique that is being widely used, especially for the production of complex shapes, is known as *selective laser sintering* (*SLS*), *3D-printing,* or *rapid prototyping*, described later (Section 6.1.4).

There are numerous methods of heating the initial pressed shape. The simplest is to place the object in an ordinary electrically heated furnace or kiln. Other methods include heating the object while subjecting it to high pressure (*hot pressing*) or heating it, either under pressure or not, by the passage of an electric current, provided the solid is a reasonable conductor.

Sintering is brought about by a variety of reactions. The first stage, which occurs at lower temperatures, as the body heats up, is an initial reduction in the surface roughness of the individual particles, followed by a period in which the particles start to join together (Figure 6.1). This often results in a coarsening of the porous grain structure rather than overall densification. At higher temperatures diffusion through the bulk and along grain boundaries contributes to the densification of the body by the elimination of internal pores and voids. In addition, material transport by viscous flow can be important in

glasses and evaporation and condensation are important in rather volatile compounds such halides. In practice, the mechanisms that operate during sintering have an influence upon the final shape of the sintered object. Sintering that takes place by vapour phase transport of material gives a solid with little shrinkage, while sintering by way of solid-state diffusion decreases the separation between the constituent particles, often leading to significant shrinkage (Figure 6.2). It is also necessary to remember that many of the overall properties of a sintered object, such as mechanical strength, and the optical, electrical, and magnetic properties, may differ from those of a bulk material. Sintering practice must take these end properties into account when the overall production process is designed.

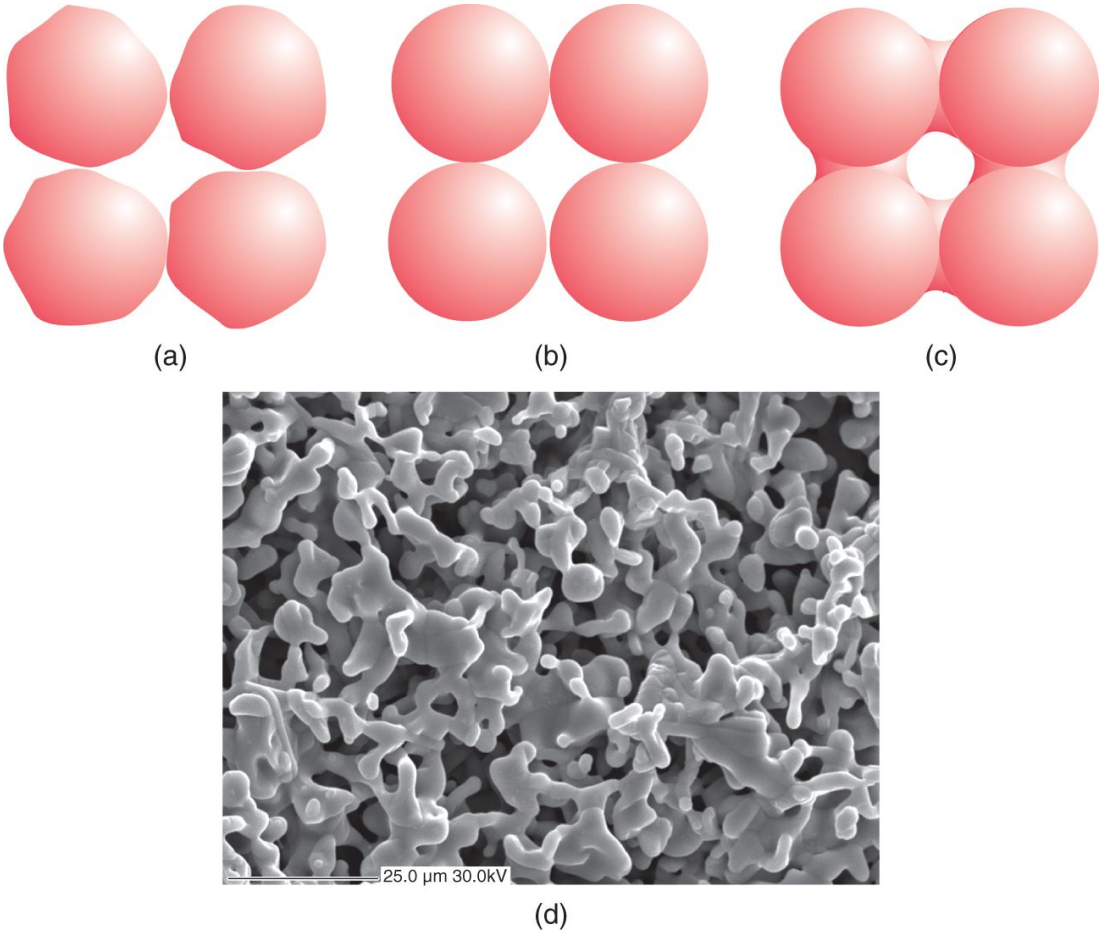


Figure 6.1 Initial stages in sintering: (a) Initial compact of slightly uneven particles; (b) decrease in surface roughness; (c) formation of a porous solid; (d) scanning electron micrograph of lightly sintered $Sr_2Ti_2O_7$ grains.

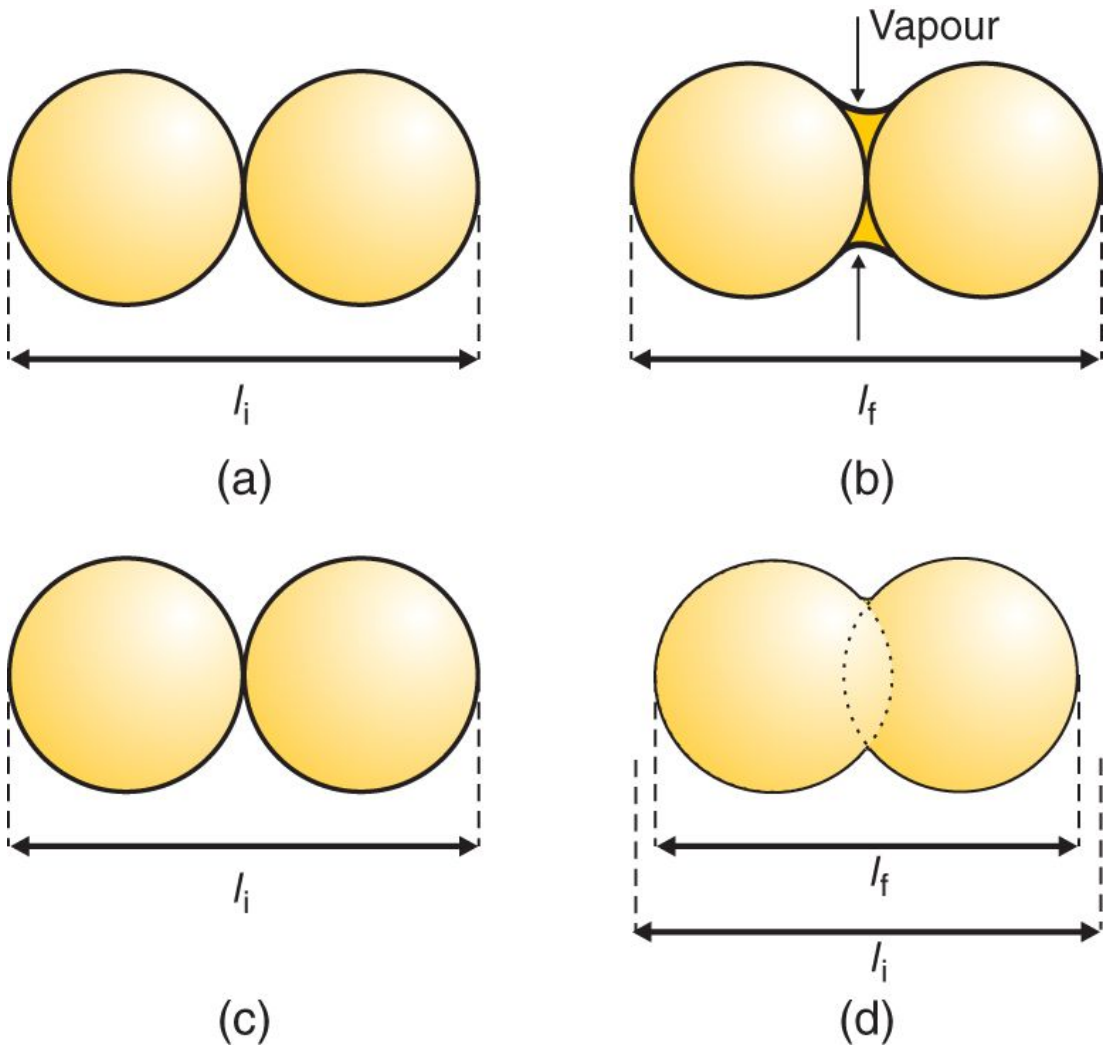


Figure 6.2 Sintering via vapour transport (a, b) does not cause shrinkage. Sintering via solid-state diffusion (c, d) causes shrinkage; l_i, l_f are initial and final lengths.

6.1.2 The Driving Force for Sintering

As sintering does not usually involve chemical reactions, the driving force is a reduction in the total surface area and the associated reduction in surface energy. This driving force can be illustrated for a flat surface that contains a spherical protuberance and a similar spherical depression, both of radius r (Figure 6.3a). The vapour pressure over a curved surface p is related to the vapour pressure over a flat surface p_0 by the *Kelvin equation*:

$$RT \ln[p/p_0] = 2V\gamma/r = \Delta G$$

where V is the molar volume of the substance, γ is the surface energy of the solid–vapour interface, and ΔG is the difference in Gibbs energy between a flat

surface and the curved surface. Thus, the vapour pressure over a protuberance will be greater than the vapour pressure over the flat surface and will increase as the radius of the curved surface decreases. In the case of a depression, the radius r is now negative and it is necessary to write:

$$RT\ \ln[p/p_0] = 2\,V\,\gamma/-r = \Delta G$$

Thus, the vapour pressure over the depression will be less than the vapour pressure over a flat surface. There will be a transfer of matter via the vapour phase from a protuberance to a depression and the surface will tend to become flat.

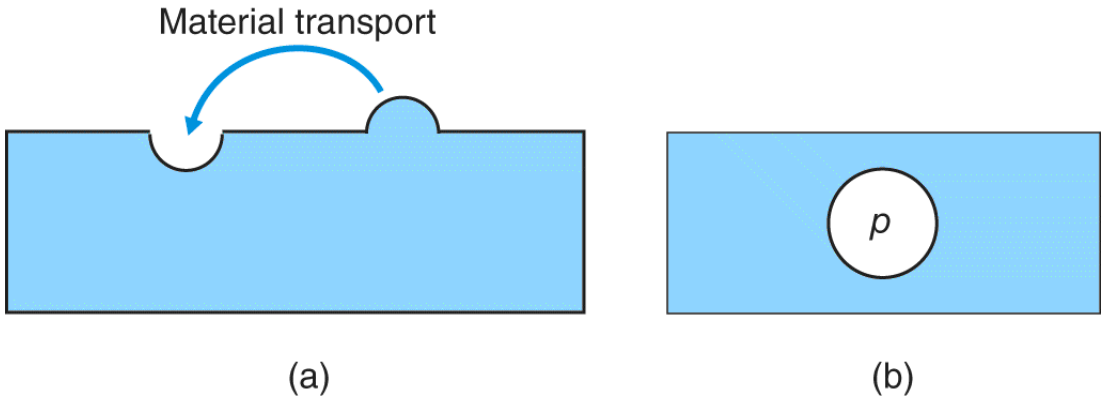


Figure 6.3 (a) Material transfers from hills to valleys during vapour-phase sintering. (b) The excess pressure, p, inside a void, balances the surface energy of the void.

Although it is important that articles fabricated by sintering do not contain large pores, because these lead to weakness and dimensional changes, the way in which powder granules link up during sintering will inevitably lead to pores forming unless the powders are carefully prepared. The size of the pores is also dependent upon the surface energy of the material. In the case of a spherical pore (Figure 6.3b), the excess pressure in the pore, p, needed to balance the surface energy is given by:

$$p = 2\gamma/r$$

where r is the radius of the sphere and γ is the surface energy. If the pressure is greater than that given, the pore will expand, while if it is smaller the pore will shrink, until equilibrium is reached. The ratio of the initial to final radii of the pore is given by:

$$\frac{r_i}{r_f} = \sqrt{\frac{2\gamma}{p_i r_i}}$$

where r_i is the initial radius of the pore, r_f the final radius, p_i is the initial pressure in the pore, and γ is the surface energy of the material. This equation indicates that small pores will tend to shrink and large pores will grow, with no change taking place when $r_i = r_f$, that is:

$$p_i r_i = 2\gamma$$

Material transport also occurs (for a similar reason) when two spheres touch to form a neck (Figure 6.2a). The spheres will have a positive and relatively large radius of curvature and the neck region a smaller negative radius of curvature. Matter will thus tend to be transported via the vapour phase from the larger spheres into the neck region causing the particles to join.

6.1.3 The Kinetics of Neck Growth and Grain Growth

The rate of flow of material into the neck region between two spheres will depend upon the mechanism of atomic transport. Four different mechanisms were investigated by Kuczynski, some 60 or more years ago. These were viscous flow, surface diffusion, bulk diffusion, and vapour transport by evaporation and condensation. The rate of neck growth was found to be quite different from one mechanism to another. They are usually expressed in terms of the ratio of the neck radius to the sphere radius by the equations:

Viscous flow: $x^2/r = k_1 t$

Surface diffusion: $x^5/r^2 = k_2 t$

Bulk diffusion: $x^3/r = k_3 t$

Vapour transport: $x^7/r^3 = k_4 t$

In these equations x and r are the neck and sphere radii (defined in Figure 6.4), k_1, k_2, k_3, and k_4 are constants that vary with temperature and t is the time of reaction.

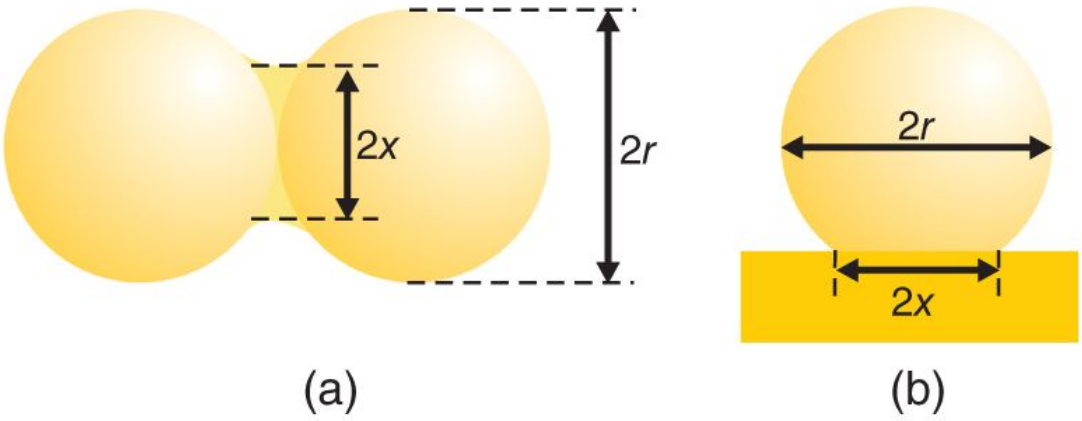


Figure 6.4 Definitions of (a) neck radius, x, and (b) sphere radius, r, used in the kinetics of sintering.

Experiments have shown that sodium chloride, NaCl, and other volatile materials sinter by a predominantly vapour transport mechanism while refractory oxides and metals sinter mainly by way of bulk diffusion.

Grain growth is often represented by an equation of the form:

$$(d^n - d_0^{\,n}) = k_0\, t\, \exp(-E/RT)$$

where d is the measured grain size, d_0 is the initial grain size, k_0 is an empirical constant, and t is the sintering time. The temperature factor is included as an Arrhenius term, where E is the activation energy for grain growth, R is the gas constant, and T the temperature (K). The value of the grain growth exponent, d, is usually taken to take a value of 2 if the grain growth is via sintering primarily involving solid-state diffusion at higher temperatures, whereas a value of 3 is used to describe sintering that occurs mainly via impurities and pores. This equation shows that the grain size is not a linear function of temperature, but an estimate of the activation energy for grain growth can be obtained over restricted temperature ranges by taking n to be 2 and plotting $\ln(d^2 - d_0^{\,2})$ vs. $1/T$.

6.1.4 Rapid Sintering

Traditionally sintering has been carried out in a furnace. For bulk applications, such a furnace, operating at typically 1200–1500 °C, will take several hours to reach the sintering temperature, many hours for the sintering reactions to occur and then several hours for the furnace to cool back to room temperature. In all a total sintering cycle can easily last 20–24 hours. This slow ramping up of the heating phase favours surface diffusion over bulk diffusion, so that neck growth and grain coarsening predominate. It is only when the temperatures reach the high sintering plateau that grain boundary and bulk diffusion,

together with densification dominate. While this is not a problem for many traditional ceramics, it does not produce satisfactory bodies for many applications. For example, in the fabrication of solid-state electrolytes used in sodium and lithium batteries, the long heating cycle causes considerable loss of the important and relatively volatile Na^+ and Li^+ species. To offset the disadvantages of long sintering cycles, a number of rapid sintering methods have been developed.

One of the first applications that made use of rapid sintering was the technique that has become known as *3D printing*, although also known as *additive manufacturing* and *rapid prototyping*. This method has been extensively used for the manufacture of complex ceramic and metal components. In this process, a thin layer of a finely divided metal or ceramic powder, either single component, coated or mixed, is laid down. The powder bed is then heated to a point close to but below the melting point of the ceramic or metal phases involved, in a very finely defined pattern, via a high-power laser controlled by a computer. In this procedure (unlike in classical sintering), some the particles may melt on the surface and fuse, being pulled together by surface tension. The relative extent of surface fusion as opposed to solid-state sintering depends upon the precise details of the materials used and the requirements of the finished component. Once one layer is treated, a new thin bed of powder is laid down on top of the first and the whole process repeated. In this way, the complex body is built up in a series of stacked solidified layers. The powder that is not laser treated remains in place to support the succeeding layers and is eventually removed at the end of the fabrication. In this way complicated three-dimensional metallic or ceramic shapes can be rapidly fabricated.

Although these 3D printing technologies have been in place for some time, the actual mechanism of the sintering/fusion reactions that take place in the laser-irradiated powder bed is still not known in detail. The laser power, the laser pulse duration, the beam shape and the scanning velocity all influence the nature of the product: whether, for example, it consists of small crystallites, larger crystals, and crystal growth habit can all influence the durability of the finished article. In addition, liquid droplets that are ejected from the melting powder bed under intense irradiation, called *scatter*, and imaged as sparks jumping from the laser beam track, can fall back onto the powder bed, resulting in irregularities and the formation of defects such as voids. These can all impair the ultimate performance of the finished piece.

Another recent technique sandwiches the green body between thin sheets of carbon. By passing a current through the carbon layers, the temperature of the ceramic precursor can be raised to approximately 3000 °C in only 30 seconds, thus avoiding the extensive neck growth and coarsening described earlier. Heating for approximately 10 seconds and then rapidly cooling the sample, again in just a few seconds, results in a fully formed sintered body, which has

retained the composition and grain size distribution of the starting material. Such a technique has been used successfully to sinter many ceramics, including alumina, stabilised zirconia, and in particular Li and Na containing ceramic solid-state electrolytes.

6.2 Phase Transitions

Phase transitions, which, in this chapter, are taken to mean changes at more or less constant chemical composition, generally take place at a definite temperature, the *transition temperature*, and at a definite pressure, the *transition pressure*. They are familiar in everyday life – ice transforms to liquid and then to vapour as the temperature rises. Here the focus is upon phase transformations that are confined to the solid state only. Broadly speaking there are two ways of classifying these solid-state transitions, thermodynamic and structural.

From a thermodynamic perspective, a phase transition will occur if the change results in a lowering of the Gibbs energy of the system. At the exact transition temperature, the Gibbs energy of both phases are equal but the way in which the free energy of the system varies as the transition point is traversed differs from one system to another. Ehrenfest used the nature of this variation to classify phase transitions as one of two types, *first-order transitions* and *second-order transitions*.

6.2.1 First-Order Phase Transitions

First-order phase transitions, which are also called *discontinuous* transitions, are those in which the molar Gibbs energy, g, of the system changes slope at the transition temperature, and the first derivative of g with respect to temperature, T $(\partial g/\partial T)_P$, and pressure, P $(\partial g/\partial P)_T$, are discontinuous (Figure 6.5). These derivatives simply measure the slope of the Gibbs energy curve and correspond to other thermodynamic variables. The slope $(\partial g/\partial T)_P$ is the molar entropy of the phase s_m and the difference in the differentials at the transition point is the molar entropy change of the transition, Δs_m:

$$(\partial g/\partial T)_P^{\text{high}} - (\partial g/\partial T)_P^{\text{low}} = s_\mathrm{m}^{\text{high}} - s_\mathrm{m}^{\text{low}} = \Delta s_\mathrm{m}$$

The slope $(\partial g/\partial P)_T$ of the g vs. P curve is the molar volume of the phase, v_m, and the difference in the slopes at the transition point is simply the change in molar volume occurring at the transition, Δv_m:

$$(\partial g/\partial P)_T^{\text{high}} - (\partial g/\partial P)_T^{\text{low}} = v_\mathrm{m}^{\text{high}} - v_\mathrm{m}^{\text{low}} = \Delta v_\mathrm{m}$$

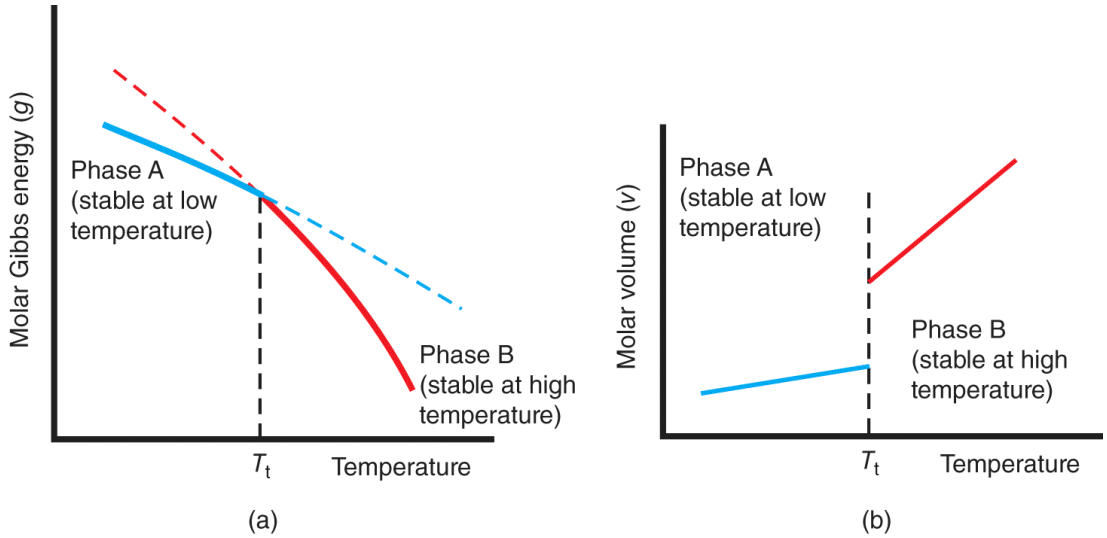


Figure 6.5 First-order transitions: (a) variation of molar Gibbs energy with temperature; (b) variation of molar volume with temperature.

It can be remarked that the volume change is invariably negative, in that the high-pressure variant has the lower molar volume. First order transitions are also characterised by an enthalpy change usually called the *latent heat* of transition and by a specific heat that is notionally infinite at the transition temperature, because all heat input is used to transform one phase to the other and none is used to raise the temperature of the body. Examples of first order solid-state transitions are the transitions between the NaCl and CsCl structure types. Rock salt itself transforms to the CsCl structure at a pressure of 298 kbar and a temperature of 298 K, and CsCl transforms to the NaCl structure at 745 K and 1 atm pressure. Similar transitions occur in many MX halides and oxides.

6.2.2 Second-Order Transitions

Second-order transitions, also called *continuous transitions*, are ones in which the Gibbs energy of the system does not change slope at the transition point and are defined by the facts that although the first derivation of the molar Gibbs energy with respect to temperature and pressure are continuous at the transition point, the second derivatives $(\partial^2 g/\partial T^2)_P$ and $(\partial^2 g/\partial P^2)_T$ are discontinuous (Figure 6.6). Thus, the molar entropy and molar volume do not change abruptly at the transition. The second derivatives measure the slopes of the $(\partial g/\partial T)_P$ and $(\partial g/\partial P)_T$ curves, and once again can be related to other thermodynamic properties. The derivative $(\partial^2 g/\partial T^2)_P$ is equal to the specific heat term $-c_P/T$, where c_P is the molar specific heat. The derivative $(\partial^2 g/\partial P^2)_T$ is equal to the volume term $-\upsilon\beta$, where β is the compressibility. Second-order

transitions include ferroelectric to paraelectric transformations (Section 9.3), ferromagnetic to paramagnetic transformations (Section 10.3), or superconducting to metallic or semiconducting transformations (Section 11.5).

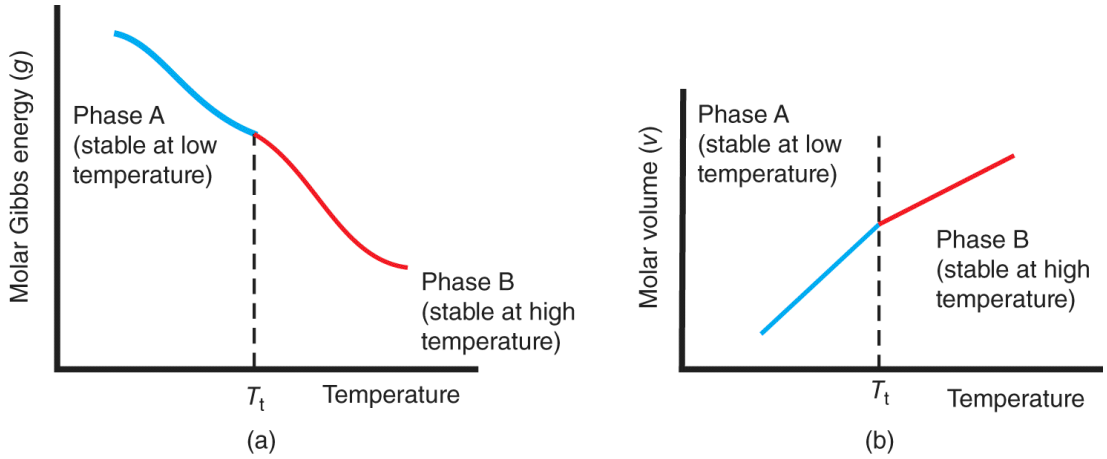


Figure 6.6 Second-order transitions: (a) variation of molar Gibbs energy with temperature; (b) variation of molar volume with temperature.

Note that it is not so easy to distinguish between first-order and second-order transformations in practice, and ideal sharp transitions are not always the rule. In such cases a clear statement of order cannot always be categorically made (see the notes on C_{60}, Section 6.4.2).

6.3 Displacive and Reconstructive Transitions

6.3.1 Displacive Transitions

In crystallographic terms it is convenient to consider a phase transformation as a structural change without any significant change in chemical composition. Displacive transitions (sometimes called *topological* transitions) are defined as changes in which one or more atoms in a structure changes its bonding slightly so as to modify bond lengths and bond directions without bond breaking (Figure 6.7). The transformations typically involve small, coordinated movements of atoms, and are usually rapid. The crystal structure of the product phase is closely related to that of the reactant. They are often triggered by temperature and are usually reversible, so that the high temperature form cannot be retained to room temperature by quenching (rapid cooling).

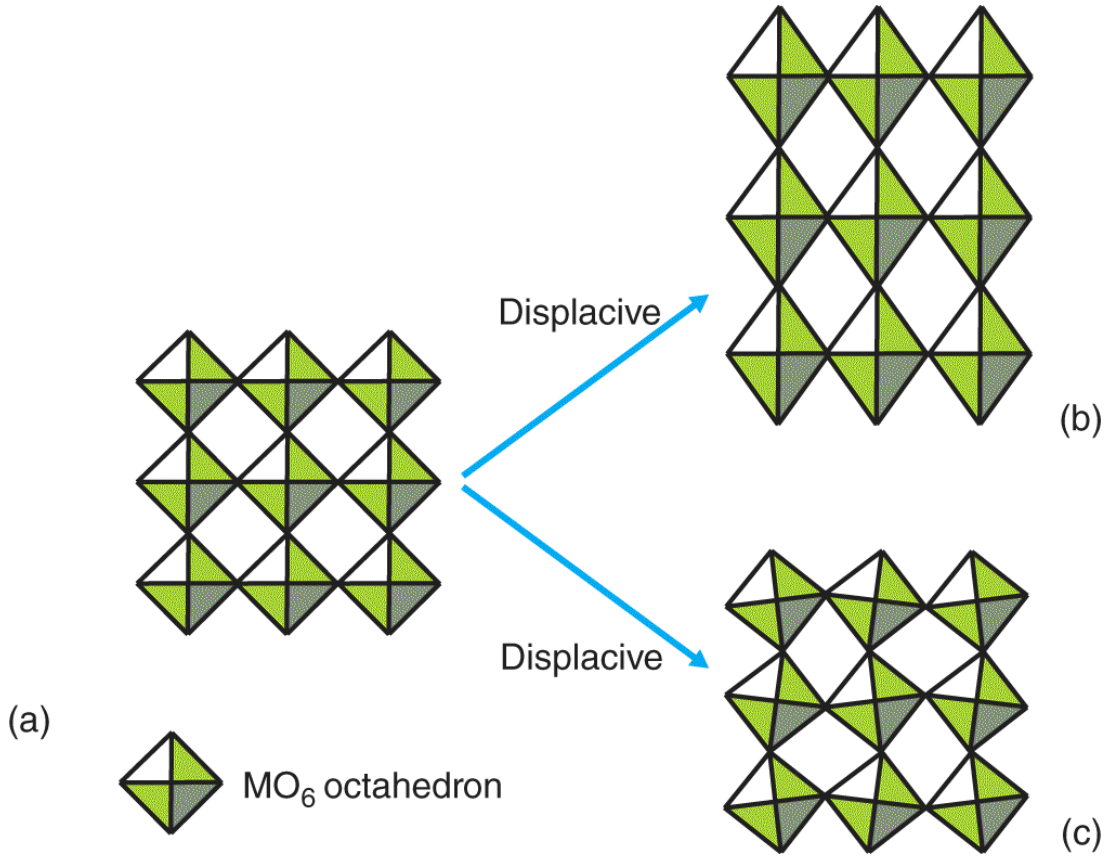


Figure 6.7 Displacive transitions, schematic: (a) undistorted octahedra; (b) octahedra lengthened axially; (c) octahedra rotated.

Displacive transitions are exemplified by bond changes in the oxide lanthanum cuprate, La_2CuO_4, which is built from sheets of CuO_6 octahedra (Figure 6.8a). At room temperature these octahedra are distorted by virtue of the *Jahn–Teller effect*, which indicates that perfect octahedral coordination of Cu^{2+} ions

is unstable with respect to a distorted octahedron. The distortion takes the form of two long (0.240 nm) bonds parallel to the unit cell **c**-axis and four short (0.189 nm) bonds parallel to the **a**- and **b**-axis (Figure 6.8b), resulting in an orthorhombic unit cell. As the temperature increases the distortion effectively decreases and at a temperature of 523 K the octahedra become regular.

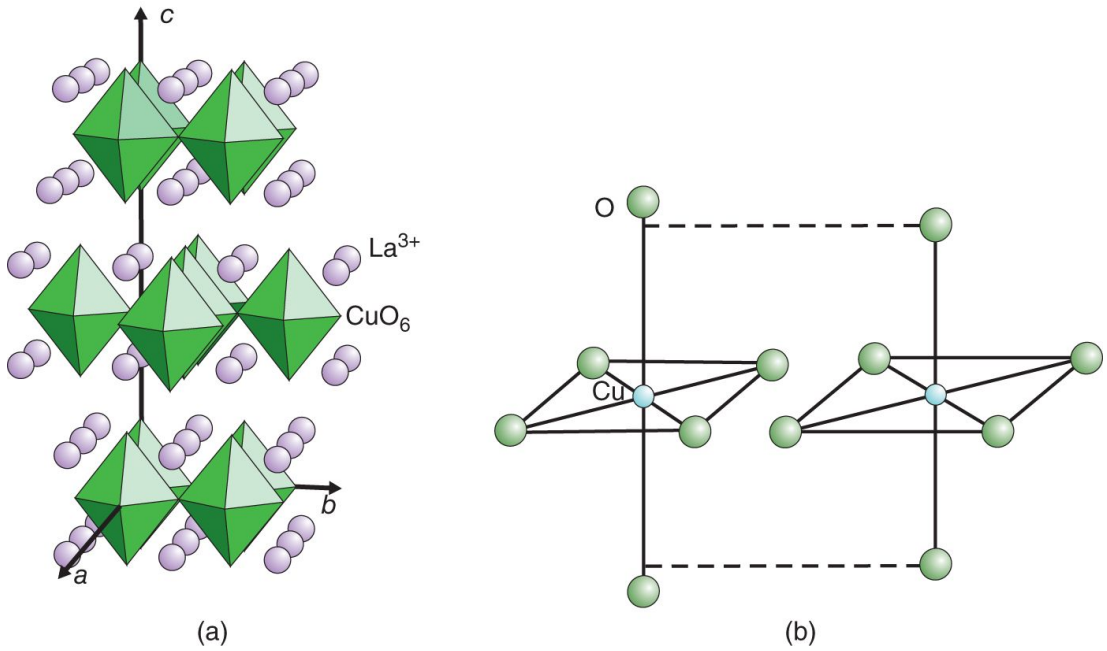


Figure 6.8 Displacive transition in La_2CuO_4: (a) The idealised structure of La_2CuO_4; (b) elongated CuO_6 octahedron below 523 K compared to a regular octahedron above that temperature.

Displacive transformations involving octahedral rotation have been studied extensively in the perovskite-structure oxides. An example is provided by $CaTiO_3$. At room temperature the structure is orthorhombic with 'rotated' and slightly irregular TiO_6 octahedra (Figure 6.9a). The phase transforms by a discontinuous (first order) transition to tetragonal at 1498 K and then to cubic (Figure 6.9b) via a continuous (second order) transition at 1634 K. This feature draws attention to the fact that displacive transitions can be first *or* second order in terms of thermodynamic classification.

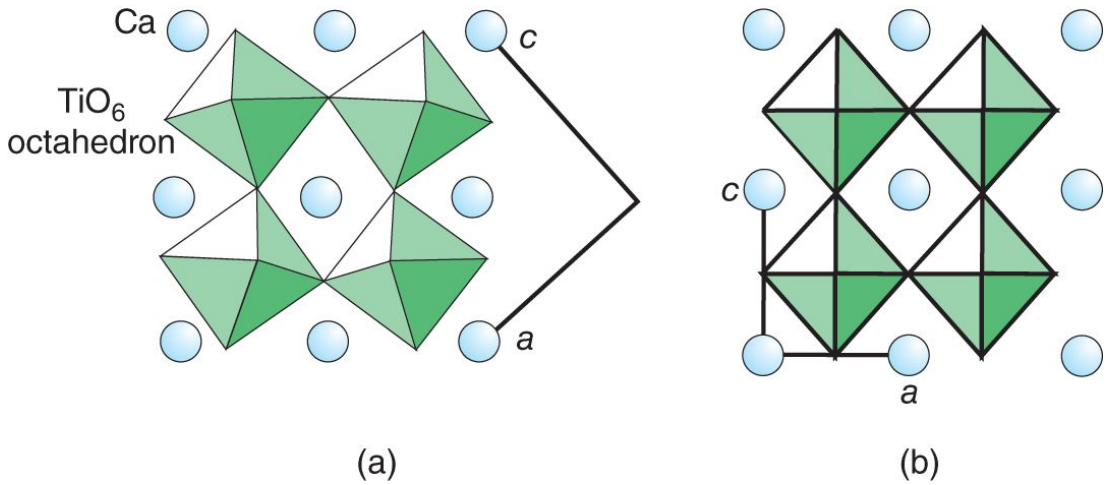


Figure 6.9 Displacive transition in $CaTiO_3$: (a) room temperature structure; (b) high temperature cubic form.

A well-documented example of a displacive transition is the transformation of the low temperature form of quartz to the high temperature form, in which both distortions and rotations of the SiO_4 tetrahedra alter but the tetrahedra themselves remain intact. Other examples include important transitions in ferroelectric ceramics that include hydrogen bonded materials in which the hydrogen in the bond is off-set to one side of centre at low temperatures, but at higher temperatures the hydrogen is found to occupy, at least statistically, a central position and the perovskite oxide $BaTiO_3$, in which distortions of TiO_6 octahedra and displacement of the central Ti^{4+} cations give rise to the ferroelectric effect.

6.3.2 Reconstructive Transitions

Reconstructive transitions describe the situation where a major reorganisation of a crystal structure occurs during which primary bonds are broken and reformed, again, with no significant change in composition (Figure 6.10). Single crystals undergoing a reconstructive transition usually fragment, and the reactions are usually rather slow. There is no *necessary* relationship between the unit cell of the parent form and that of the product, but as the composition of the two phases remains the same, crystallographic relationships between the two structures may be apparent. Although there is no coincidence between the thermodynamic definitions of phase transformation and displacive transitions, *all* reconstructive transitions are first-order thermodynamically and usually involve considerable latent heat changes.

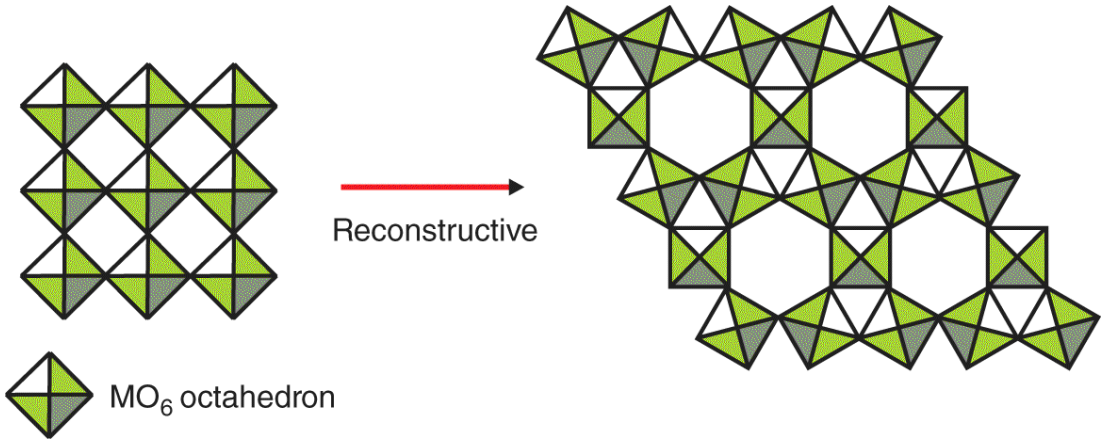


Figure 6.10 A reconstructive transition, in which components are rearranged without change of composition.

The transformation of minerals with the olivine structure to the spinel structure under the high pressures that occurs deeper within the mantle is of importance to studies of the Earth. Olivine itself is a mineral that is a solid solution between the end members Mg_2SiO_4, forsterite, and Fe_2SiO_4, fayalite, with the general composition $(Mg, Fe)_2SiO_4$. It is widespread in the Earth's upper mantle. The oxygen array in olivine is close to hexagonal (...ABAB...) closest packing and the cations are divided between octahedral and tetrahedral interstices, with the (Mg, Fe) pair occupying half of the available octahedral sites and the Si occupying ⅛ of the available tetrahedral sites in an ordered fashion. Spinel has the formula $MgAl_2O_4$ and the spinel structure is the cubic analogue of the olivine structure (Figure 3.12). The oxygen array in spinel is close to cubic (...ABCABC...) closest packing and the cations are again divided between octahedral and tetrahedral interstices, with the Mg^{2+} ions occupying ⅛ of the available tetrahedral sites and the Al^{3+} ions occupying half of the available octahedral sites in an ordered fashion. Many minerals that adopt the olivine structure at ordinary pressures (for instance Mg_2SiO_4, Fe_2SiO_4, Ni_2SiO_4, Al_2BeO_4, Mg_2GeO_4) transform to spinel structure analogues when subjected to high pressures.

This transformation is undoubtedly reconstructive in nature, but there is uncertainty about the mechanism of the transformation. Two principal ways of going from one structure to the other have been suggested. The first of these supposes that a minute volume (a nucleus) of the spinel structure variant forms at a suitable place in the olivine matrix and then subsequently grows into a significant crystallite. This would produce a new phase with to no correspondence between the crystal structures of the parent and product material. An alternative mechanism involves converting the (...AB...) packing in olivine into (...ABC...) packing by a shearing of the structure followed by subsequent diffusion of the cations into new sites (Figure 6.11). In this case the

oxygen planes in the two structures would remain parallel and (001) in olivine becomes (111) in the spinel so that there is a strong geometrical relationship between the two structures although the transition is reconstructive. Reactions in which the crystallographic orientation of the reacting phase is preserved into the product phase are called *topotactic* reactions. If this holds in the olivine–spinel transformation, it would be expected that lamellae of spinel would fit coherently into the olivine matrix and partly transformed crystals should show this signature. Despite this distinction there is still considerable uncertainty about the transformation mechanism. For example, the olivine to spinel transformation in Mg_2GeO_4 has been described as occurring by both nucleation and shear. It may be that the mechanism of transformation is not absolutely fixed but changes in response to local conditions, including the nature of any impurities and defects present in the sample.

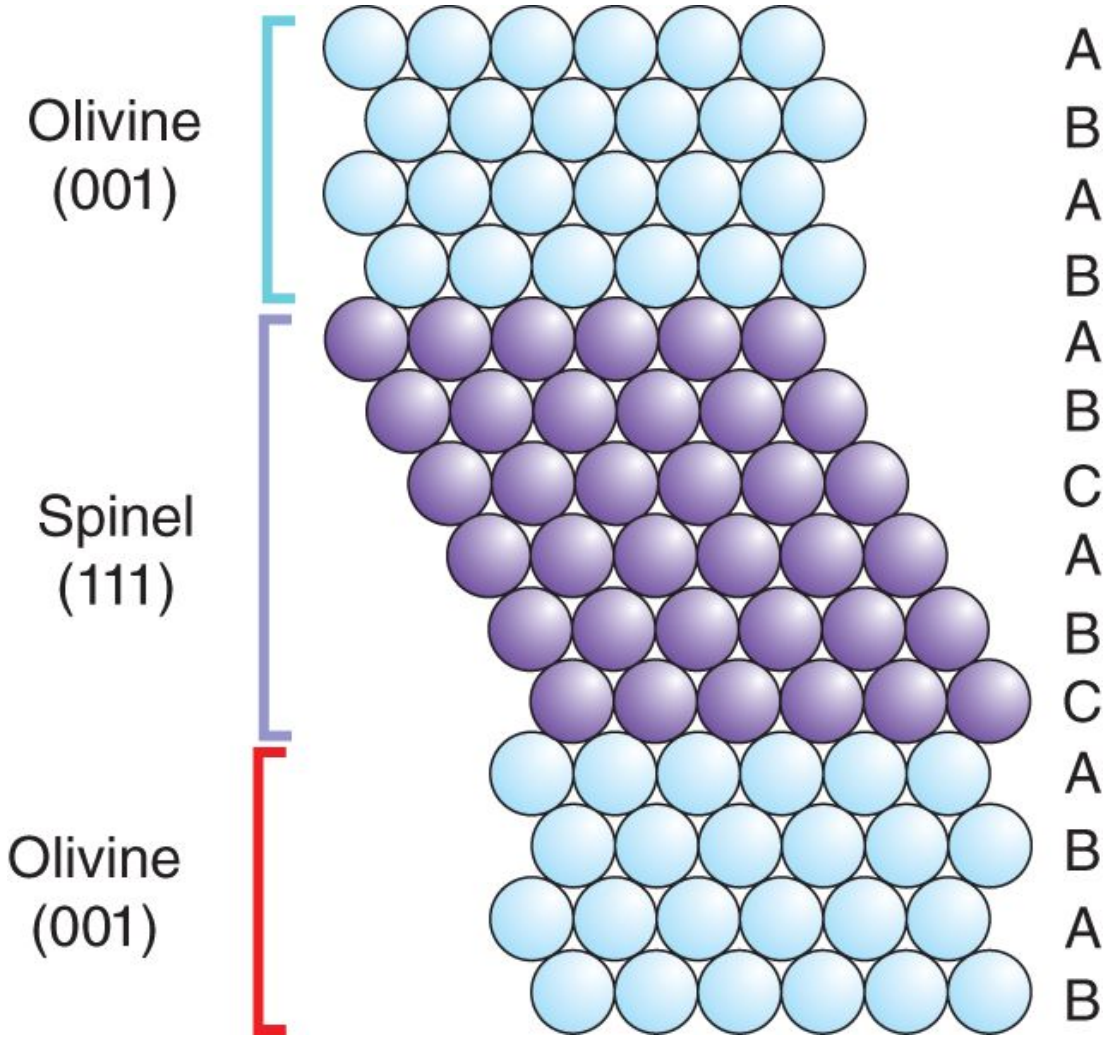


Figure 6.11 A topotactic reconstructive transformation in which the …ABAB… packing of O planes in olivine shear to produce …ABCABC… packing in spinel.

6.4 Order–Disorder Transitions

An *order–disorder transition* is a transition in which the degree of order in a phase changes. Such transitions occur frequently in both metal alloys and ceramics. The degree of order in a phase can be quantified *by an order parameter* that is a number usually set at 0 for total disorder and 1 for perfect order. There are three main types of order-disorder transition: positional ordering/disordering; orientational disordering; and ordering of electronic or nuclear spin states, giving rise to magnetism, described in Chapter 10.

6.4.1 Positional Ordering

If two or more types of atom occupy the same sites in a crystal then ordering can take place. Positional order–disorder is legion in minerals because geological timescales provide scope for long annealing at moderate temperatures. Positional ordering is also widespread in alloys between similar elements if these are annealed at moderate temperatures, and has been described earlier with respect to the gold–copper alloys (Figure 4.3). Similar order–disorder occurs in many non-metallic compounds. For example, the oxide $LiFeO_2$, prepared at high temperatures, adopts the NaCl structure in which the cations are distributed at random over the available octahedral sites. On annealing this phase at lower temperatures, the cations order, leads to a tetragonal unit cell. A similar transformation has been recorded with the phase Li_2NiTiO_4. At high temperatures this phase adopts a disordered NaCl structure in which layers of cations fill all the octahedral interstices between every (111) plane of oxygen ions at random. The ordered structure forms below approximately 550 °C. This phase has a monoclinic unit cell in which the cation layers now alternate between Li-rich and Li-poor sheets lying between the same oxygen ion layers.

A more complex example is given by a number of oxide spinels, which have a general formula of AB_2O_4. Here cations occupy tetrahedral ( ) or octahedral [ ] sites. There are two principal arrangements of cations, $(A)[B_2]O_4$, the *normal spinel structure*, and $(B)[AB]O_4$, *inverse spinel structure*. In reality, very few spinels have exactly the normal or inverse structure and the cation distribution between the sites is described by an order parameter (the *occupation factor*), λ, which gives the fraction of B^{3+} cations in tetrahedral positions.

In many compounds it is convenient to think of defect ordering rather than atomic or ionic ordering. The titanium sulfides related to TiS_2 can be considered to contain Ti interstitials. The structure of these phases consists of sheets of composition TiS_2 (Figure 6.12). If the octahedral sites between these sheets are completely filled with Ti atoms, the compound corresponds to TiS

(Ti_2S_2), while if they are empty the composition is TiS_2. In between these end points any composition Ti_xS_2 can be synthesised with x taking values from 1 (TiS_2) to 2 (TiS). At high temperatures these additional Ti atoms are disordered over the available empty octahedral sites. On the other hand, if the samples are annealed at low temperatures, many variants occur in which interstitial Ti atoms order over the octahedral sites. For example, the nominal phase Ti_2S_3 has been shown to disproportionate into many ordered structures including $Ti_{1.156}S_2$, $Ti_{1.170}S_2$, and $Ti_{1.198}S_2$ when annealed at lower temperatures.

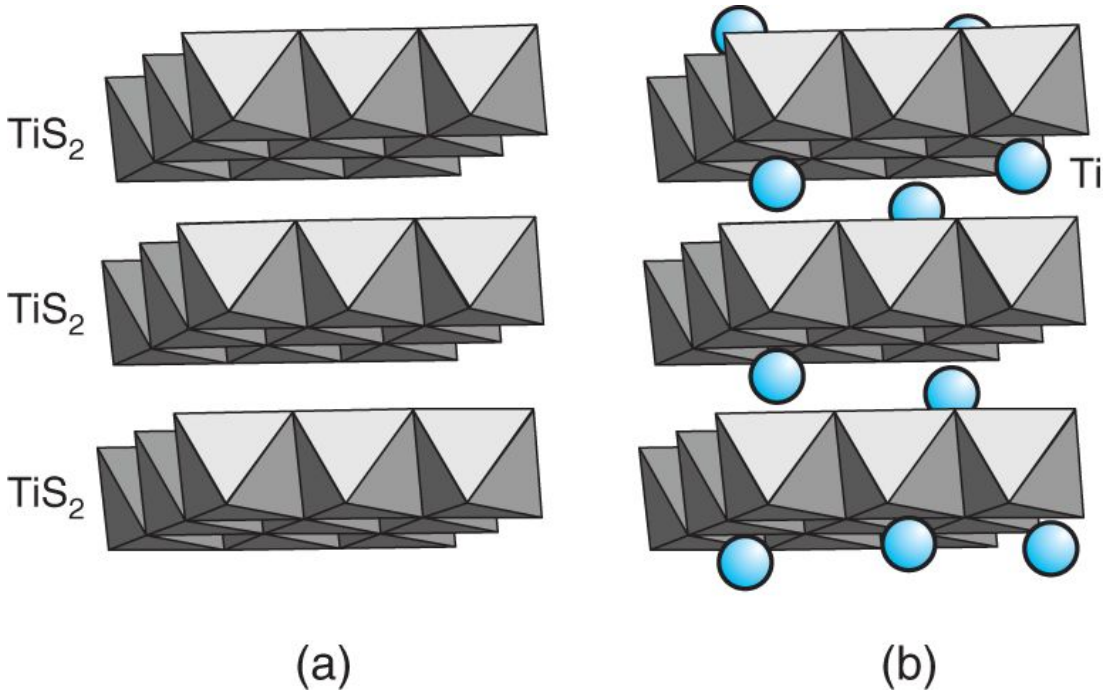


Figure 6.12 The idealised structure of (a) TiS_2; (b) phases with compositions between TiS_2 and TiS.

Non-stoichiometric oxides with the fluorite structure are often characterised by populations of ordered oxygen vacancies. These latter phases are typified by 'PrO_2'. When prepared at about 930 °C the composition of this oxide is close to $PrO_{1.833}$ and the phase supports a broad range of composition. If the homogeneous oxide is annealed at lower temperatures, it disproportionates into a series of phases given by the formula Pr_nO_{2n-2} (n = 7, 9, 10, 11, 12), all of which can be described in terms of the parent fluorite structure containing ordered oxygen vacancies.

6.4.2 Orientational Ordering

Orientational order–disorder frequently occurs in phases containing groups such as the angular $NH_4^+, NO_3^{-3}, NO_2^-$ or linear H–Cl or CN^-. At the lower temperatures the groups are in a fixed orientation. As the temperature increases orientational disorder can occur in a number of ways and several distinct phases and orientational phase transitions may occur before a high temperature form is reached. These variations often endow the various structures with ferroelectric properties. Orientational disorder takes the form of random orientation in the case of angular groups. For example, in $NaNO_2$, the NO_2^- groups align parallel to the +**b**-axis at low temperatures but orient randomly in either the +**b** or –**b** direction at high temperatures. In the case of the ammonium halides related to NH_4Cl, it was long supposed that the NH_4^+ groups rotated freely at high temperatures. This is not the case and the groups take on two alternative orientations, at random, in the structure (Figure 6.13a,b). Linear groups, such as polar HX molecules in the halogen halides, also tend to adopt a single direction at lower temperatures and random directions, constrained only by the crystal structure, at higher temperatures (Figure 6.13c,d). A similar orientational order–disorder transformation is seen in compounds allied to KCN. In this phase the CN^- groups are ordered at lower temperatures and give rise to an orthorhombic unit cell. At higher temperatures the CN^- groups are directed at random along any of the possible [111] directions, resulting in a cubic unit cell analogous to NaCl.

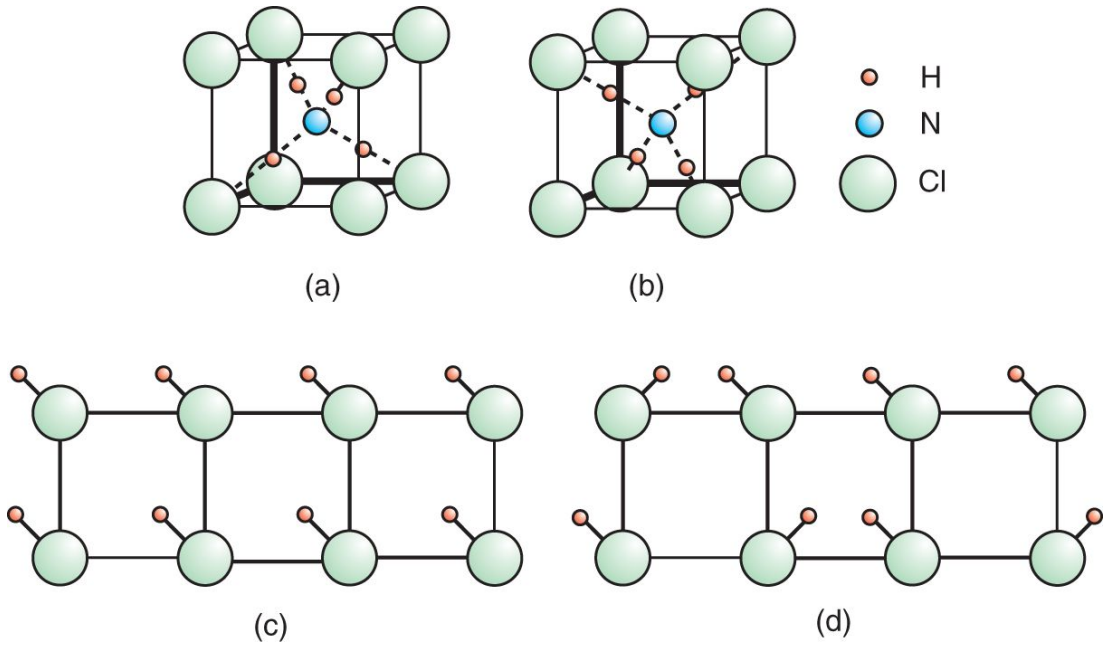


Figure 6.13 Orientational transformations: (a, b) alternative orientations of NH_4^+ tetrahedra in ammonium halides; (c, d) alternative orientation of linear HCl molecules in hydrogen halides.

The carbon allotrope C_{60} (Buckminsterfullerene), which is of interest because of its electronic properties, also shows an orientational order–disorder transition. The molecule C_{60} has a diameter of 0.071 nm diameter. The C atoms are situated at the vertices of a truncated icosahedron, to produce a shape that resembles a soccer ball (Figure 2.6a). At room temperature these molecules pack together to form a crystalline solid with a face-centred cubic unit cell and a C_{60} unit at each lattice point. In this phase the molecules show free rotation. At temperatures below approximately −20 °C, the free rotation is lost. The C_{60} units become locked in place and the unit cell changes to simple cubic. The change is an order–disorder transition that has been described as ‘weakly first order’.

6.5 Martensitic Transformations

Martensitic transformations are essentially solid-state displacive transitions that take place at a constant temperature and composition due to synchronised atomic displacements over distances smaller than the normal interatomic distances in the parent phase. The name is derived from a characteristic microstructure of hard steels called *martensite*, first identified optically by the metallurgist Martens. Martensitic transitions take place very rapidly, because atomic diffusion does *not* occur and for this reason are frequently called *diffusionless* transitions.

Although the term martensitic transition is now applied to large numbers of diffusionless transitions in metal and ceramic systems, it is useful to start with steels in this context as the characteristics of the change in steels is similar to that in other systems.

6.5.1 The Austenite–Martensite Transition

Rapidly cooling the homogeneous iron-carbon alloy austenite (the γ phase, face-centred cubic structure, Figure 2.18), to room temperature (a process called *quenching*), leads to the formation of *metastable* martensite, which is termed the α′ phase. This has a tetragonal structure that is a slightly distorted body-centred cubic form of the ferrite α phase. The lattice parameters of the α′ phase vary with carbon content, a_t taking a value of approximately 0.285 nm and c_t taking values between 0.292 and 0.300 nm. The ratio of c_t/a_t increases linearly with carbon content to a maximum of 1.08 in the highest carbon steels. The formation of martensite is extremely rapid because only a slight displacement of atoms takes place. In particular, the carbon atoms have no time to diffuse any great distance. These atoms, normally present up to about 1.5 wt%, are situated in octahedral sites in austenite and this does not change on transformation so that they remain in octahedral sites in body-centred martensite. Once a nucleus of a martensite crystal forms, it will grow to its final size in a time of the order of 10^{-7} seconds. This is as fast as the speed of sound in iron. The rate of formation seems to be constant down to liquid helium temperatures.

Despite this rapidity, the transition rarely goes to completion and transformed steels consist of grains, which are mixtures of austenite and martensite. The resultant solid is made up of a set of interlocking martensite platelets with the **c**-axes aligned at random along the original austenite cubic axes, surrounded by an amount of untransformed austenite, which makes martensite both the hardest and the most brittle constituent of quenched steels. This effect is easily demonstrated. Heat an initially flexible steel piano wire to redness and then rapidly cool it by plunging it into cold water. The cold wire will be brittle and can be snapped by hand. The broken ends will scratch glass.

The simplest description of the crystallographic relationship between the two structures is given by the *Bain model*. The face-centred cubic γ phase, austenite, lattice parameter a_c, can be represented in terms of a body-centred tetragonal unit cell with $a_t = a_c/\sqrt{2}$; $c_t = a_c$ (Figure 6.14). To convert this prototype martensite unit cell into the true body-centred cubic unit cell of martensite, a_t must increase by about 22% at the same time as c_t shrinks by about 14% to produce an ideal lattice parameter of about 0.286 nm. If this transformation is correct, the following orientational relationships hold between face-centred cubic γ, austenite and body-centred cubic α′, martensite:

$$(111)_\gamma \rightarrow (011)_{\alpha'}$$

$$[110]_\gamma \rightarrow [100]_{\alpha'}$$

$$[101]_\gamma \rightarrow [111]_{\alpha'}$$

$$[112]_\gamma \rightarrow [011]_{\alpha'}$$

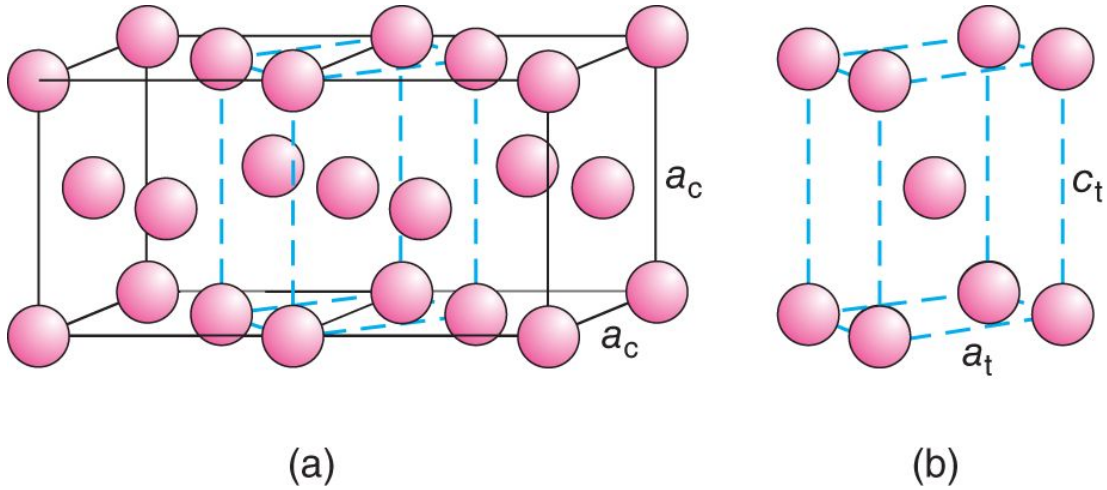


Figure 6.14 (a) Two unit cells of the cubic face-centred cubic structure, the alternative body-centred tetragonal unit cell is outlined and drawn separately in (b).

Experimentally the orientational relationships given in the Bain model are not fully consistent with experimental observations but are a starting point for understanding the transformation.

The characteristic microstructure of the martensite that forms is composition dependent. Below about 0.6 wt% carbon the martensite forms in long blades and is called *lath martensite*. At compositions above about 1.0 wt% carbon, the form is more lens shaped in form, and is called *plate martensite*. At

compositions between 0.6 and 1.0 wt% carbon, a mixture of the two forms is found. During the transformation there is no preference as to which of the cubic axes elongates and which shortens during martensite formation, so that the **c**-axis of the martensite can lie along any of the original cubic axes (Figure 6.15) and usually several different *variants* or *domains* (i.e. orientations) occurs in an austenite grain.

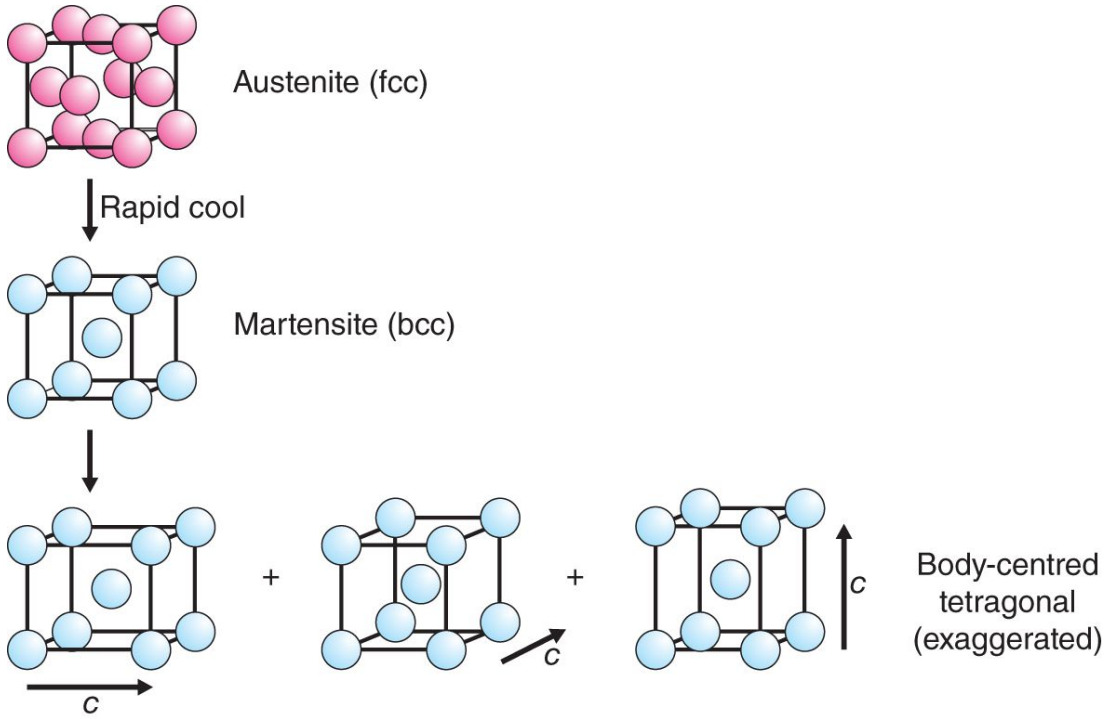


Figure 6.15 Schematic crystallographic changes occurring during the formation of martensite from austenite.

The production of each martensite platelet involves a shear parallel to the plate and usually an expansion normal to it. These distortions cause considerable strain in the two-phase matrix that contributes to the fact that the transition rarely goes to completion. It also has consequences for the microstructure of the martensite itself. An unconstrained transformation will simply deform the overall structure of the sample (Figure 6.16a). The platelets of martensite are, however, closely constrained by the surrounding austenite matrix. The first consequence of this is that plate martensite takes on the characteristic lens shape (Figure 6.16b). In addition, the platelets show extensive slip lines and large numbers of closely spaced twin planes. These features arise because the deformation of the platelets (Figure 6.17a,b) must be accommodated within the surrounding relatively inflexible matrix at as low an energetic cost as possible. The two simplest ways of achieving this are by either repeated slip, which may be brought about by the passage of dislocations through the crystal (Figure 6.17c), or by *polysynthetic twinning*, the formation of many narrow

twin slabs (Figure 6.17d). These alternatives and the underlying transition mechanism are still under active study.

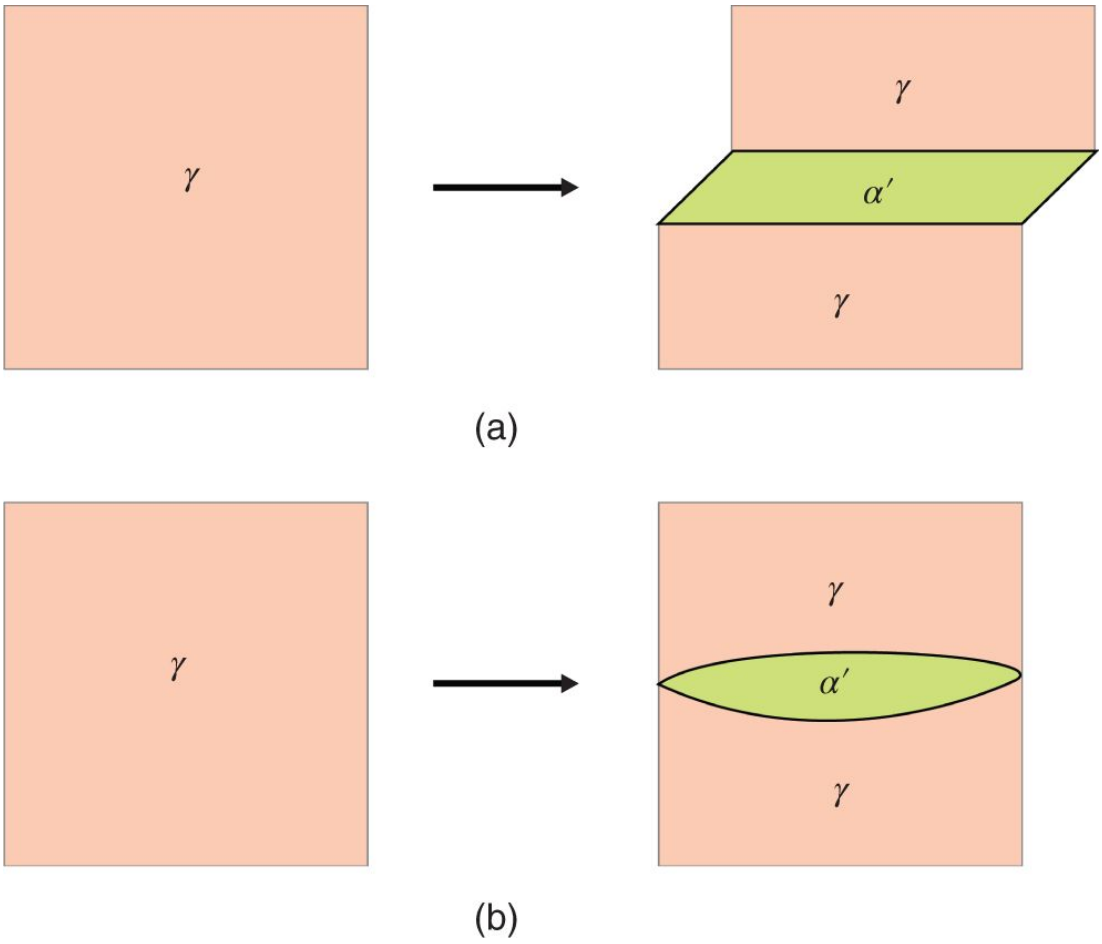


Figure 6.16 (a) The unconstrained transformation of austenite (γ phase) to martensite (α′ phase); (b) the constrained transformation produces lens-shaped volumes.

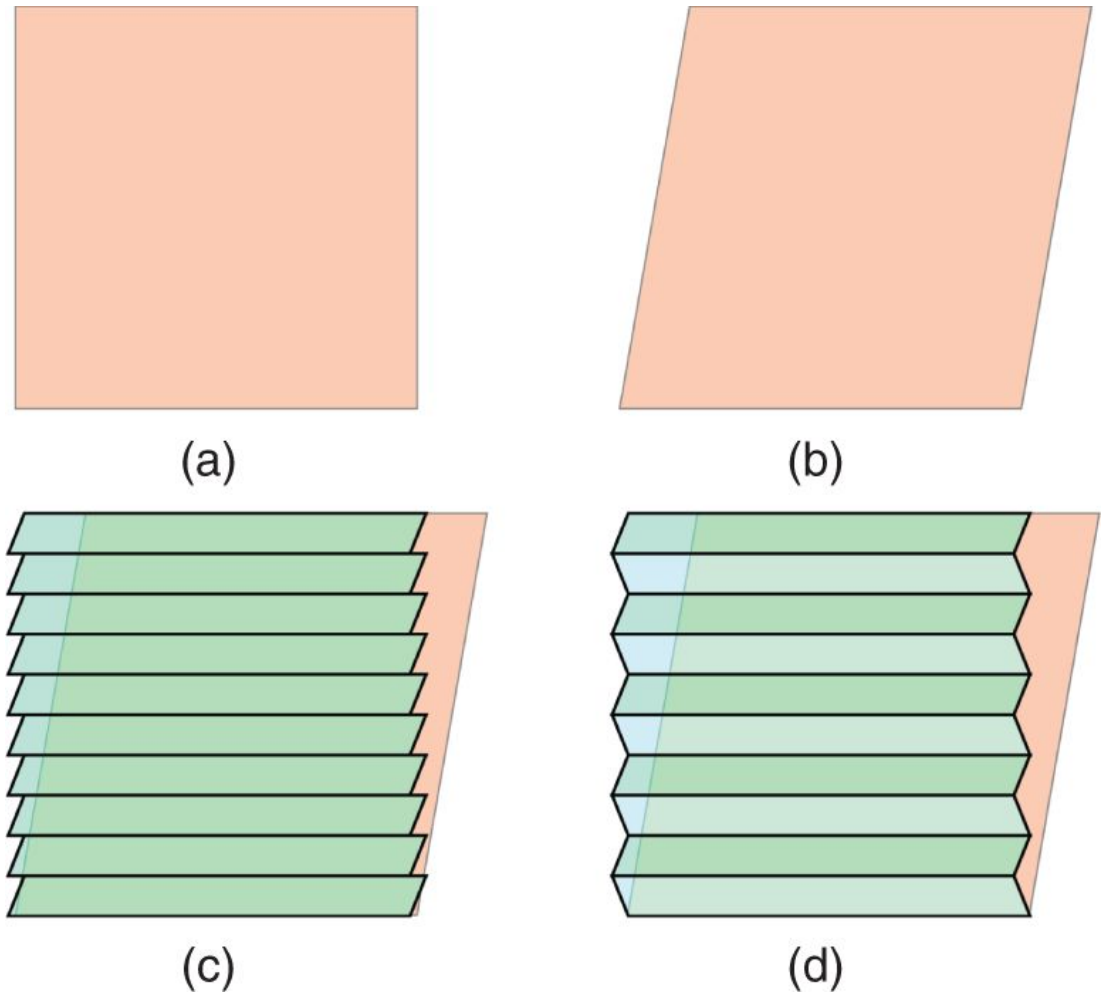


Figure 6.17 The shape-change caused by deformation (a, b) can be minimised by repeated slip (c) or polysynthetic twinning (d).

The temperature at which the transformation to martensite takes place is found to be composition dependent. Martensite starts to form when the temperature reaches about 700 °C for the lowest carbon content steels but not until a temperature of about 200 °C for austenite with a carbon content of 1.2 wt%. The temperature at which the martensite starts to form is usually labelled M_s, the *martensite start temperature*, and the temperature at which the transformation is complete is labelled M_f, the *martensite finish temperature*. If martensitic steel is reheated, the process is reversed and austenite begins to form at A_s, the *austenite start temperature* and is complete at A_f, the *austenite finish temperature*. These temperatures are also composition dependent. M_s and M_f invariably differ from A_s and A_f and the cycle between austenite and martensite is an example of *hysteresis*, commonly found in solid-state transformations (Figure 6.18). The martensite start temperature for a steel depends on a large number of parameters, including the composition,

austenite grain size and shape, and mechanical stress and strain. As the transformation is of prime importance for the control of the properties of steel, considerable effort is currently invested into the construction of computer models to predict M_s for a wide range of steel-related alloys.

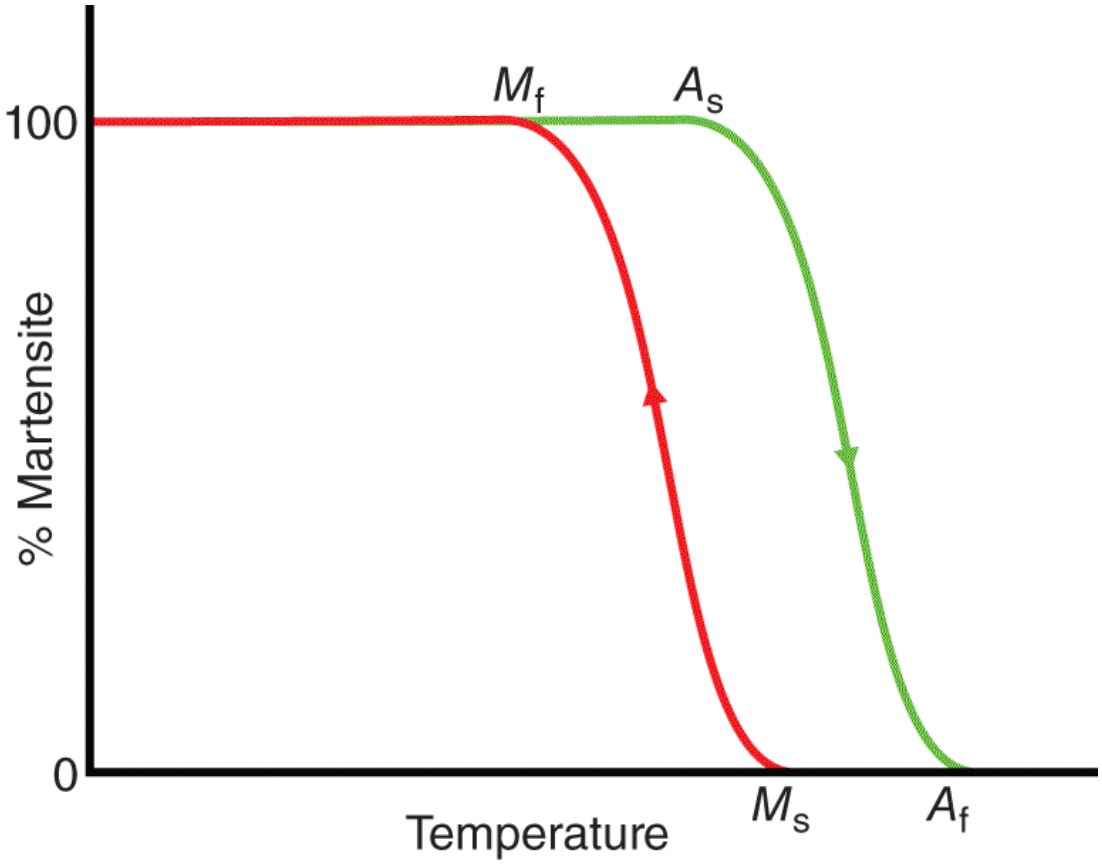


Figure 6.18 Hysteresis (schematic) between martensite start and finish temperatures, M_s and M_f and austenite start and finish temperatures A_s and A_f, in a typical steel.

Martensite can also revert to ferrite and cementite when the steel is heated at a temperature below the austenite stability region and above that at which carbon diffusion becomes possible. This process is called *tempering*. During tempering, the components present revert to those formed during slow cooling of austenite (see Section 6.6). However, the *microstructure* of the steel will be different from that achieved by slow cooling as the intermediate martensite microstructure will be quite different to that of pure austenite. The properties of steels are controlled by choosing a cycle of slow cooling, quenching, and tempering to optimise the microstructure formed with respect to the end use.

6.5.2 Martensitic Transformations in Zirconia

At room temperature, zirconia, ZrO_2, and isostructural, HfO_2, both adopt the monoclinic baddeleyite structure. These phases show a progressive transformation from monoclinic to tetragonal and finally adopt the cubic fluorite structure as the temperature rises:

$$ZrO_2\text{: baddeleyite (monoclinic)} \xrightarrow{1170^\circ C} \text{(tetragonal)} \xrightarrow{2370^\circ C} \text{fluorite (cubic)}$$

$$HfO_2\text{: baddeleyite (monoclinic)} \xrightarrow{1720^\circ C} \text{(tetragonal)} \xrightarrow{2600^\circ C} \text{fluorite (cubic)}$$

The crystal structures of the tetragonal and monoclinic phases are slightly distorted versions of the high temperature cubic fluorite structure form. This high temperature structure cannot be quenched to room temperature but can be stabilised by doping with cations such as Ca^{2+} and Y^{3+} (Section 5.13.3). The monoclinic–tetragonal phase transition is a displacive martensitic transition and it is the dilation of 4–5% and the shear strain of 14–15% that cause the fragmentation of pure zirconia and hafnia bodies on thermal cycling, thus preventing their use in refractory components. The tetragonal to monoclinic transition is typically fast, with the phase boundary moving at about the speed of sound.

The lattice parameters of the zirconia phases are:

Monoclinic: $a = 0.5193$ nm, $b = 0.5204$ nm, $c = 0.5362$ nm, $\beta = 99.2^\circ$

Tetragonal: $a = 0.5132$ nm, $c = 0.5228$ nm

Cubic: $a = 0.5144$ nm

The structural correspondence (Figure 6.19) indicates that $[010]_t$ is parallel to $[010]_m$ and $[100]_t$ is parallel to $[100]_m$. Shear on $(001)_t$ parallel to the **a**-axis distorts the tetragonal to the monoclinic form. In the transition, Zr^{4+} cations appear to be displaced by a shear; rapidly followed by the oxygen ions then shuffling into new positions, leading to the growth of heavily twinned martensitic plates in the untransformed matrix.

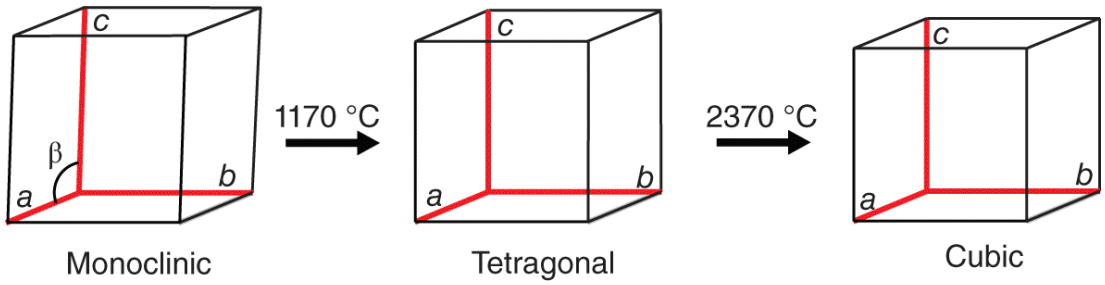


Figure 6.19 The structural relationship between the forms of zirconia.

The tetragonal to monoclinic transition has been studied in zirconia nanoparticles. Roughly spherical tetragonal particles transform to rod-like monoclinic particles with (001) twin planes. In this context it is of interest to note that the tetragonal form of zirconia can be stabilised to room temperature in nanoparticles as the relative increase in surface energy compensates for the energy gained during the transformation in large grains. The critical size for stabilisation is somewhere between 10 and 30 nm diameter.

6.5.3 Martensitic Transitions in Ni–Ti Alloys

An important martensitic transformation occurs in the titanium–nickel alloy TiNi (Figure 6.20), called Nitinol, used in shape-memory alloys (Section 6.5.4). At temperatures above 1090 °C, TiNi has a body-centred cubic structure in which the atoms are distributed at random over the available sites in the crystal. Below 1090 °C, this structure orders to form the cubic CsCl structure (Figure 6.21a). If this latter phase is quenched to room temperature, the structure transforms via a martensitic transformation into a monoclinic crystal. The atomic displacement that occurs is essentially a shear in the {101} planes of the cubic structure. The {101} planes are made up of a triangular net of atoms, with alternating rows of Ni and Ti atoms (Figure 6.21b,c). If these shear so as to stack in the hexagonal closest packing arrangement ...ABAB... the resulting structure is the AuCd type, with typical orthorhombic lattice parameters of a = 0.4767 nm, b = 0.3164 nm, c = 0.4885 nm. In Nitinol the orthorhombic AuCd structure is slightly deformed into a monoclinic unit cell with typical monoclinic lattice parameters of a = 0.4120 nm, b = 0.315 nm, c = 0.4623 nm, β = 96.8°.

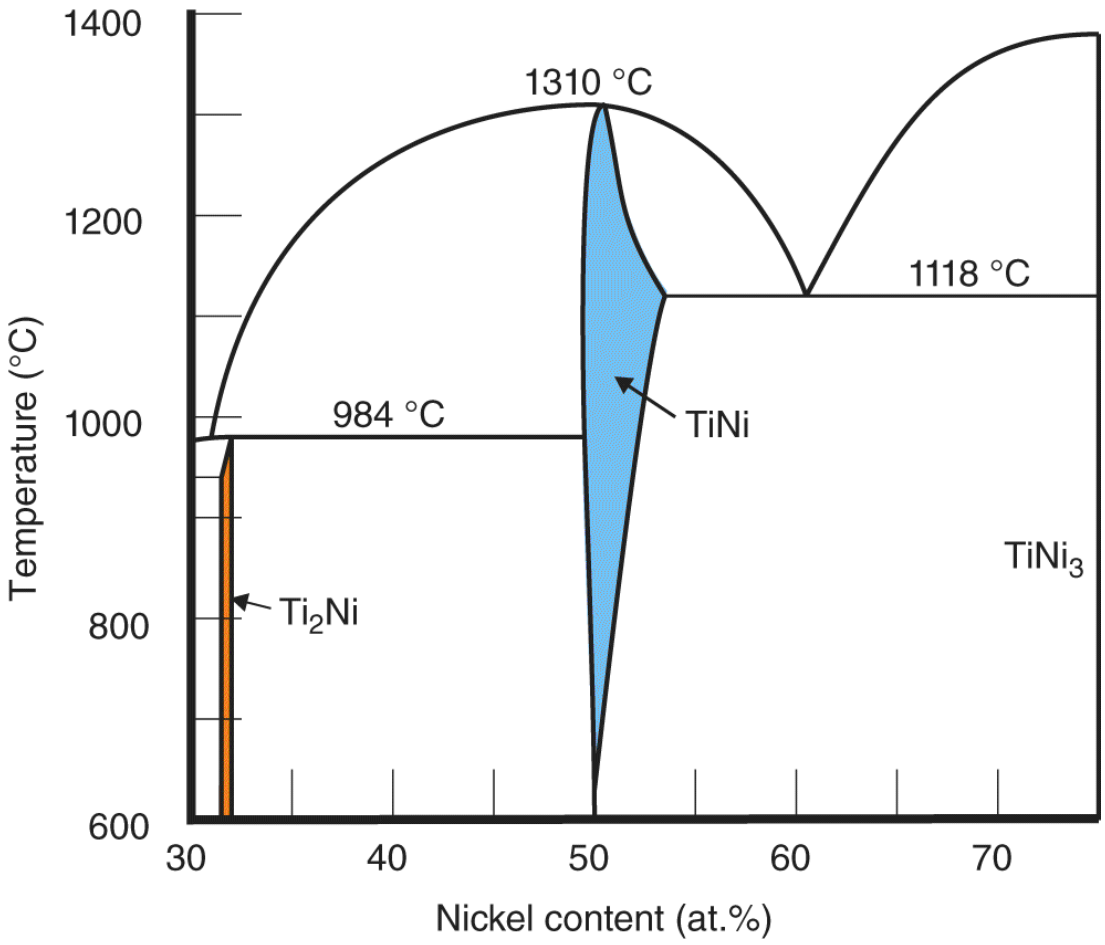


Figure 6.20 Part of the phase diagram of the Ti–Ni system.

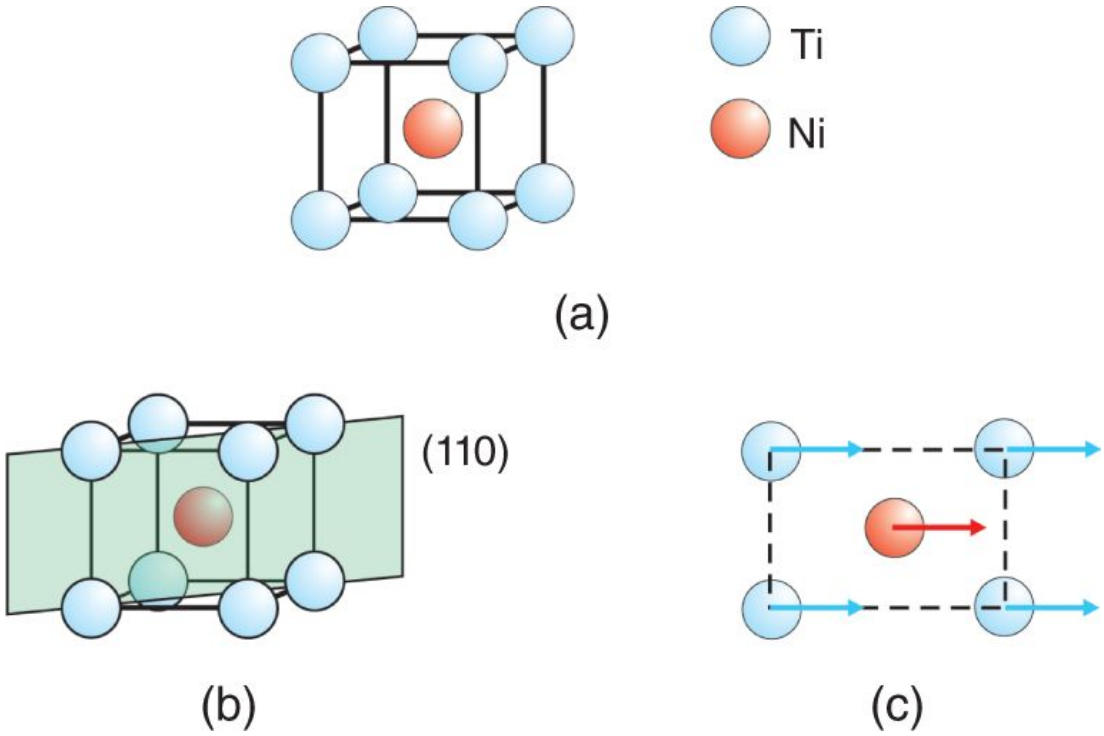


Figure 6.21 The CsCl structure of the high-temperature form of TiNi: (a) cubic unit cell; (b) atoms in a (110) plane; (c) shear displacement that occurs upon transformation to martensite.

On cooling, the transformation starts at the martensite start temperature M_s, 60 °C and is complete by the martensite finish temperature M_f, 52 °C. As with the majority of martensitic transformations, the shearing process nucleates at a number of points within the crystal as it cools, and each grows to form a domain or variant. Extensive twinning helps to minimise the strain in the crystal and when the transformation is completed the material is highly twinned, but maintains the same bulk shape as the original material (Figure 6.22).

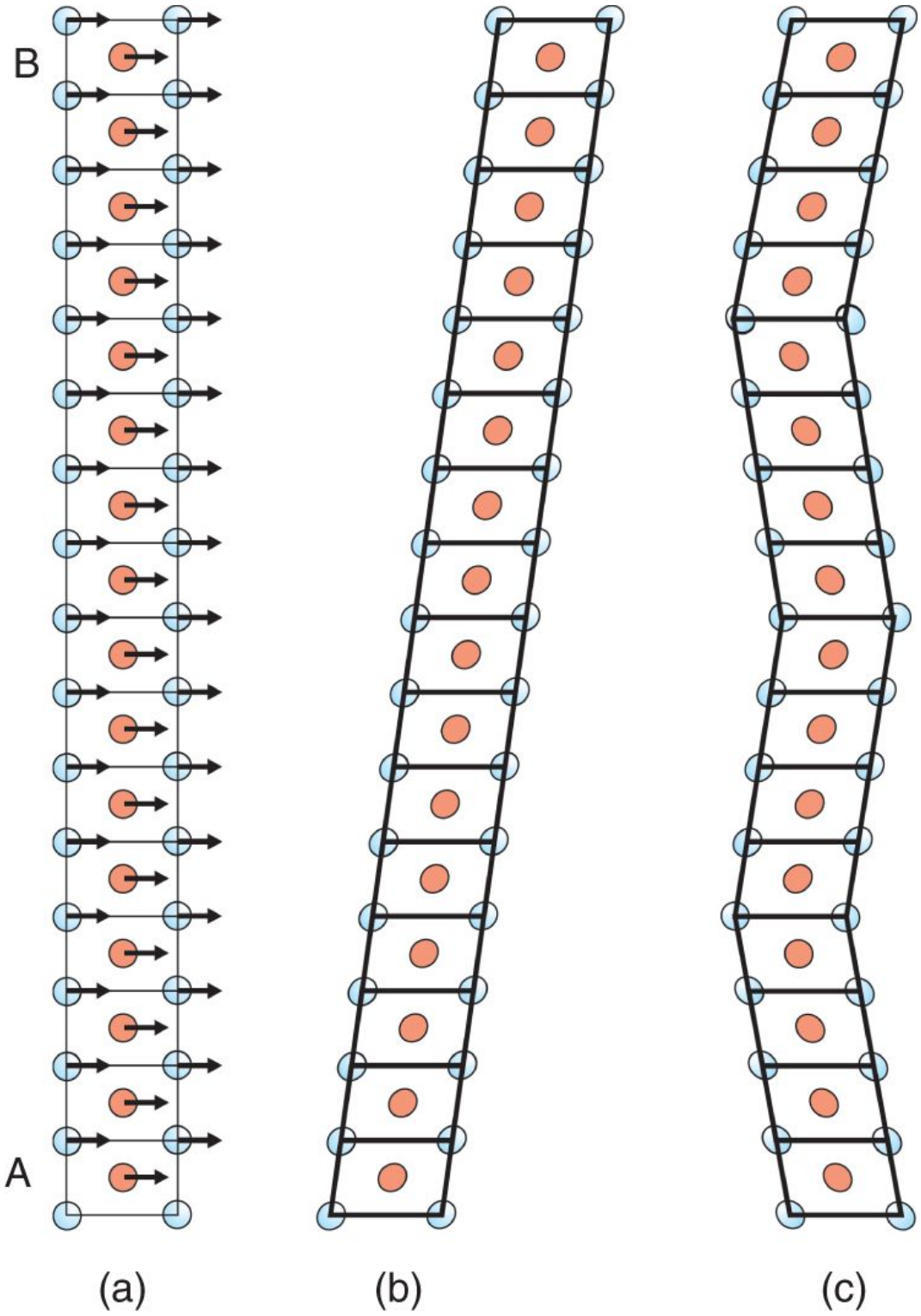


Figure 6.22 (a) A stack of unit cells of TiNi projected down [110]. Shear displacement is applied to each cell in sequence, starting at A and ending at B to produce a martensitic structure (b). (c) The overall shape of the original stack is more nearly maintained by repeated twinning.

The formation of a heavily twinned material on cooling can be reversed by an increase in temperature, which causes the material to transform to the untwinned pre-martensite state. The transformation starts at the austenite start temperature A_s, 71 °C and is complete at the austenite finish temperature A_f, 77 °C. (These terms are adopted directly from the nomenclature for steels although no austenite is present.) As in steels, M_s and M_f differ from A_s and A_f and the system shows considerable hysteresis.

The transformation can be greatly modified by the addition of other alloying elements or by variation of the heat treatment to which the alloy is subjected. For example, TiNi has an extension in composition from $Ni_{1.0}Ti_{1.0}$ to about $Ni_{0.86}Ti_{1.14}$ at 1118 °C. This phase range narrows at lower temperatures, so that cooling an alloy with a composition slightly richer in nickel than NiTi results in the formation of precipitates of $TiNi_3$ in the TiNi matrix. These considerably modify the mechanical properties of the martensitic phase. Moreover, this modification is dependent on precipitate morphology and size, and therefore on the rate of cooling. The formation of precipitates also influences M_s and it has been found that a change in composition of as little as 1 at.% can move M_s by more than 100 K. The martensitic transformation in this alloy is thus easily open to modification by heat treatment and alloying, which gives these microstructures great flexibility from the point of view of engineering design.

6.5.4 Shape-Memory Alloys

Shape-memory alloys are a group of metallic materials that can regain their original shape after deformation. This is a rather remarkable property and has been used in a wide variety of devices, which range from antennae on space craft, that can be crumpled into a small volume for launch, and then unravel into a dish form on deployment, to spectacle frames that can be returned to their original shape after being sat on!

Shape-memory alloys show a *thermoelastic* martensitic transformation. This is a martensitic transformation, as just described, but which, in addition, must have only a small temperature hysteresis, some 10's of degrees at most, and mobile (easily moved) twin boundaries. Additionally, the transition must be crystallographically reversible. The unusual elastic behaviour is described in terms of *superelasticity* (Section 8.1.4).

To initiate the shape-memory effect, the alloy is formed into the desired final geometry at temperatures above M_s. On cooling, this original form is maintained, but the material transforms to a heavily twinned state. If this shape is now deformed (crumpled up, say) the twin boundaries move to accommodate the stress. It is for this reason that the twin boundaries must be

mobile. In effect, the size of the individual twins increases, and the number of twins decreases. Reheating above A_f causes the structure to return to the high temperature state with the original shape (Figure 6.23). A crystallographically reversible transition is a requirement for this stage. The process by which the original form is recovered so that the material has 'remembered' its original shape is simply one of twin-plane elimination (Figure 6.24).

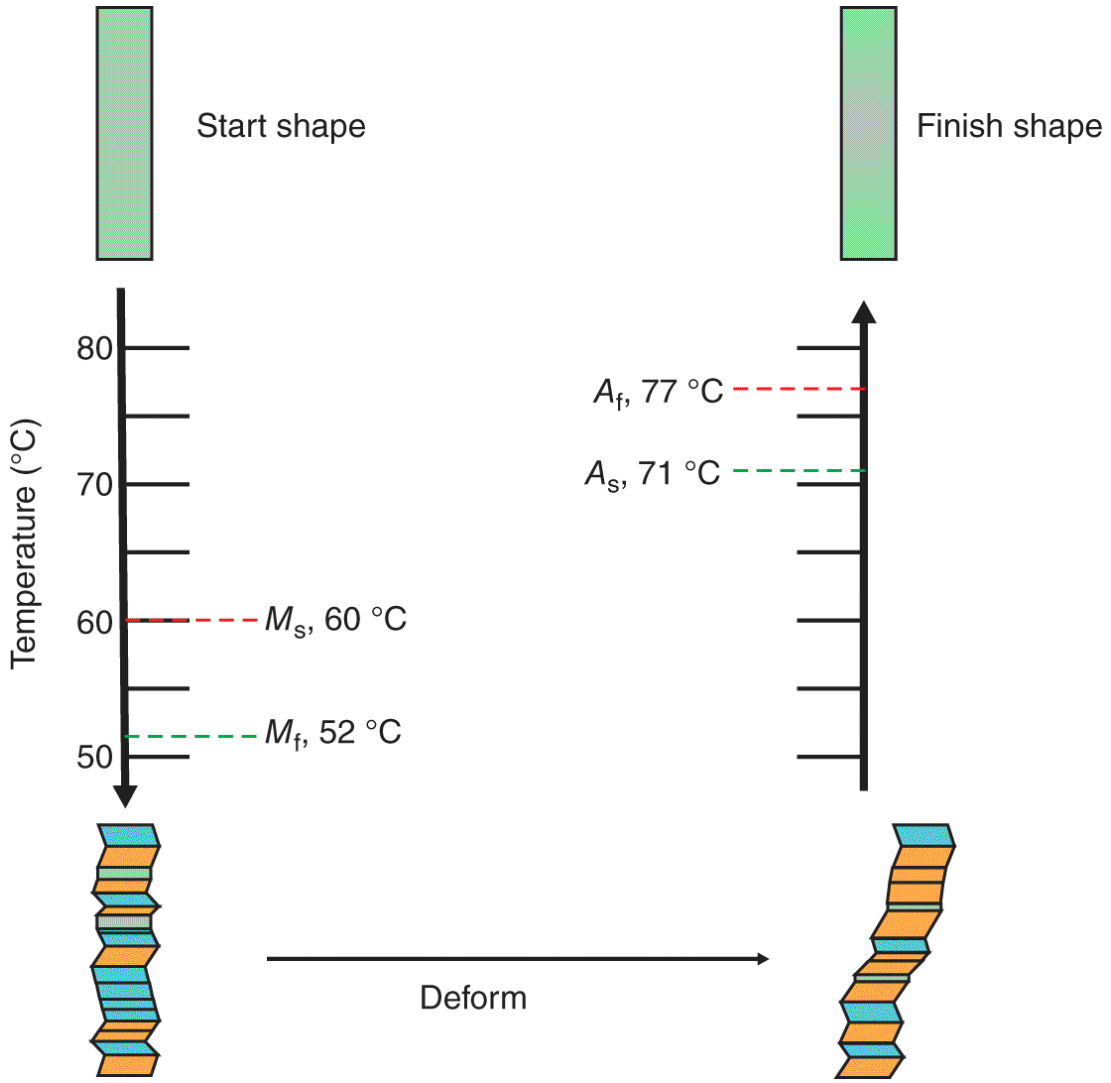


Figure 6.23 The sequence of events taking place during deformation and shape-recovery of a shape-memory alloy. The temperatures are appropriate for TiNi.

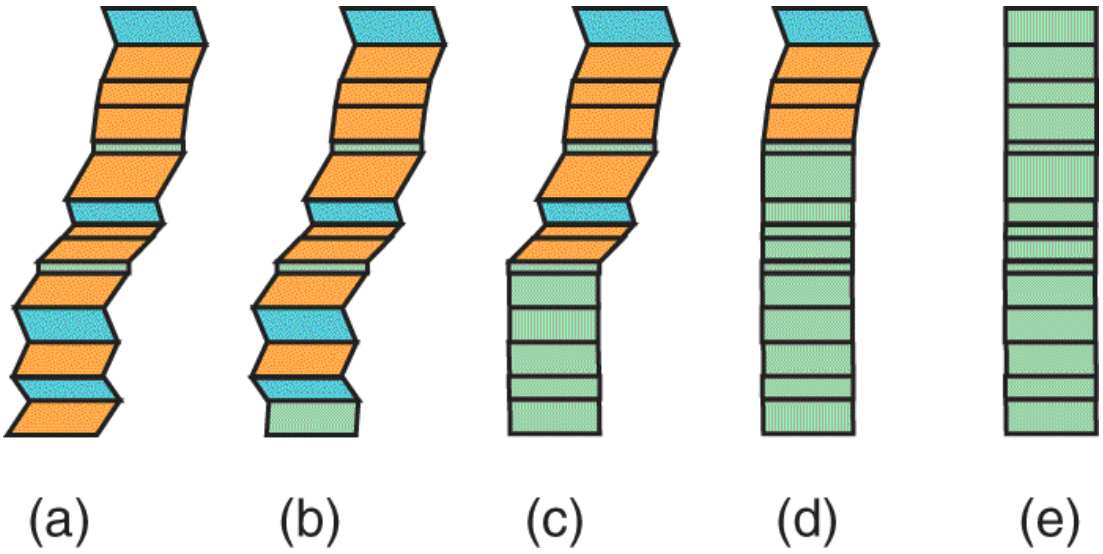


Figure 6.24 (a–e) The progressive recovery of the original shape of a deformed rod by twin removal.

In the shape-memory transformation described, only the shape of the parent phase is 'remembered'. It is called the *one-way shape-memory effect*. It is also possible to produce alloys that display two-way shape-memory effects. In these materials, both the shape of the parent phase and the martensitic phase are 'remembered'. This reversible effect is caused by the fact that the nucleation of the martensite is very sensitive to the stress field. Introduction of lattice defects such as precipitates can restrict the number of variants that form and the positions where they nucleate. Such materials generate the martensitic shape on cooling below the M_f temperature. Cycling between higher and lower temperatures causes the alloy to switch alternately between the two shapes.

Nitinol only has a useful temperature working range, when changes due to composition are included, of from approximately −20 to 80 °C. This is rather limiting, as space applications need an application range of approximately −150 to 120 °C and automotive applications require a range of approximately −50 to 150 °C. Alloys in the Fe–Ni–Co–Al system can be useful from approximately −80 to 40 °C, and alloys in the Fe–Mn–Al–Ni system have shown shape-memory effects over the increased temperature range of −200 to 150 °C. The shape-memory effect is also apparent in zirconia ceramics. Practical applications of the transformation are rather limited because of the brittleness of these materials. However, the transformation takes place at a much higher (several hundred degrees) higher than in Nitinol, making it of potential use in extreme conditions.

6.6 Phase Diagrams and Microstructures

6.6.1 Equilibrium Solidification of Simple Binary Alloys

One of the most important phase transitions occurs when a liquid transforms into a solid. A great deal of information concerning the microstructure of the solid so formed can be obtained from a consideration of the phase diagram of the material, even though phase diagrams refer to equilibrium conditions and solidification is rarely carried out so slowly as to be an equilibrium process. For example, view the solidification of a simple nickel–copper alloy from the point of view of the phase diagram. In the liquid state any nickel–copper alloy is homogeneous. Solid will begin to form as soon as the temperature reaches the liquidus, at T_1 (Figure 6.25a). The initial composition of the liquid is l_1 (virtually equal to c) and that of the solid is s_1. The solid is rich in nickel compared with the original liquid composition. If the material is held at a temperature of T_1 for long enough, a *dynamic equilibrium* will be achieved. In this state, although the system appears to be static, the solid is continually dissolving and reforming. The atoms that are in the liquid and solid phases are continually being exchanged.

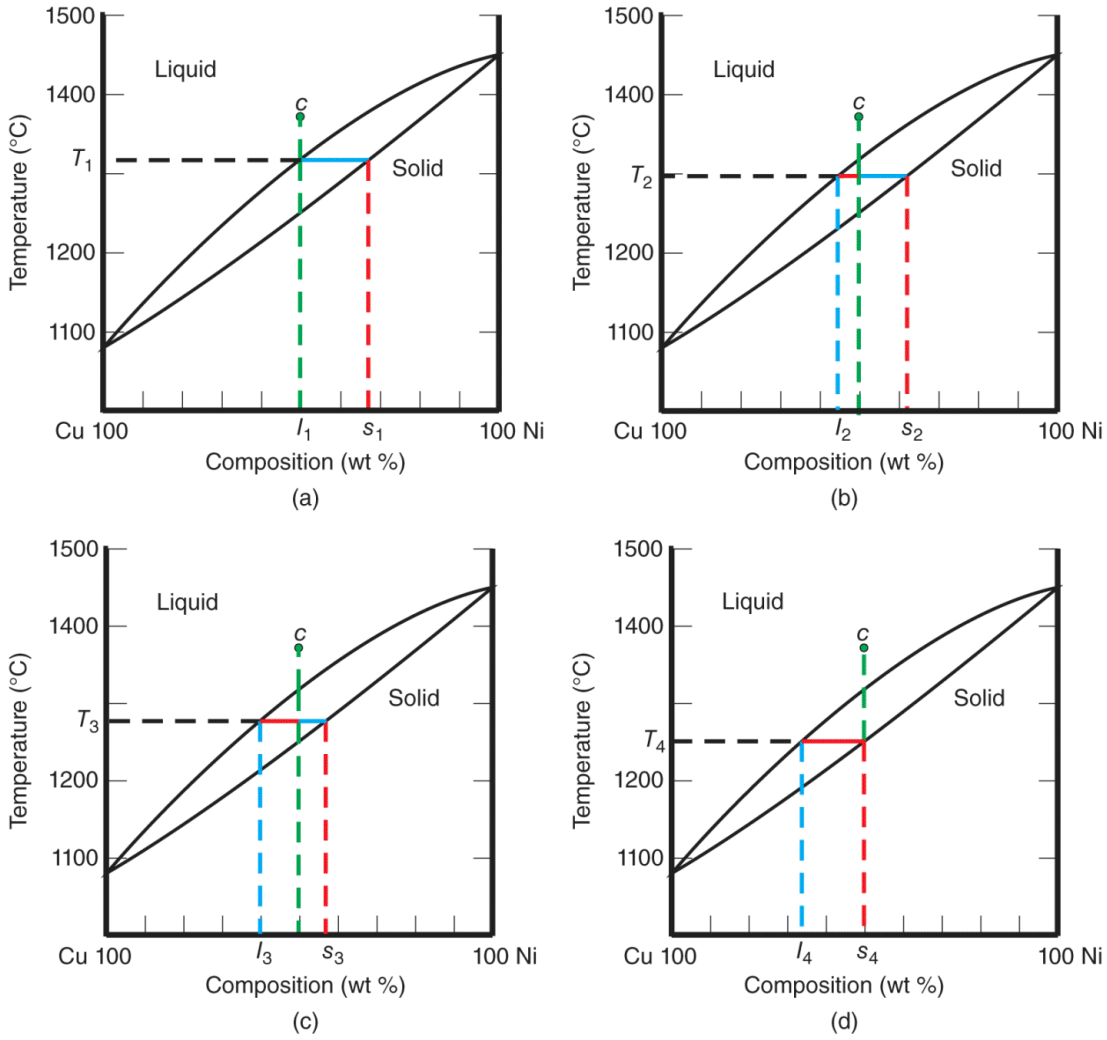


Figure 6.25 Cooling a sample of composition c in the Cu–Ni system: (a) liquid composition l_1, solid composition s_1; (b) liquid composition l_2, solid composition s_2; (c) liquid composition l_3, solid composition s_3; (d) liquid composition l_4, solid composition s_4.

As the mixture is slowly cooled, this exchange leads to a continuous change in the composition of the solid and liquid phases. When the temperature drops slightly to T_2 (Figure 6.25b), the original solid of composition s_1 will be replaced by a solid of composition s_2 (much exaggerated in the figure). The new composition of the liquid in equilibrium with the solid is l_2. At all times dynamic equilibrium holds and the atoms in the solid are dissolving in the liquid while other atoms in the liquid are crystallising to form solid. Over a period of time the crystal present will always have the composition s_2, although the actual atoms that comprise the solid are forever changing. The same is true of the liquid. Further slow cooling, to temperature T_3, will cause the

composition of the solid to change gradually to s_3, in equilibrium with liquid of composition l_3 (Figure 6.25c). This imaginary process can be continued until all of the original composition c is identical to the point on the solidus s_4, at temperature T_4 (Figure 6.25d). The final trace of liquid in composition with this solid has a composition l_4.

Thus, during equilibrium cooling, the composition of the solid will run down the solidus line, s_1 to s_2 to s_3, etc. and the composition of the liquid in equilibrium with the solid runs down the liquidus from l_1 to l_2 to l_3, etc. as the liquid cools. The composition of the solid phase when all of the liquid has solidified will be equal to that of the original liquid phase. Not only does the composition of the solid and liquid phases change continuously as the temperature falls through the two-phase region, but also the number of small crystals present also increases. When temperature T_4 is reached, the microstructure of the solid consists of crystallites or grains (Figure 6.26a). Further cooling in the solid will not initiate change of composition.

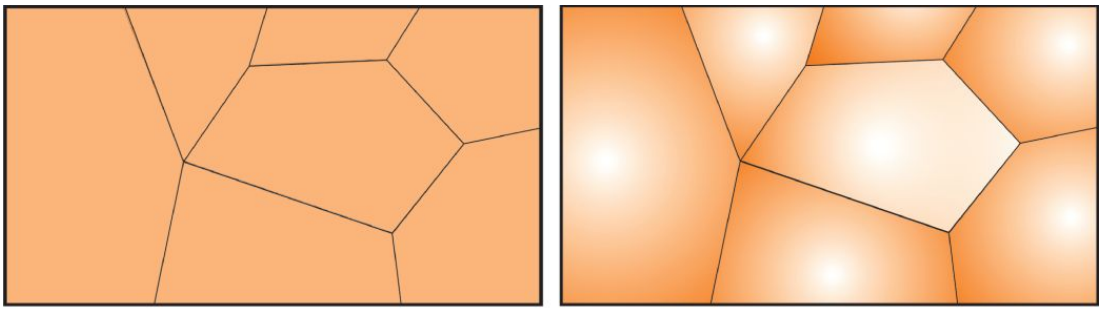


Figure 6.26 Microstructures of a solidified Cu–Ni allot: (a) slow cooling; (b) rapid cooling.

6.6.2 Non-equilibrium Solidification and Coring

During normal processing, cooling is usually rather fast and solidification is rarely an equilibrium process. Solidification in these conditions is extremely complex. If cooling is not too fast, the first material to precipitate will still have a composition s_1 (Figure 6.25a). However, there will be insufficient time on further cooling for the original solid to equilibrate, and new material of composition s_2 will start to form on the nucleus of composition s_1. Ultimately the solid will consist of a core that is richer in nickel than the outer regions and there will be a gradation of composition of the alloy from the inside to the outside (Figure 6.26b). This is called *coring*. Coring can occur in all crystallites formed rapidly.

These non-equilibrium structures can be removed by annealing solid at a temperature below that of the solidus. Naturally this is only effective if the atoms can diffuse in the solid at the annealing temperature.

6.6.3 Solidification in Systems Containing a Eutectic Point

The composition and microstructure of a solid formed by slow cooling in a system with a eutectic point depends critically upon the composition of the liquid with respect to the eutectic composition. The situation will be explained using the tin–lead phase diagram (Figure 2.14).

In a liquid tin–lead alloy, the two atom species are mixed at random. A liquid alloy with the same composition as that of the eutectic point, c_e, called the *eutectic composition*, is unique. On cooling slowly through the eutectic point, it will pass directly into the solid state at a temperature 183 °C, without traversing a two-phase solid + liquid region (Figure 6.27a). However, the solid that forms is not homogeneous but must contain two phases, solid α and solid β. Thus, the random arrangement of atoms in the liquid must separate into the appropriate solid compositions on solidification. It is found each grain contains a characteristic microstructure that consists of thin alternating lamellae of the two phases called *eutectic* α and *eutectic* β (Figure 6.27b). The actual thickness of the lamellae and their shapes will depend upon the relative diffusion coefficients of the two species. Note, though, that the *average* composition of the solid will be the same as that of the eutectic point.

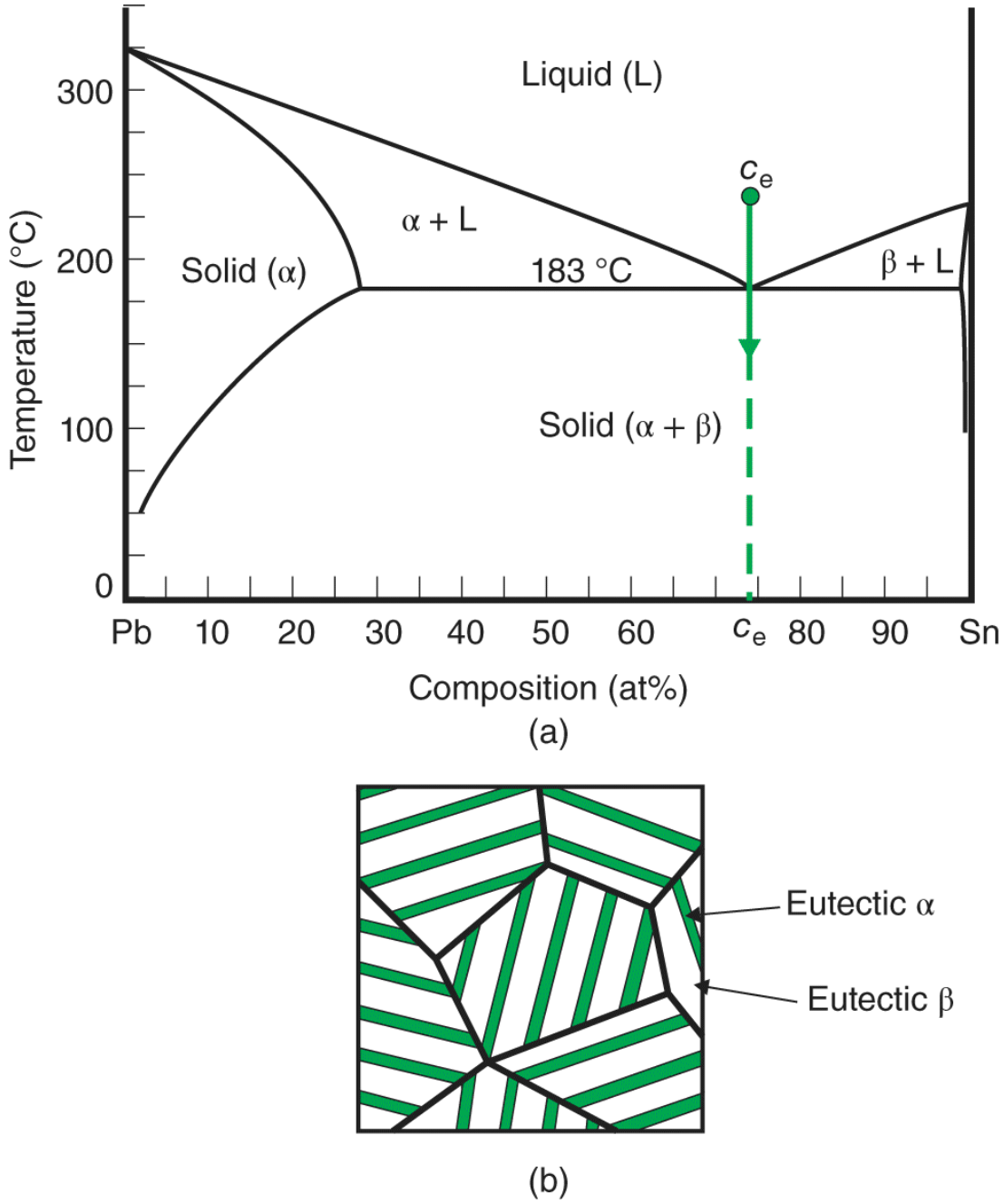


Figure 6.27 The microstructure of solidified Pb–Sn alloys: (a) the Pb–Sn phase diagram showing the eutectic composition c_e; (b) grain structure consisting of eutectic α and eutectic β.

The idealised microstructure of a solid formed when other liquid compositions slowly cool reflects the presence of both liquid and solid. Suppose that the liquid with composition c_1, richer in lead than the eutectic composition (Figure 6.28a), is slowly cooled. At temperature T_1, below the liquidus, some solid α-phase will have nucleated (Figure 6.28b). The composition of the solid

phase is s_1 and that of the liquid phase is l_1. As cooling continues, the composition of the solid α crystallites will move along the solidus, as described earlier for the nickel–copper alloys. At the same time, the composition of the liquid phase in contact with the crystallites will move along the liquidus. For example, at temperature T_2 the solid has a composition of s_2 and the liquid a composition of l_2 (Figure 6.28a). Ultimately the horizontal solidus line will be reached at the eutectic temperature. At this point, any further drop in temperature will cause the remaining liquid to solidify. The microstructure of this latter material will be the same as the characteristic eutectic structure described above. The solid is a mixture of eutectic α and β and crystallites of the α-phase that formed in contact with the liquid, called *primary* α (Figure 6.28c). A similar situation will occur for compositions on the tin-rich side of the eutectic point. In this case, the microstructure will consist of precipitates of *primary* β in a eutectic matrix.

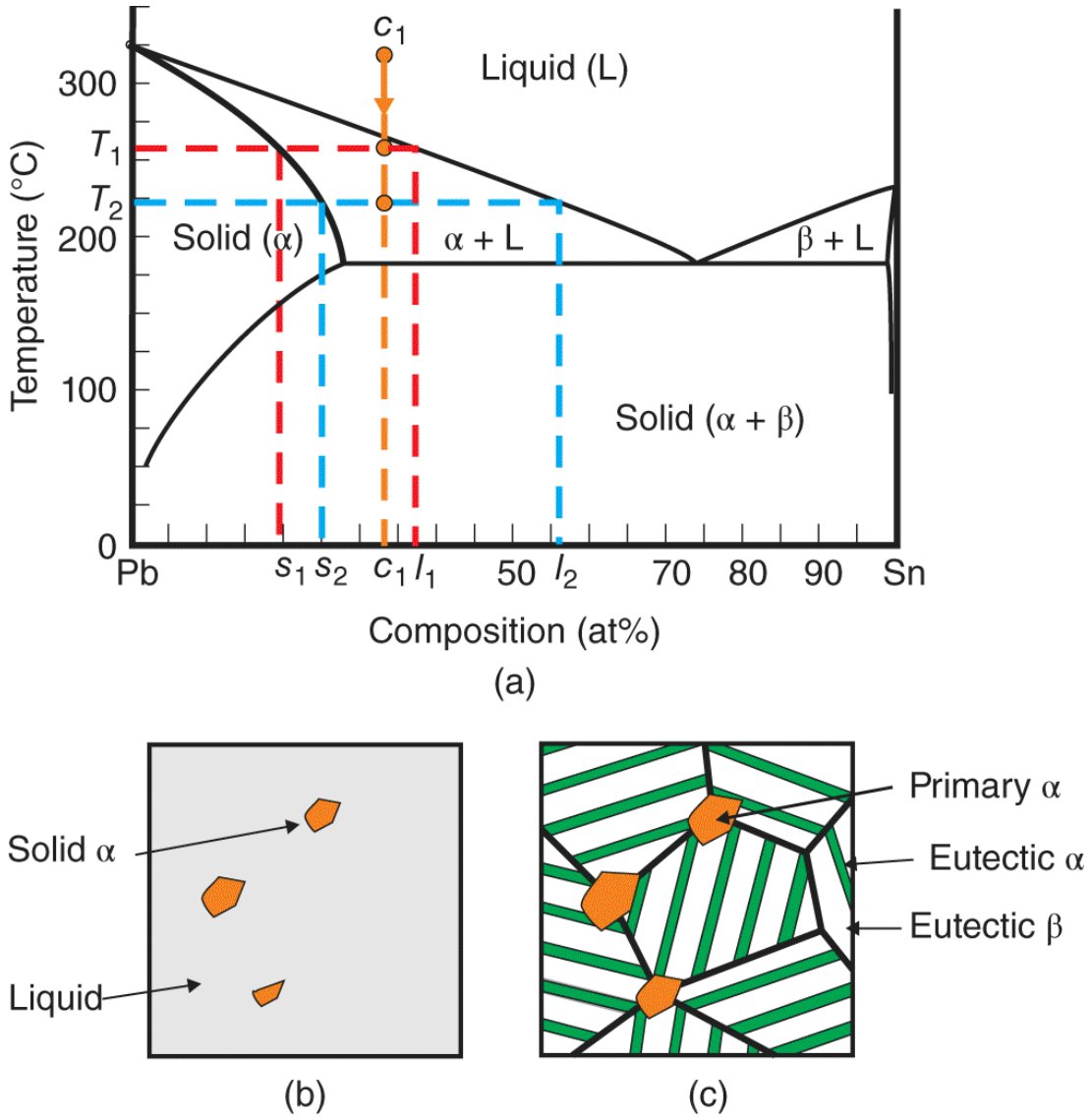


Figure 6.28 The microstructure of solidified Pb–Sn alloys: (a) the Pb–Sn phase diagram showing the Pb-rich composition c_1; (b) cooling initially produces solid and liquid phases; (c) solidification produces α, eutectic α, and eutectic β.

In the case of a very lead-rich composition such as c_2 (Figure 6.29a), the first phase to form as the temperature passes the liquidus is solid α in liquid. Ultimately the temperature will fall below the solidus, and at a temperature T_1, for example, the solid will consist of grains of α-phase (Figure 6.29b). On cooling further, at temperature T_2, for example, the temperature will fall below the solvus, and the solid will consist of precipitates of the β-phase in grains of the α phase (Figure 6.29c).

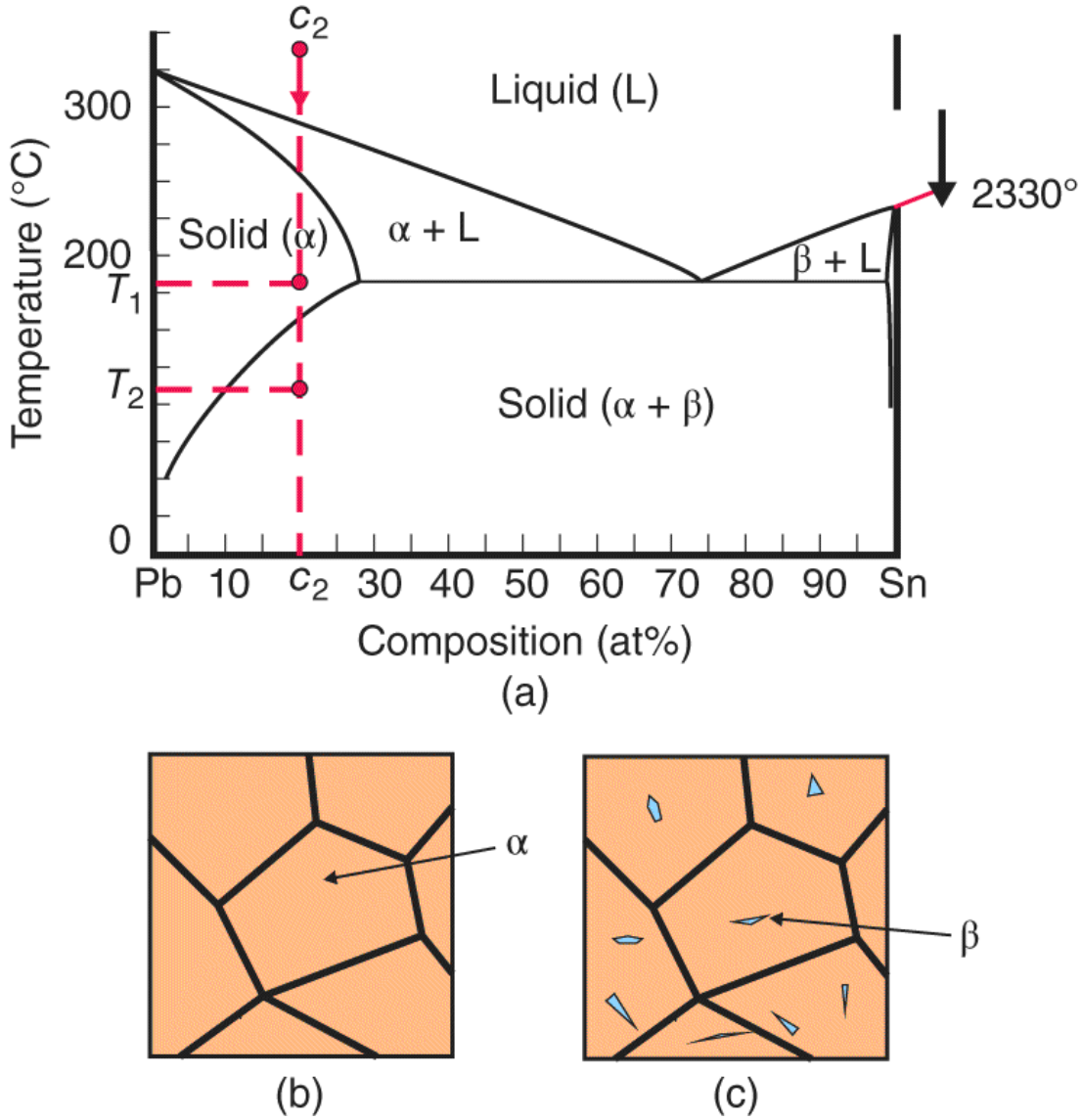


Figure 6.29 The microstructure of solidified Pb–Sn alloys: (a) the Pb–Sn phase diagram showing the Pb-rich composition c_2; (b) cooling initially to T_1 produces a homogeneous solid α phase; (c) cooling to T_2 produces additional grains of the β phase.

The compositions and amounts of the phases present at all times can be calculated by use of tie lines and the lever rule (Section 1.2.1).

6.6.4 Equilibrium Heat Treatment of Steel in the Fe–C Phase Diagram

The changes brought about by quenching of austenite (Section 6.5) are quite different from those produced by near equilibrium cooling of austenite. The most important transformation-point in this region of the Fe–C phase diagram

from the point of view of carbon steels is the eutectoid point. This is at 0.76 wt% C (0.034 at.% C) and a temperature of 727 °C. Above this temperature (T_1, Figure 6.30a), an alloy with the eutectoid composition, c_e, is single γ phase (austenite) (Figure 6.30b). The face-centred cubic structure of austenite is strained by the interstitial carbon atoms and this increases substantially as the temperature slowly falls and the face-centred cubic unit cell contracts. Ultimately the austenite transforms into a two-phase mixture of α phase (ferrite) and cementite (Fe_3C) at a temperature of 727 °C (T_2, Figure 6.30a), with internal strain likely to provide the driving force.

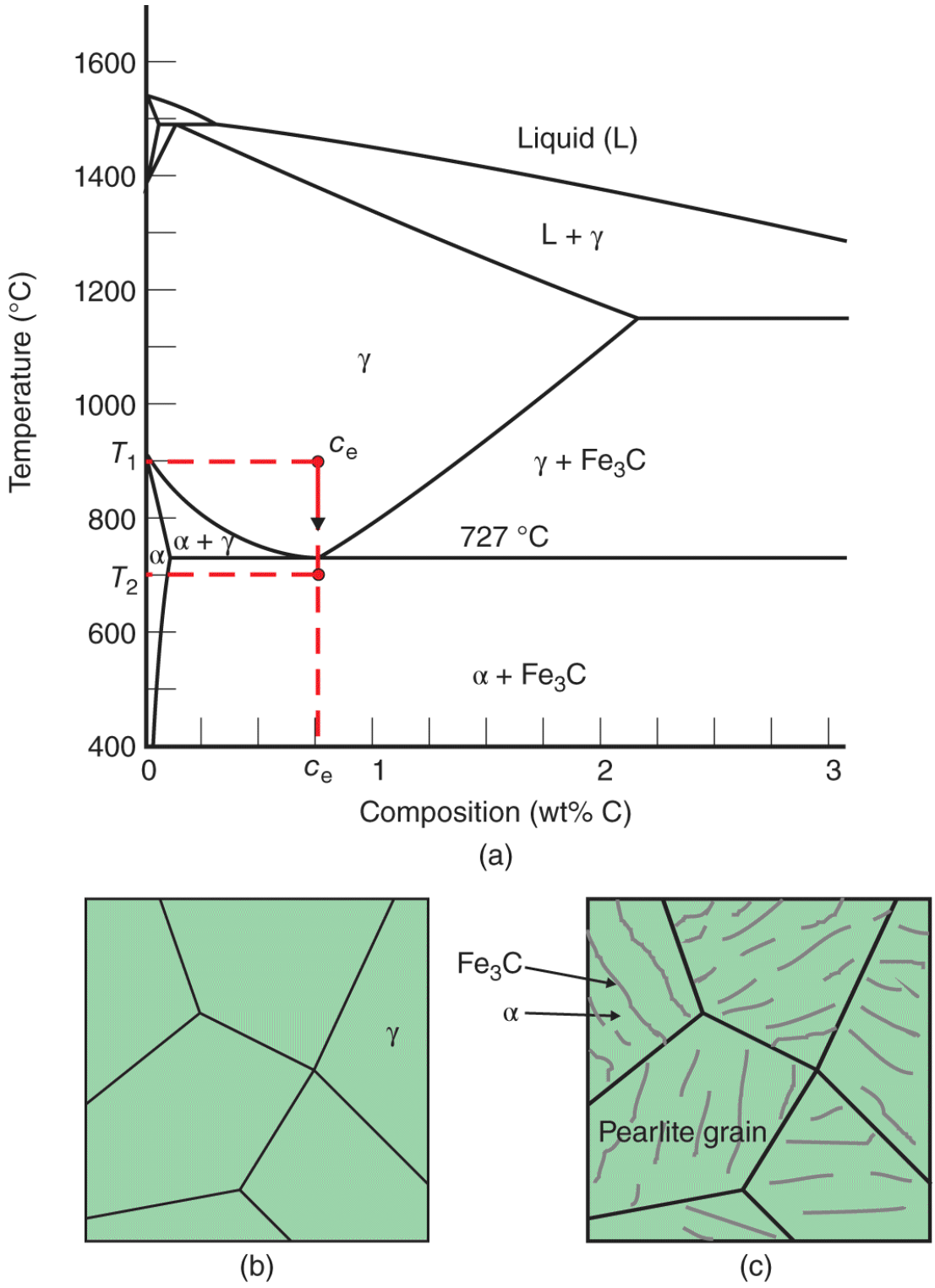


Figure 6.30 (a) The existence diagram of the Fe–C system. (b) Microstructure of the solid with overall composition c_e at temperature T_1. (c) Microstructure of the solid with overall composition c_e at temperature T_2.

The transformation is complex. The phase diagram shows that body-centred cubic ferrite, one of the phases existing below the eutectoid temperature, is hardly able to dissolve any carbon. The transformation requires diffusion of the carbon to create carbon-poor regions that become ferrite and carbon-rich regions that become cementite. At the same time the iron atoms must rearrange considerably. The transformation of the face-centred cubic austenite array into the body-centred cubic ferrite array requires considerable shuffling of the iron atoms. The rearrangement of the iron atoms in austenite to that found in cementite is much more complex. The structure of cementite is made up of hexagonal close packed layers of Fe atoms twinned on every third $(11\bar{2}2)$ plane with respect to the hexagonal close packed unit cell. The transformation thus involves changing the ...ABC... packing of austenite into multiply twinned ...ABAB... packing. This twinning is a method of minimising the internal strain in the structure while maintaining the overall shape and volume of the solid (Section 6.5). The carbon atoms lie in the twin planes of the cementite structure, where most room occurs.

The newly formed cementite nucleates at many sites simultaneously. The resulting solid, made up of thin lamellae of cementite and α-ferrite side by side, is called *pearlite* because it has a lustrous appearance in an optical microscope. Pearlite is not a compound or single phase, but a *microstructure* (Figure 6.30c) and in this context the phases are called *eutectoid ferrite* and *eutectoid cementite*.

The microstructures are formed at compositions away from the eutectoid point mirror those previously described for eutectic transformations. Compositions to the iron-rich side of the eutectoid are called *hypoeutectoid alloys*. For example, a hypoeutectic composition c_1, at temperature T_1 (Figure 6.31a), consists of homogeneous grains of austenite (Figure 6.31b). As these cool slowly, the austenite region is exchanged for a two-phase region (temperature T_2, Figure 6.31a). This material is transformed to ferrite plus austenite. The ferrite often forms at grain boundaries, as the reaction is kinetically favoured in the disordered regions at grain boundaries (Figure 6.31c). It also has an almost constant composition and holds only a very small amount of dissolved carbon in its body-centred cubic structure. The remaining austenite thus becomes carbon rich. As cooling continues the compositions of the two phases run down the phase boundaries present. Ultimately the temperature reaches the eutectoid temperature, 727 °C. Further cooling causes the austenite to transform to pearlite (temperature T_3, Figure 6.31a). The microstructure now consists of pearlite (*eutectoid ferrite* and *eutectoid cementite*), together with the precipitates of the ferrite formed earlier, called *proeutectoid ferrite* (Figure 6.31d).

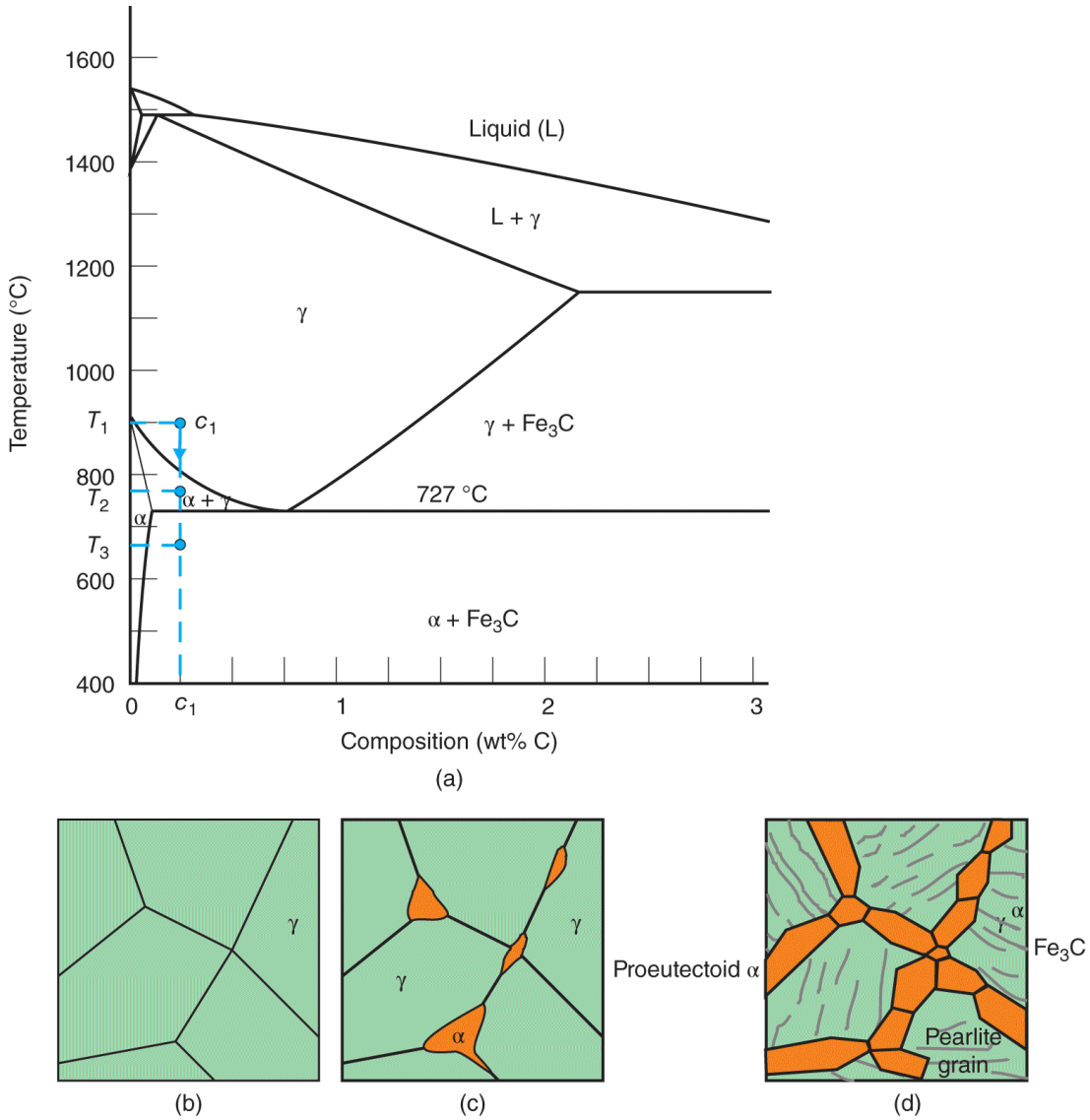


Figure 6.31 (a) The existence diagram of the Fe–C system. (b) Microstructure of the solid with overall composition c_1 at temperature T_1. (c) Microstructure of the solid with overall composition c_1 at temperature T_2. (d) Microstructure of the solid with overall composition c_1 at T_3.

Compositions on the carbon-rich side of the eutectoid are called *hypereutectoid alloys*. Once again, as such an alloy is cooled slowly it will pass from single-phase austenite into a two-phase region. Consider a hypereutectic composition c_2 at temperature T_1 (Figure 6.32a). The microstructure will consist of grains of austenite (Figure 6.32b). On cooling to temperature T_2 (Figure 6.32a), cementite separates from the austenite, forming preferentially at the grain boundaries (Figure 6.32c). One reason for this is that the strain generated in the complex transformations taking place is more easily relieved at grain boundaries. This material is called *proeutectoid cementite*. As cooling

continues, more proeutectoid cementite will form in the grain boundaries, and the composition of the remaining austenite will run down the phase boundary to the eutectoid point. Further cooling, below the eutectoid temperature, will cause any remaining austenite to transform into pearlite (temperature T_3, Figure 6.32a). The microstructure of the final solid will consist of proeutectoid cementite and pearlite, which itself consists of eutectoid cementite and eutectoid ferrite (Figure 6.32d).

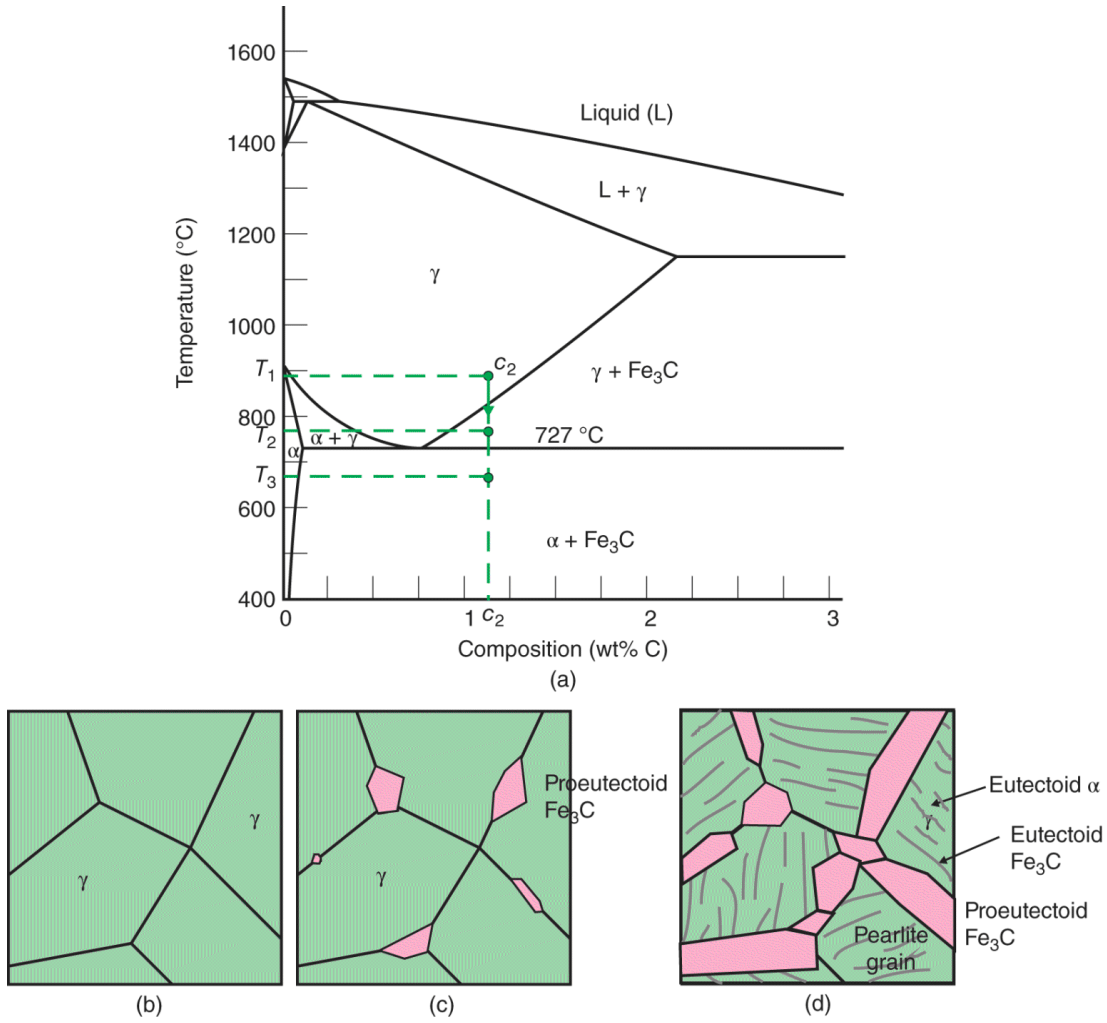


Figure 6.32 (a) The existence diagram of the Fe–C system. (b) Microstructure of the solid with overall composition c_2 at temperature T_1. (c) Microstructure of the solid with overall composition c_2 at temperature T_2. (d) Microstructure of the solid with overall composition c_2 at T_3.

6.7 High Temperature Oxidation of Metals

Corrosion of metals takes place by way of a variety of chemical reactions. At ordinary temperatures, this process is often called *tarnishing*, and at high temperatures, *scale formation*. In this section, the reaction of metals with dry gases at *relatively high temperatures* is considered. This is referred to as *direct corrosion*, to distinguish it from many common forms of corrosion, including rust formation on iron, which need the presence of water (Section 7.4).

6.7.1 Direct Corrosion

Direct corrosion occurs when a gas in the environment reacts directly with a metal. These reactions are of importance in many diverse applications: turbine blade performance in hot engines, thermocouple voltage stability in nuclear reactors, separators in high temperature fuel cells, and so on. The discussion that follows applies to all gases, but reaction with oxygen will be used to illustrate the phenomenon.

The driving force for these reactions is a decrease in the Gibbs energy of the system and oxidation reactions are influenced by the equilibrium oxygen pressure surrounding the metal. Calculation of the equilibrium partial pressures over metal oxides shows that, with few exceptions (notably the precious metals), values lie between approximately 10^{-7} atm, to 10^{-40} atm. As the oxygen partial pressure in air is about ⅕ of an atmosphere it is clear that metals will have a tendency to oxidise when exposed to air and there is always a considerable driving force for reaction.

The simplest cases of oxidation are those in which only one stable oxide exists, such as in the binary Al–O and Ca–O systems. In oxygen or air both aluminium and calcium react to form an oxide, even at room temperature:

$$4Al + 3O_2 \rightarrow 2Al_2O_3$$

$$2Ca + O_2 \rightarrow 2CaO$$

Many oxides, especially the transition metals, display several oxidation states. For instance, below approximately 375 °C copper oxidises to CuO and the oxidation reaction is:

$$2Cu + O_2 \rightarrow 2CuO$$

Between the temperatures of 375 and 1065 °C, the stable oxide is Cu_2O and the reaction is:

$$4Cu + O_2 \rightarrow 2Cu_2O$$

In many cases a certain amount of oxygen can dissolve in the metal before an oxide can be identified and it is this phase that is in equilibrium with the oxide rather than the pure metal. Thus, at a temperature of approximately 1000 °C copper metal takes into solid solution 0.33 at.% O before Cu_2O formation starts. These solid solutions are often of significance in high temperature corrosion mechanisms.

6.7.2 The Rate of Oxidation

One of the most important aspects of direct corrosion is the rate at which the reaction takes place. Clearly, if the rate is low enough, oxidation may cease to have practical consequences, as is the case with stainless steel. In reality the rates of formation of oxide films vary widely. For thin layers, nominally less than approximately 100 nm in thickness, four rate laws have been established experimentally. They are:

$$x = \sqrt[3]{k_c t} \qquad \text{Cubic law}$$

$$x = k_1 - k_2 \ln t \qquad \text{Logarithmic law}$$

$$1/x = k_1' - k_2' \ln t \qquad \text{Reciprocal logarithmic law}$$

$$x = \sqrt{k_p t} \qquad \text{Parabolic law}$$

where, in each equation, x is the thickness of the film, t is the time of reaction, and k_c, k_p, etc. are experimentally determined constants. In the case of thick layers, nominally over 100 nm, two laws have been observed:

$$x = kt \qquad \text{Linear}$$

$$x = \sqrt{k_p t} \qquad \text{Parabolic}$$

The constant k_p is called the *parabolic rate constant*. A linear rate is usually found when the film is porous or cracked while a parabolic equation is observed when the film forms a coherent, impenetrable layer. As the rate of film growth, dx/dt diminishes with time for parabolic rate law, this latter equation is associated with *protective kinetics*.

6.7.3 Oxide Film Microstructure

The initial step of an oxidation reaction is usually the adsorption of oxygen onto the metal surface. At this stage the adsorbate will consist of oxygen molecules that are weakly bound to the surface (Figure 6.33a). This is called *physical adsorption* or *physisorption*. On most metals, the oxygen molecules rapidly dissociate into oxygen atoms, and the resulting layer of atoms is more strongly bound to the metal surface (Figure 6.33b). This stage is called *chemical adsorption* or *chemisorption*.

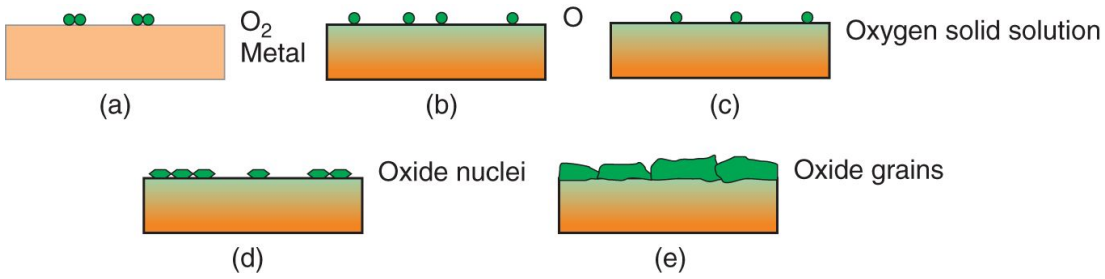


Figure 6.33 Oxide formation on a metal surface, schematic: (a) physical absorption of O_2; (b) dissociation of O_2 into separate atoms; (c) penetration of some atoms into the metal; (d) formation of oxide nuclei; (e) formation of surface film.

The oxidation begins by the diffusion of oxygen atoms into surface layers of the metal to form a dilute solid solution, while more molecules continue arrive at the surface and dissociate (Figure 6.33c). Saturation of the surface and subsurface occurs, leading to nucleation of chemically recognisable oxides. At high temperatures and low partial pressures of oxygen, this takes place at random (Figure 6.33d). The oxide formed might be the normal oxide, MgO, for example, on magnesium, or, if a number of oxides form, it might be a 'lower' oxide, FeO on iron, for example. Sometimes metastable oxides are also found that do not occur normally in the bulk. Continued growth proceeds so that the nuclei enlarge to form islands of oxide. Under some conditions, growth perpendicular to the surface might be much greater than lateral growth, in which case *whiskers* can form. Lateral growth is most frequent in dry conditions, and as this continues, the islands grow together to give a surface layer consisting of randomly oriented grains of oxide (Figure 6.33e). Thereafter, further oxidation requires transport of material across the oxide film.

The simplest way for this to happen is via cracks in the film or along porous grain boundaries. Cracks may appear during heating and cooling cycles if the thermal expansion coefficient of the oxide is very different to that of the parent metal. Another factor is the volume of oxide produced by oxidation of a given volume of metal. This is called the *Pilling–Bedworth* ratio, X_{PB}, which is the

molar volume of oxide formed divided by the molar volume of metal consumed. For the general oxidation:

$$mM + n/2\,O_2 \rightarrow M_mO_n$$

$$X_{PB} = m_o\,\rho_m / m\,m_m\,\rho_o$$

where m_o is the molar mass of the oxide M_mO_n, ρ_o the density of the oxide M_mO_n, m_m the molar mass of the metal, and ρ_m the density of the metal.

If the Pilling–Bedworth ratio is less than 1, the oxide cannot cover the metal completely and the oxide film has an open or porous structure. Oxidation takes place continuously, and the oxidation kinetics tends to be linear. This type of behaviour is found for the alkali and alkaline earth metals. In the rare cases where the Pilling–Bedworth ratio is equal to 1, a closed layer can form, which is stress free. When the Pilling–Bedworth ratio is greater than 1, a closed layer will form, but with a certain amount of internal compressive stress present.

The location of this stress depends upon the mechanism of the reaction, as discussed below. In cases where further formation of oxide is on the outer side of the layer, the stresses are easily relieved and the layer remains coherent. This results in protective oxidation with parabolic reaction kinetics. The oxide film is called *protective scale*. If the new oxide film forms between the metal and an outer oxide layer, the stresses cannot be easily relieved, the oxide layer can crack, and *spalling* (fragmentation) can occur. Spalling is also possible when the value of X_{PB} is significantly greater than one, because the additional volume that has to be accommodated generates stresses that lead to cracking. In general, cracks lead to faster oxidation, called *accelerating* or *breakaway* oxidation. The kinetics typical of breakaway oxidation initially follows a parabolic rate law, but this changes to a linear rate law when the film begins to fissure.

6.7.4 Film Growth via Diffusion

In cases where a continuous and coherent layer of oxide film is present, further reaction can only proceed by diffusion of some of the reactants across the film. In many solids, the passage of neutral atoms is less likely than the transport of ions. Perhaps the most obvious mechanism for oxide formation is the diffusion of M^{n+} cations outward from the metal towards the gas atmosphere (Figure 6.34a). If this occurs, a large negative charge would remain at the metal–metal oxide interface. This negative charge would act to slow the diffusing positively charged cations, and this would bring the reaction to a halt. To maintain electrical neutrality in the system and to allow the reaction to continue, cation diffusion must be accompanied by a parallel diffusion of an

appropriate number of electrons. In such cases overall charge neutrality needs to be maintained at all times, so that the transport of the ions and electrons remain balanced – a process called *ambipolar diffusion*. When the electrons arrive at the surface, they react with oxygen molecules arriving on the oxide surface to form O^{2-} ions. These are incorporated into the oxide film and together with the arriving M^{n+} cations allow the film to grow at the *outer surface* of the film. Note that film growth cannot continue if the oxide is an electrical insulator, typified by the oxidation of aluminium. The initial reaction of aluminium with oxygen is via cation and electron transport, and results in a thin continuous and well-attached film of oxide. However, this metal appears to be corrosion resistant because the aluminium oxide film is insulating and prevents reaction from continuing. As the thin film is also transparent, the metal remains shiny and gives the overall impression that of no corrosion has occurred.

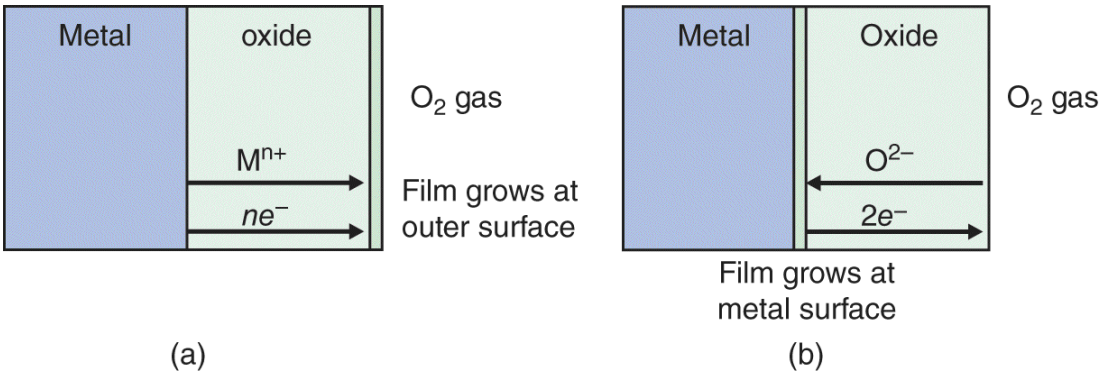


Figure 6.34 Growth of an oxide film on a metal surface: (a) diffusion of metal cations and electrons; (b) counter-diffusion of anions and electrons.

An alternative mechanism envisages that oxygen ions diffuse across the film from the outer surface towards the metal (Figure 6.34b). The oxygen ions cannot be generated spontaneously, and once again, it is necessary for electrons to move through the oxide layer from the metal to make the ionisation possible. This leaves M^{n+} cations behind at the metal–metal oxide boundary. These cations are able to combine with the arriving O^{2-} anions to extend the oxide film at the metal–metal oxide *inner boundary*. Film growth will again be curtailed if the film is an electrical insulator.

A third possibility, in which counter-diffusion of M^{n+} cations and O^{2-} anions occurs, is rare in oxide film formation, and will be discussed in detail in Section 6.8.

The kinetics of oxidation will depend upon the exact mechanism of the reaction. As an example, the oxidation of copper metal in a low partial pressure of oxygen is described. Under these conditions, the reaction produces Cu_2O by a mechanism involving diffusion of Cu^+ cations and electrons. The reaction is:

$$2Cu + \tfrac{1}{2} O_2 \rightarrow Cu_2O$$

The initial reaction results in the formation of a continuous film of oxide that is firmly attached to the metal surface. The rate of growth of the film is controlled by the slow diffusion of the Cu^+ ions compared with electron transport. The reaction rate can be derived from Fick's first law of diffusion:

$$J = -D\,(\mathrm{d}c/\mathrm{d}x)$$

where D is the diffusion coefficient of the slowest moving ion, c is the concentration of these ions, x is the position in the film, and J is the net flow of these ions through the film. If the film increases in thickness by Δx in time Δt, the rate of increase in the thickness of the film, $\mathrm{d}(\Delta x)/\mathrm{d}(\Delta t)$, will be proportional to the net flow of ions J:

$$\mathrm{d}(\Delta x)/\mathrm{d}(\Delta t) \propto J$$

The concentration gradient $(\mathrm{d}c/\mathrm{d}x)$ is given by the concentration of the ions at the inner and outer boundaries of the film divided by the film thickness $(c_1 - c_2)/x$ or $\Delta c/x$ hence:

$$\mathrm{d}(\Delta x)/\mathrm{d}(\Delta t) = k\,\Delta c/x$$

where k is a constant of proportionality that incorporates the diffusion coefficient of the ions. Integrating and rearranging this equation gives the parabolic rate law:

$$x^2 = k_\mathrm{p}\,t$$

where the constants have been incorporated into k_p, the parabolic rate constant. The units of k are the same as the units of the diffusion coefficient and generally:

$$k \propto D_\mathrm{A}$$

where D_A is the diffusion coefficient of the ionic species.

6.7.5 Alloys

The oxidation of alloys is complex and depends upon a variety of factors including the relative free energies of oxide formation of the components, the temperature, the oxygen pressure, the concentrations of the alloying elements, and the degree of oxygen solubility in the alloy. For a homogeneous alloy of several components A, B, C, and so on, it is found that one element, often

described as the *less noble* metal, will oxidise preferentially and form a film as described above, say, B_mO_n on the surface. If this film is protective the underlying alloy will be conserved; a situation that occurs in stainless steel, where a thin film of insulating Cr_2O_3 protects the iron alloy from continuous corrosion. Otherwise, the initial thin film will only continue to grow if the alloy component that is most easily oxidised is able to diffuse through the alloy to the thin film itself. If this is not possible, or if diffusion is slow, the other alloy components may begin to oxidise and generate a composite film consisting of, for instance, a layer of A_mO_n between the alloy and the outer B_mO_n film. In cases where there is significant oxygen solubility in the alloy, small precipitates of oxide B_mO_n may appear within the alloy, a phenomenon called *internal oxidation*. Although these complexities are commonplace, the oxidation kinetics often follow a parabolic rate law if diffusion is rate controlling.

6.8 Solid-State Reactions

The reactions of solids require an input of energy to transform a stable initial state into other products. This has led to reactions being loosely classified as thermochemical, those requiring heat energy input; electrochemical, those requiring electrical energy; photochemical, reactions triggered by light; and mechanochemical, reactions induced by the absorption of mechanical energy. Electrochemical reactions, which involve oxidation and reduction, are described in Chapter 7, particularly with respect to the operation of batteries, fuel cells, and electrolysis. In this section, the other three reaction pathways are outlined. As an example of a typical thermochemical reaction, the formation of oxide spinels, compounds of formula AB_2O_4, is described. Solid reactions triggered by light are described with respect to polymeric photoresists, widely used in the manufacture of electronic circuit boards. Mechanochemical reactions are illustrated with reference to the formation of new phases as a result of grinding or milling of powders.

6.8.1 Spinel Formation

Reactions between two solids generally take place at high temperatures, and are typical of thermochemical processes. In the main, following the initial reaction, the product separates the reactants. Further reaction is dependent upon the transport of material across this barrier. Cracking, porosity, and volume mismatch can all help in this. In this section, the case when a coherent layer forms between the two reactants will be considered. The *mechanism* of the reaction may depend upon whether electron transport is possible in the intermediate phase, while the *rate* of reaction will be controlled by the rate of diffusion of the slowest species at the reaction temperature. To illustrate the problems encountered a typical solid-state reaction, the formation of oxide spinels is described.

Spinel itself is a mineral with a composition $MgAl_2O_4$. The spinel formation reaction can be represented by the chemical equation:

$$MgO + Al_2O_3 \rightarrow MgAl_2O_4$$

Suppose that a crystal of Al_2O_3 is placed in close contact with a crystal of MgO to form a *reaction couple*. Initial reaction will result in the separation of the MgO and Al_2O_3 by a thin layer of $MgAl_2O_4$. Continued reaction will depend upon transport of reactants across the spinel layer. As in the case of metal oxidation, a number of mechanisms can be suggested, but because $MgAl_2O_4$ is an insulator, electron transport is not possible, and so only mechanisms involving ions are permitted.

One such mechanism requires counter-diffusion of equal numbers O^{2-} anions and Mg^{2+} cations (Figure 6.35a,b). The electrical charges on the ions are equal and opposite so no charge-balance problems arise. New spinel growth will take place at the Al_2O_3 interface. Alternatively, the diffusion of O^{2-} anions accompanied by an parallel diffusion of Al^{3+} cations might occur (Figure 6.35c). Because of the difference in the ionic charges, the diffusion of two Al^{3+} cations have to be balanced by the transport of three O^{2-} anions to maintain charge neutrality. Spinel growth will now take place at the MgO boundary. The counter-diffusion of Mg^{2+} and Al^{3+} is also possible (Figure 6.35d). To maintain charge neutrality, the diffusion of three Mg^{2+} cations must be balanced by the diffusion of two Al^{3+} cations. In this case, the spinel layer forms on either side of the initial boundary.

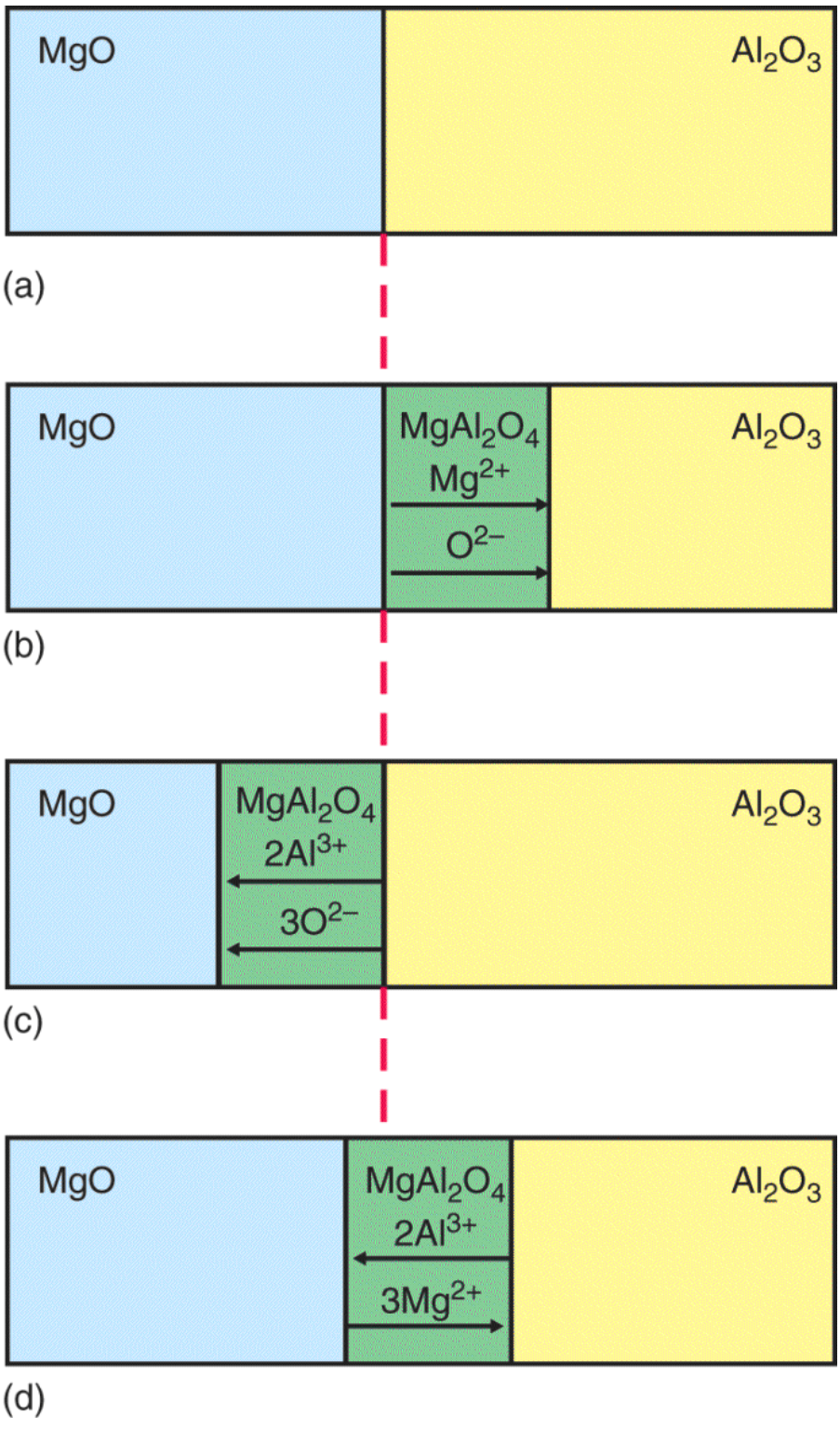


Figure 6.35 Mechanisms of spinel formation: (a) initial state; (b) diffusion of Mg^{2+} and O^{2-} ions; (c) diffusion of Al^{3+} and O^{2-} ions; (d) counter-diffusion of Mg^{2+} and Al^{3+} ions.

It has been found that the reaction between MgO and Al_2O_3 follows this latter mechanism. The reactions at the boundary between Al_2O_3 and spinel are:

$$Al_2O_3 \rightarrow 2Al^{3+} + 3O^{2-}$$

$$3Mg^{2+} + 3O^{2-} + 3Al_2O_3 \rightarrow 3MgAl_2O_4$$

The reactions at the boundary between MgO and spinel are:

$$3MgO \rightarrow 3Mg^{2+} + 3O^{2-}$$

$$3O^{2-} + 2Al^{3+} + MgO \rightarrow MgAl_2O_4$$

These equations indicate that the spinel layer grows in an asymmetrical fashion. For every three Mg^{2+} ions that arrive at the Al_2O_3 boundary, three $MgAl_2O_4$ units result, while for every two Al^{3+} ions which arrive at the MgO boundary, only one $MgAl_2O_4$ unit forms. The mechanism can be checked by placing inert markers at the initial boundary. In this case the markers will be buried in the spinel layer which will form in a ratio of 1 : 3 on either side of the markers, with the thicker part on the Al_2O_3 side (Figure 6.36). In general, the ratio of the layer thickness on each side of the boundary will depend upon the charges on the counter-diffusing cations and marker position is a sensitive indicator of the spinel formation mechanism.

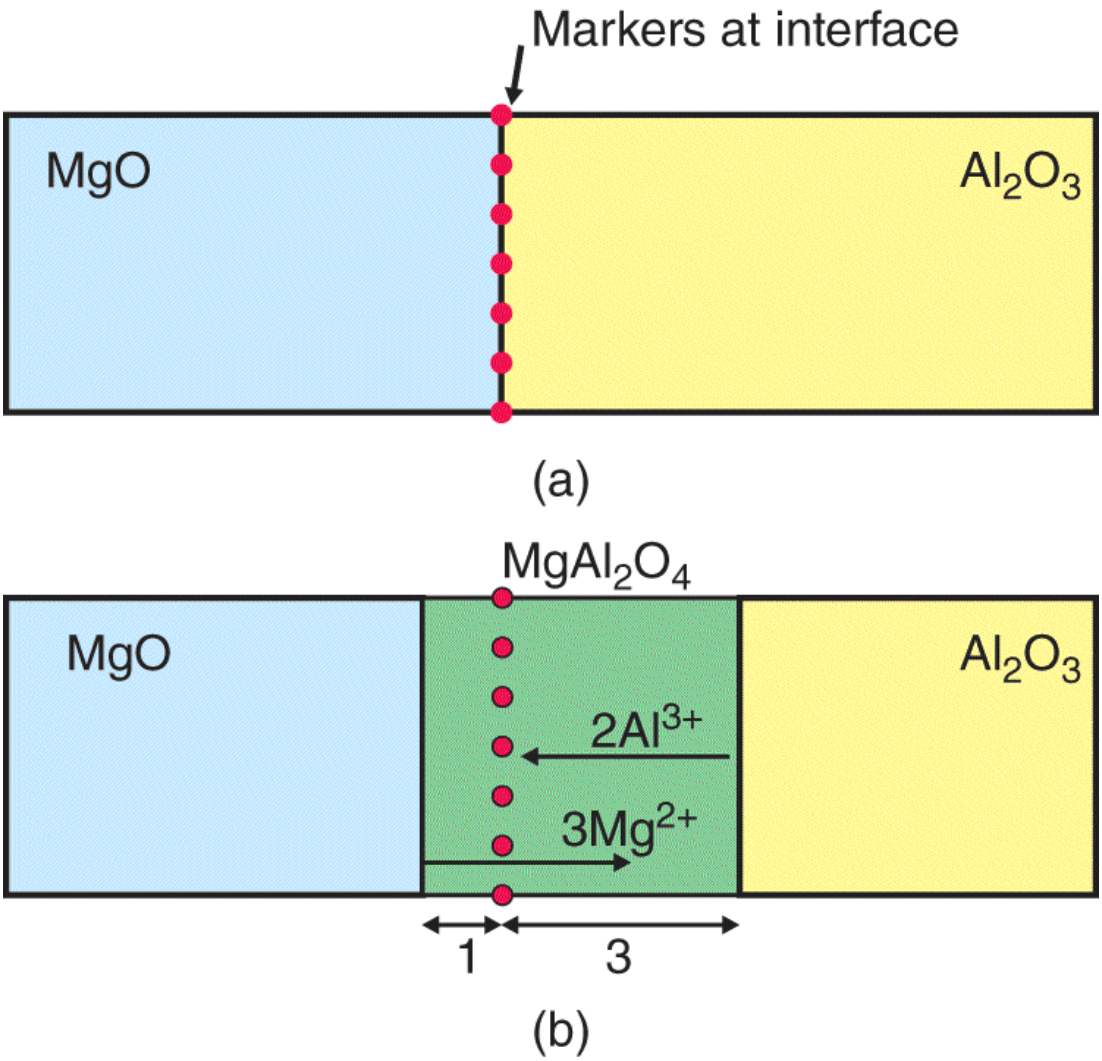


Figure 6.36 Markers used to determine the mechanism of $MgAl_2O_4$ formation: (a) inert markers at original interface; (b) markers after reaction.

The rate at which the total thickness of the spinel layer grows is controlled by the speed of diffusion of the slowest cation and follows a typical parabolic rate law:

$$x^2 = 2k_p t$$

where x is the thickness of the spinel layer, k_p is called the *practical reaction rate constant* or *parabolic rate constant*, and t is the reaction time. Many solid-state reactions give a parabolic rate law for the growth of an internal phase and such a parabolic rate law is taken as evidence that the reaction is diffusion controlled. As before, one can write:

$$k \propto D_A$$

where D_A is the diffusion coefficient of the ionic species that diffuses at the slowest speed.

6.8.2 Photoresists

There are large numbers of reactions that are driven by light irradiation in its various forms, including many important processes that occur in the natural world, such as photosynthesis. The complex metal and ceramic parts produced by 3D printing (Section 6.1.4) use light to control both the physical extent of the reaction and to induce the sintering reactions that produce a solid body. Here a group of reactions in which light interaction with solids is central to the production of electronic components, photoresists, is outlined.

From the middle of the twentieth century, microelectronics has evolved largely as a result of the ability of manufacturers to add more and more complex electronic circuitry to computer memory chips and components, as well as the fabrication of more sophisticated microprocessors. The key to this advance has been the use of photoresists to achieve ever-increasing densities of electronic components onto silicon wafers.

Photoresists are light sensitive polymers or polymer precursors that undergo reaction when exposed to light. These are coated as a thin film on the electronically important substrate, such as a silicon chip (Figure 6.37). The exposure is made through a mask, so that only selected parts of the photoresist are irradiated. The detail on the final chip is thus defined by the mask and by the wavelength of the irradiating light. Smaller details become achievable as the wavelength of the light decreases from infra-red, wavelengths greater than approximately 750 nm, to ultraviolet, with wavelengths much shorter than 400 nm. Commercially very short wavelength ultraviolet light or electron beam radiation is used to give the degree of resolution needed for high-density electronic circuits.

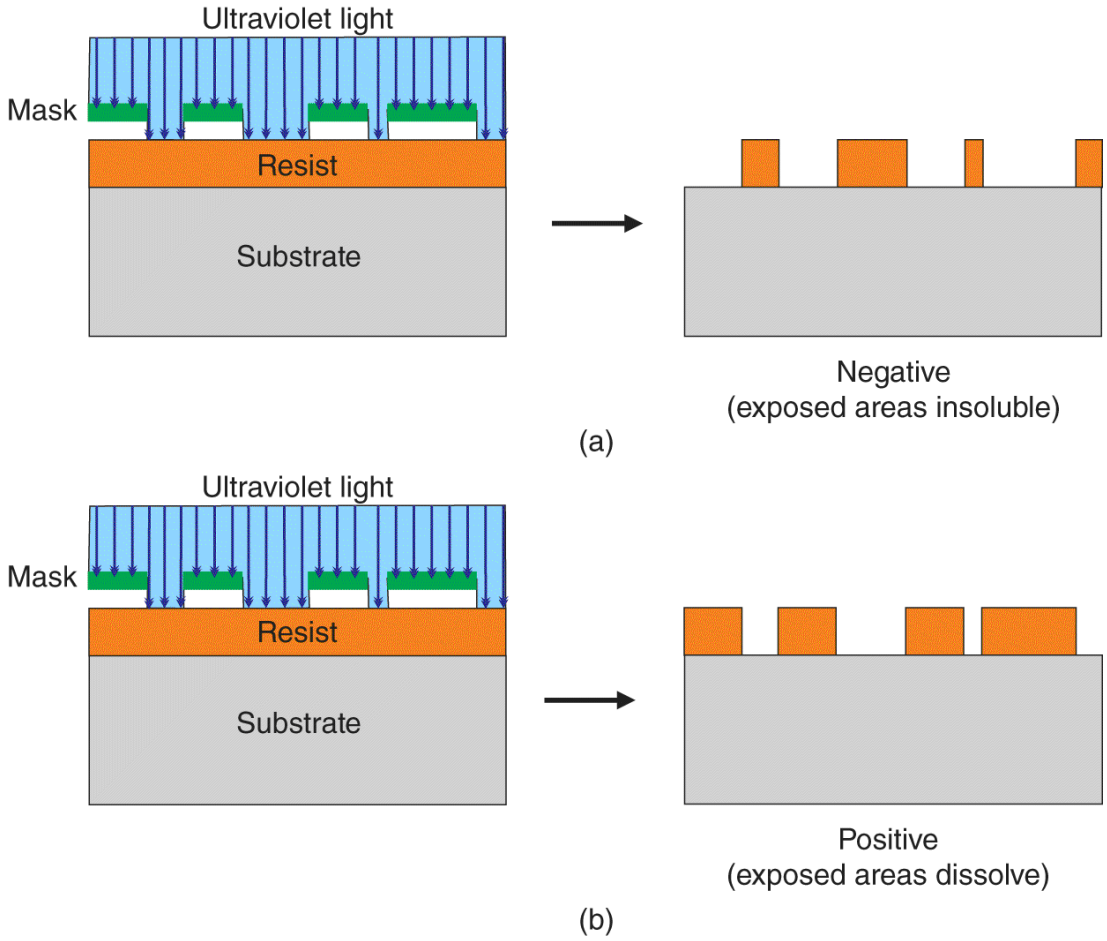


Figure 6.37 Photoresists, schematic: (a) negative photoresist; (b) positive photoresist.

There are large numbers of photoresists commercially available, and most consist of a variety of additives and sensitisers, blended with solvents, mixed with the primary reactants, so as to achieve reproducible results. Although the precise mechanism of operation of many of these materials is not known in detail, they fall into three broad categories: materials that polymerise on exposure to the appropriate wavelength of light, materials that decompose on exposure to the appropriate wavelength of light, and materials that undergo cross-linking on exposure to the appropriate wavelength of light.

Photoresists that function by polymerisation of a monomer produce an insoluble material in those regions that have been exposed to the light. Following exposure, the surface film is treated with a solvent which removes all of the unreacted photoresist. The pattern of the polymer that remains is an image of the light pattern that fell onto the substrate, and such materials are called *negative photoresists* (Figure 6.37a). A number of other negative photoresists are known that behave in a similar fashion, to produce a solvent

resistant pattern on light irradiation, but the molecules in the photoresist form cross-links rather than polymerise.

The opposite behaviour is shown by *positive photoresists*. In these materials, the regions exposed to light dissolve when treated with a solvent (Figure 6.37b). Positive photoresists initially form a resinous film on application to the substrate. Light breaks the bonds in this material or in an additive, which generates molecular fragments that are susceptible to dissolution. Treatment with an appropriate solvent washes away these exposed areas, leaving the original coating patterned by the light, but in the opposite way to that of the negative photoresists.

The precise mechanisms by which many commercial photoresists function are not known at a molecular level. Negative photoresists are often molecules containing a double bond joining two C atoms. Light energy, absorbed by the molecules, breaks the double bond to give reactive free radicals (Section 4.4). These unite to form a polymer in the irradiated regions. An example of this type of reaction is given by the transformation of methyl methacrylate, CH_2□$C(CH_3)COOCH_3$, into poly(methyl methacrylate), known by the acronym PMMA, by the opening of the C□C double bond as a result of light irradiation. Positive photoresists tend to consist of a polymeric resinous phase together with an unstable additive. Light irradiation causes the additive to decompose into fragments, which weaken the film chemically. Subsequent treatment by a solvent removes these weakened areas, thus removing those regions that have been irradiated.

6.8.3 Mechanochemistry

Mechanochemistry describes the instigation of chemical reactions via a mechanical energy input, generally milling or grinding. In the commonest technique, ball-milling, the reactants are placed in an inert container, which may be steel, a rigid plastic or an inorganic oxide such as alumina or agate, together with inert balls of (usually) the same material. The container is then vigorously shaken or spun so that the balls smash into the contents of the container. When no chemical reaction takes place, the contents are reduced to a fine powder and this method has long been used to prepare homogeneous powders with particle sizes down to nanometers, of, for example, oxides such as alumina. In such small particles, the outer shell is often amorphous, while the inner core of the particle retains a crystalline structure. In other cases, the repeated impact of the balls engenders a chemical reaction between the contents of the mill. As such, mechanochemistry has been widely used in the instigation of many reactions that are not readily carried out using conventional chemical techniques, including both main group inorganic chemistry and organic and organometallic reactions.

Apart from dry milling, which may be advantageous in some cases where solvent addition is not desirable, the technique allows for the addition of other materials, generally in small quantities, into the milling chamber. Thus, liquid assisted grinding (which is essentially derived from kneading bread) is widely used, as is polymer assisted grinding, in which a small quantity of a polymer is included, catalyst assisted grinding, in which small quantities of a metal catalyst is added. In addition, if the mill is transparent, light from LEDs can be shone onto the contents during milling to promulgate reactions that are triggered by light. Apart from these modifications, reaction paths can be influenced by the milling time, whether the material is milled once only or several times, and the overall temperature of the system.

Carbon nanotubes and fullerenes are not readily incorporated into solvents and reactions between them are difficult to achieve using normal chemical techniques. Milling, however, does instigate reactions that are not readily duplicated otherwise. For example, milling C_{60} will result in the dimer C_{120} (Figure 6.38), a material that can be fabricated, up to the present, *only* by mechanochemical synthesis. In outline, C_{60} is placed in a stainless-steel mill with stainless-steel balls and a small amount of a catalyst, typically KCN or KOH, and is milled for 30 minutes under a nitrogen atmosphere. The result is a conversion of approximately 20–30% of the C_{60} to the dimeric C_{120}.

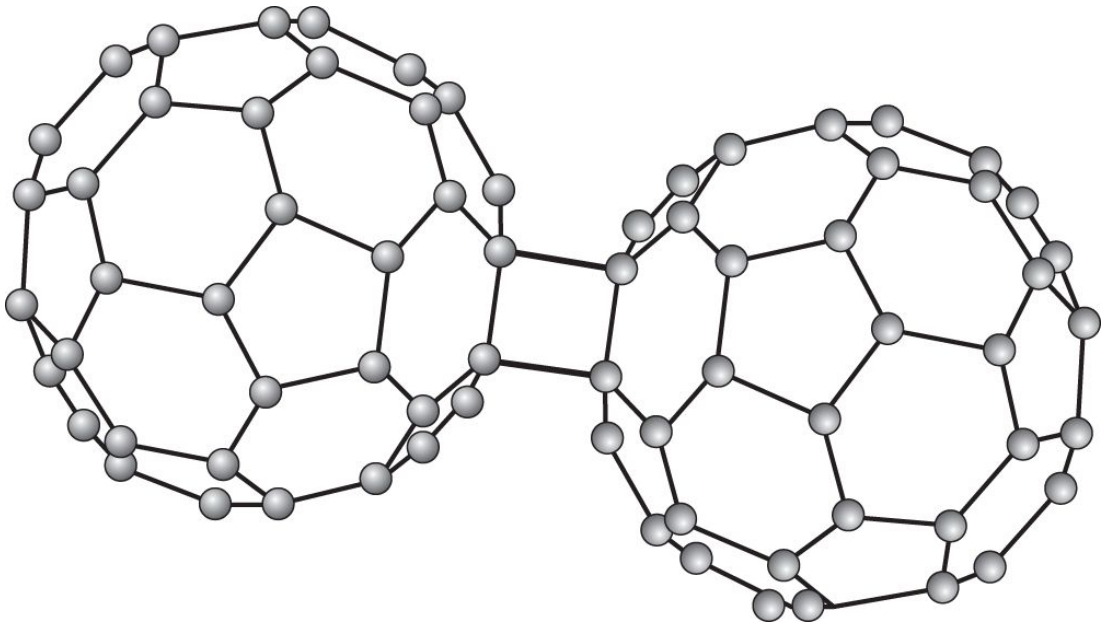


Figure 6.38 C_{120} dimers produced by milling C_{60} monomers.

A number of oxide reactions that cannot easily be obtained using thermochemical reactions have also been induced via mechanical milling. For example, the spinel phases produced by heating described earlier invariably have a normal or an inverse cation distribution (Section 3.2.2.8). The phase

$(Zn)[Al_2]O_4$ is a normal spinel when prepared by thermally initiated reaction between ZnO and Al_2O_3, as described for the formation of $MgAl_2O_4$ previously, but milling the normal spinel for a short time will transform the structure to a metastable form $(Zn_{0.88}Al_{0.12})[Zn_{0.12}Al_{1.88}]O_4$, with a degree of inversion, λ of 0.12. Similarly, milling the normal spinel $(Zn)[Fe_2]O_4$ produces the metastable partly inverse $(Zn_{0.59}Fe_{0.41})[Zn_{0.41}Fe_{1.59}]O_4$., with a degree of inversion, λ of 0.41 Note that this change in cation distribution will also affect the magnetic properties of the phase.

Solid-state reactions including spinel formation can also be achieved by milling at essentially room temperature. An example if given by the reaction of ZnO and SnO_2 to form zinc stannate, Zn_2SnO_4:

$$2ZnO + SnO_2 \rightarrow Zn_2SnO_4$$

This normally high-temperature reaction can be carried out at room temperature by milling. Milling for only two hours induces the formation of an excess of zinc stannate and the reaction is complete after approximately four hours at room temperature.

Further Reading

Sintering and 3D Printing:

Kang, S.-J.L. (2005). *Sintering, Densification, Grain Growth and Microstructure*. Oxford: Elsevier Butterworth-Heinemann.

Khairallah, S.A., Martin, A.A., Lee, J.R.I. et al. (2020). Controlling interdependent meso-nanosecond dynamics and defect generation in metal 3D printing. *Science***368**: 660.

Wang, C., Ping, W., Bai, Q. et al. (2020). A general method to synthesize and sinter bulk ceramics in seconds. *Science* **368**: 521.

Lars Röntzsch. Computer simulations of sphere to sphere and sphere to plate sintering are to be found at www.roentzsch.org.

High Temperature Oxidation and Solid-State Reactions:

Khanna, K.S. (2002). *High Temperature Oxidation and Corrosion*. Materials Park, OH: ASM International.

Young, D. (2008). *High Temperature Oxidation and Corrosion of Metals*. Oxford: Elsevier.

For Mechanochemistry See:

Friščič, T., Mottlilb, C., and Titi, H.M. (2020). Mechanochemistry: mechanochemistry for synthesis. *Angew. Chem. Int. Ed.***59**: 1018.

Šepelák, V., Düvel, A., Wilkening, M. et al. (2013). Mechanochemical reactions and synthesis of oxides. *Chem. Soc. Rev.***42**: 7507.

Zhu, S.-E., Li, F., and Wang, G.-W. (2013). Mechanochemistry of fullerenes and related materials. *Chem. Soc. Rev.***42**: 7535.

Problems and Exercises

Quick Quiz

6.1 The transformation of a compacted powder into a solid by heating is called:

a. Annealing.

b. Tempering.

c. Sintering.

6.2 Liquid phase sintering involves using:

a. Only liquids for the reaction.

b. Solids mixed with liquids for the reaction.

c. An additive that melts during the reaction.

6.3 The driving force for sintering is:

a. Reduction in surface area.

b. Reduction in total volume.

c. Reduction in Gibbs energy of reaction.

6.4 During slow cooling of an Ni–Cu alloy, the composition of the crystallites:

a. Is constant.

b. Varies continuously.

c. Has the composition of the liquid.

6.5 When an Ni–Cu alloy is cooled fairly rapidly, the core of the crystallites will have a composition:

a. Richer in the lower melting point parent phase.

b. Richer in the higher melting point parent phase.

c. Identical to that of the liquid phase.

6.6 When a liquid alloy of lead and tin is cooled through a eutectic point, the microstructure of the solid produced contains:

a. Eutectic α and eutectic β.

b. Primary α and eutectic β.

c. Eutectic α and primary β.

6.7 When a liquid alloy of lead and tin, with a composition between the eutectic point and the α-phase is cooled, the microstructure of the solid produced contains:

a. Primary α and eutectic β.

b. Primary α, eutectic α, and primary β.

c. Primary α, eutectic α, and eutectic β.

6.8 When austenite is cooled slowly through the eutectoid point, the material that forms is called:

a. Cementite.

b. Ferrite.

c. Pearlite.

6.9 The microstructure of pearlite is composed of:

a. Eutectoid ferrite plus austenite.

b. Eutectoid ferrite plus eutectoid cementite.

c. Eutectoid cementite plus austenite.

6.10 A hypoeutectoid steel has a composition that lies to the:

a. Iron-rich side of the eutectoid.

b. Carbon-rich side of the eutectoid.

c. Carbon-rich side of austenite.

6.11 Slow cooling of a hypereutectoid steel composition produces a microstructure composed of:

a. Proeutectoid ferrite plus eutectoid cementite plus eutectoid ferrite.

b. Proeutectoid cementite plus proeutectoid ferrite plus pearlite.

c. Proeutectoid cementite plus eutectoid cementite plus eutectoid ferrite.

6.12 Martensite is produced from austenite by:

a. Quenching the austenite.

b. Annealing the austenite.

c. Tempering the austenite.

6.13 A martensitic transformation is described as a:

a. Reconstructive transformation.

b. Diffusionless transformation.

c. Equilibrium transformation.

6.14 Shape-memory alloys utilise:

a. Martensitic transformations.

b. Eutectoid transformations.

c. Reconstructive transformations.

6.15 The shape-memory effect in shape-memory alloys requires that the solid contain:

a. Fixed twin boundaries.

b. Mobile twin boundaries.

c. No twin boundaries.

6.16 The shape-memory alloy Nitinol has an approximate formula:

a. $NiTi_3$.

b. NiTi.

c. Ni_3Ti.

6.17 Direct corrosion of a metal requires the presence of:

a. A gas plus water.

b. A dry gas.

c. Air.

6.18 The parabolic rate law for oxidation arises when:

a. The oxide film is cracked.

b. The oxide film is thin.

c. Diffusion controls the reaction.

6.19 The Pilling–Bedworth ratio is used to predict the:

a. Likelihood of corrosion of a metal.

b. Reactivity of a metal with a gas.

c. Rate of corrosion of a metal.

6.20 The formation of spinel, $MgAl_2O_4$, involves the:

a. Diffusion of Mg^{2+} and O^{2-} ions.

b. Diffusion of Al^{3+} and O^{2-} ions.

c. Diffusion of Mg^{2+} and Al^{3+} ions.

6.21 A parabolic rate constant is characteristic of a reaction between two solids, which is controlled by:

a. Sintering.

b. Ionic diffusion.

c. Vapour transport.

6.22 Rapid prototyping (3D printing) utilises:

a. Rapid heating in a micro-furnace.

b. Spraying liquid droplets from a heated nozzle.

c. Laser heating of a powder layer.

6.23 A negative photoresist is:

a. Weakened on exposure to light.

b. Polymerised on exposure to light.

c. Removed by exposure to light.

Calculations and Questions

6.1 The surface energy of solid Al_2O_3, at 1850 °C is 0.905 J m^{-2} and the density is 3970 kg mol^{-1}. Calculate the relative pressure, p/p_0, over a hemispherical 'hill', of diameter 1×10^{-7} m, on the surface of a flat plate of corundum, at this temperature.

6.2 The surface energy of solid MgO, at 1500 °C is 1.2 J m^{-2} and the density is 3580 kg mol^{-1}. Estimate the pressure differential between a spherical pit of diameter of 5×10^{-7} m and that over the surface of a flat plate of magnesia, at this temperature.

6.3 Assuming that the ideal gas law ($pV = nRT$) holds, deduce the formula

$$\frac{r_i}{r_f} = \sqrt{\frac{2\gamma}{p_i r_i}}$$

where r_i and r_f are the initial and final radii of the pore in the material, p_i is the initial gas pressure in the pore, and γ is the surface energy of the material.

6.4 A spherical pore in a soda-lime glass at 1200 K contains trapped gas. Assuming the gas trapped in the pore is at atmospheric pressure, and the surface energy of the glass at 1200 K is 0.350 J m^{-2}, calculate the

equilibrium pore size when the initial pore is of diameter (a) 0.4 μm; (b) 4 μm; (c) 40 μm.

6.5 Using the data in Question 6.4, what is the equilibrium size of a pore that will neither shrink nor expand?

6.6 Solid titanium nitride, TiN, has a surface energy of 1.19 J m^{-2} at 1200 °C. If the voids in this material contain nitrogen gas at the same partial pressure as found in the atmosphere, approximately 8×10^4 Pa, estimate the maximum void size to ensure that voids in a sintered ceramic will shrink.

6.7 The gold–silver system forms a complete solid solution. The melting point of gold is 1064.43 °C and that of silver is 961.93 °C

a. During rapid cooling of a 50 at.% gold:50 at.% silver mixture, which phase will be richest in the core of a grain?

b. Sketch the microstructure of the solid at a temperature of 800 °C if the melt is cooled very slowly.

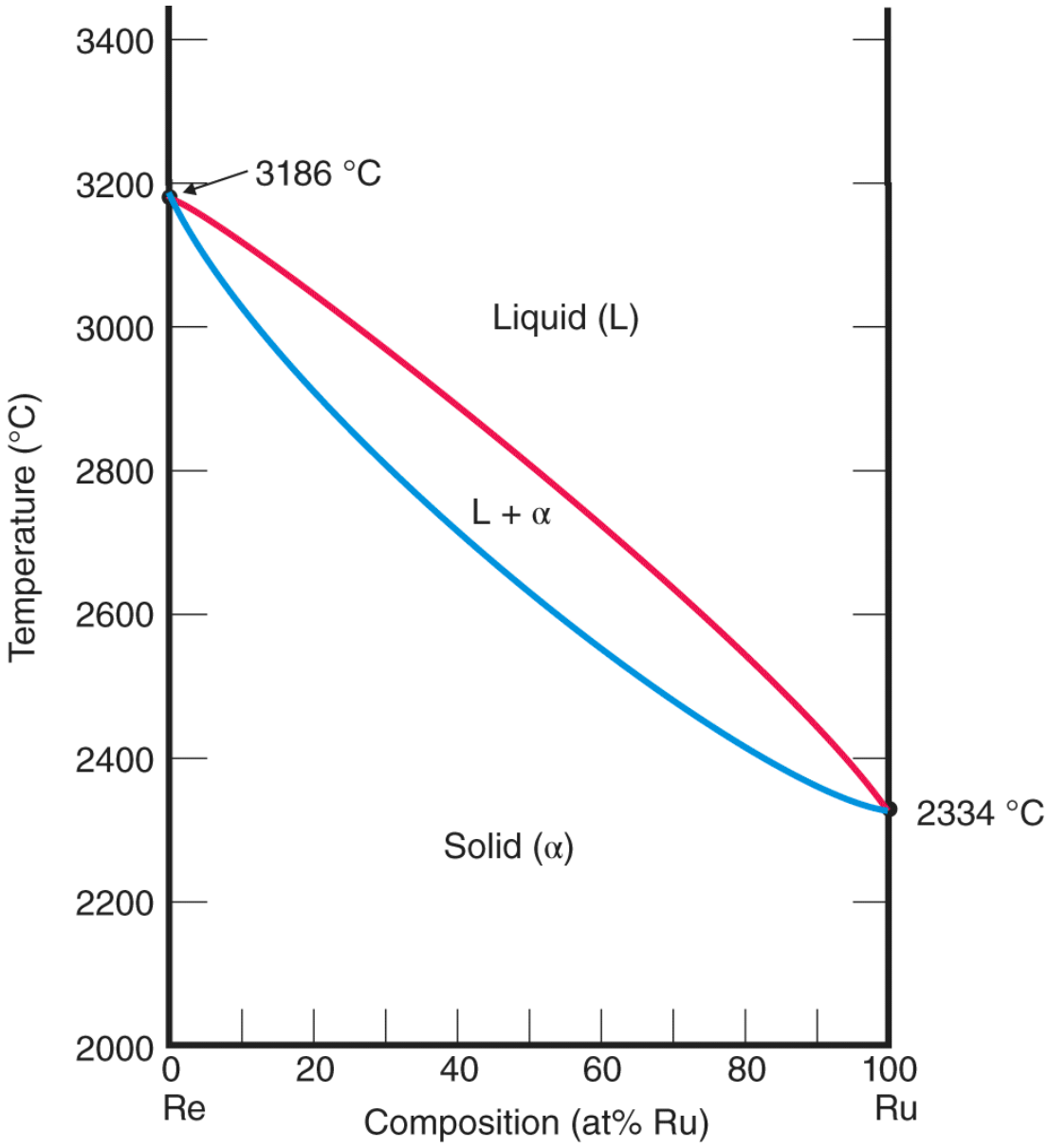
c. Sketch the microstructure of the solid at a temperature of 800 °C if the melt is cooled quickly.

6.8 The ruthenium–rhenium system forms a complete solid solution (see figure). The melting point of ruthenium is 2334 °C and that of rhenium is 3186 °C.

a. During rapid cooling of a 45 at.% Re:55 at.% Ru mixture, which phase will be richest in the core of a grain?

b. Sketch the microstructure of the solid at a temperature of 2000 °C if the melt is cooled very slowly.

c. Sketch the microstructure of the solid at a temperature of 2000 °C if the melt is cooled quickly.

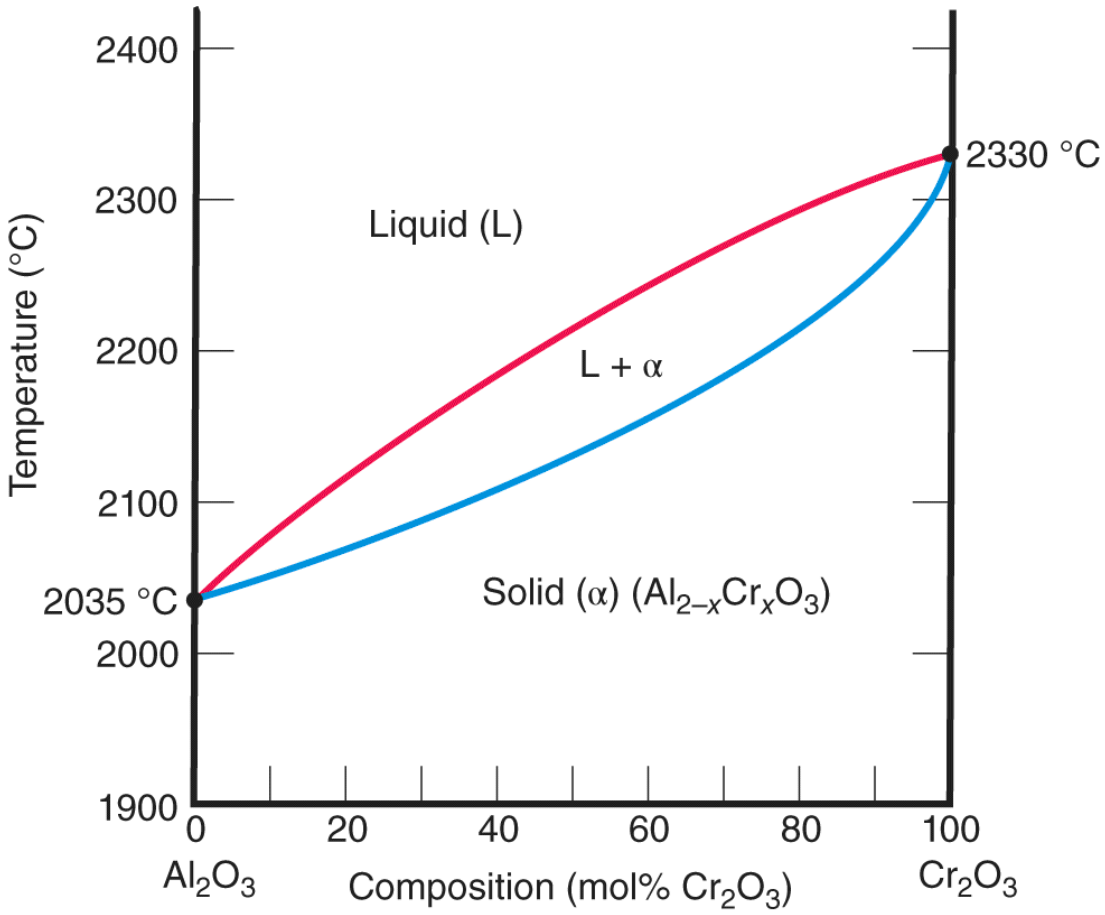


6.9 The Al_2O_3–Cr_2O_3 system forms a complete solid solution (see figure). The melting point of Al_2O_3 is 2035 °C and that of Cr_2O_3 is 2330 °C.

a. During rapid cooling of a 33 mol% Al_2O_3:67 mol% Cr_2O_3 mixture, which phase will be richest in the core of a grain?

b. Sketch the microstructure of the solid at a temperature of 2000 °C if the melt is cooled very slowly.

c. Sketch the microstructure of the solid at a temperature of 2000 °C if the melt is cooled quickly.

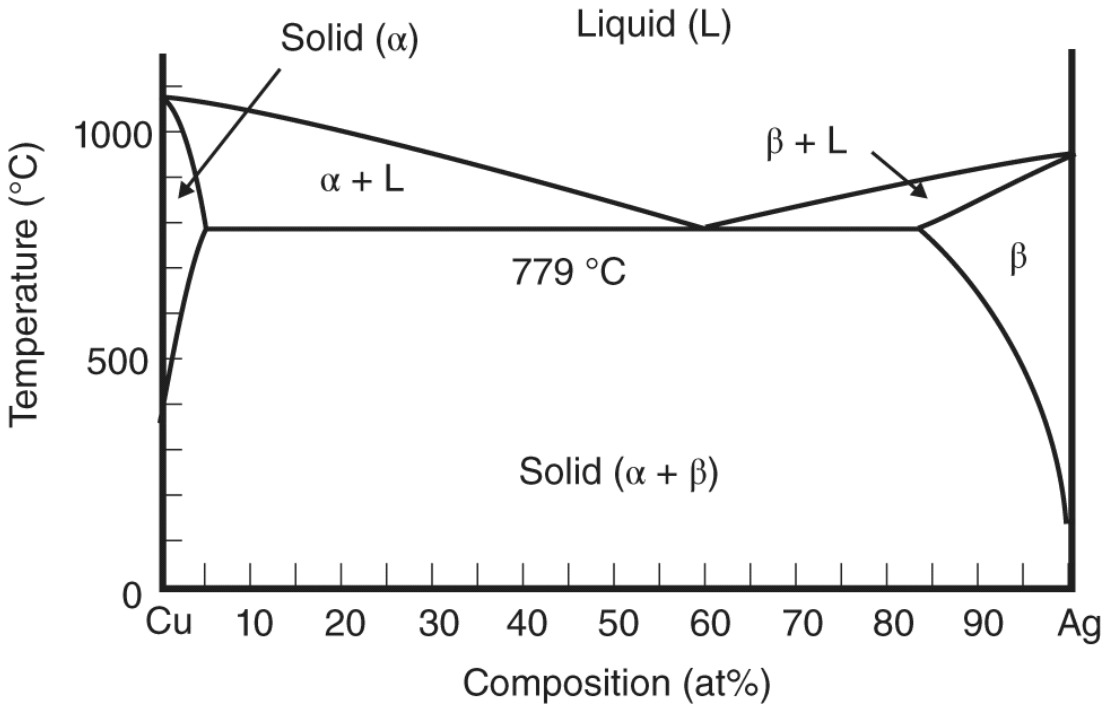


6.10 With respect to the phase diagram of the copper–silver system:

a. Sketch and label the microstructure of a solid containing 2.5 at.% Ag (about half way across the α phase field at 800 °C), when the melt is slowly cooled to 800 °C, and then to 400 °C.

b. Sketch and label the microstructure of a solid containing 30 at.% Ag when the melt is slowly cooled to 700 °C.

c. Sketch and label the microstructure of a solid formed by slowly cooling the eutectic composition containing 60 at.% Ag to 700 °C.

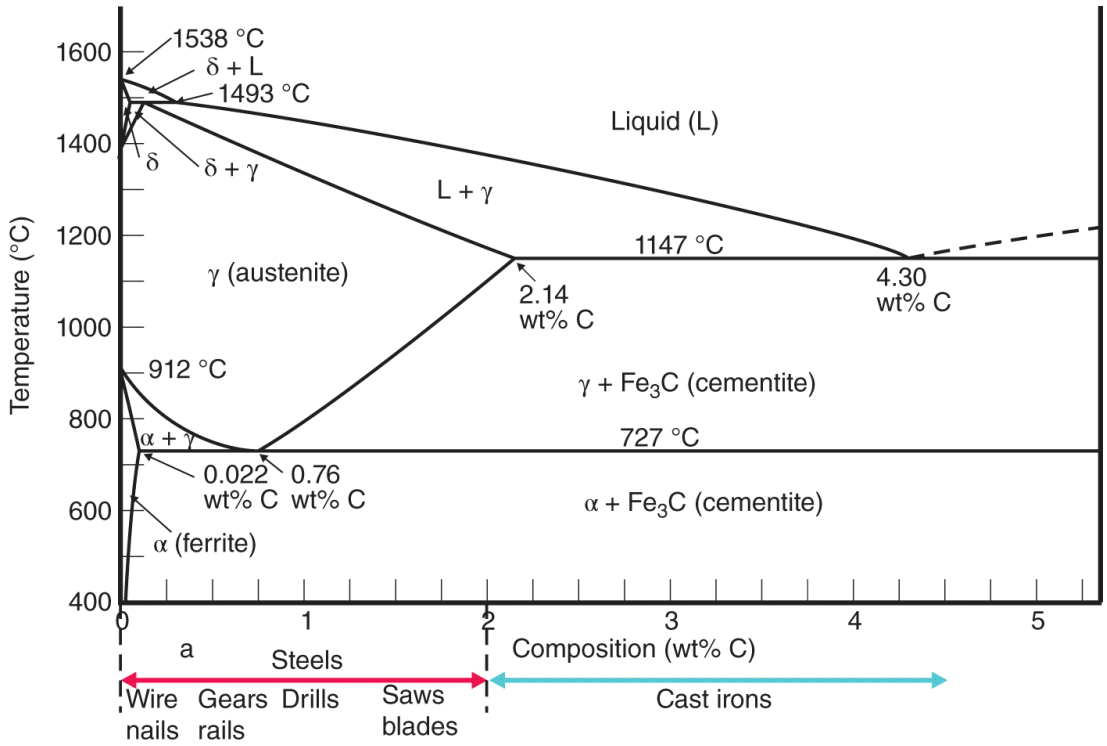


6.11 With respect to the following iron–carbon phase diagram:

a. Sketch and label the microstructure of a solid containing 3.44 wt% C when the melt is slowly cooled to 800 °C. What phases will occur if the sample is further cooled to 600 °C?

b. Sketch and label the microstructure of a solid containing 6.2 wt% C when the melt is slowly cooled to 800 °C.

c. The alloy in (b) is further cooled to 600 °C. What phases will now be present? How much of each is present?



6.12 With respect to the iron–carbon phase diagram (see Question 6.11):

a. Sketch and label the microstructure of a solid containing 1.0 wt% C when the melt is slowly cooled to 1000 °C, and then to 750 °C and finally to 700 °C.

b. Sketch and label the microstructure of a solid containing 0.3 wt% C when the melt is slowly cooled to 1000 °C, and then to 750 °C and finally to 700 °C.

c. Sketch and label the microstructure of a solid containing 0.76 wt% C when the melt is slowly cooled to 1000 °C, and then to 750 °C and finally to 700 °C.

6.13 Which of the metals in the table are likely to be protected from corrosion by the formation of a protective oxide film.

Metal/metal oxide	Density of metal (kg m^{-3})	Density of oxide (kg m^{-3})
Cu/CuO	8933	6315
Fe/Fe_2O_3	7873	5240
K/K_2O	862	2320
Ti/TiO_2	4508	4260
Al/Al_2O_3	2698	3970

Metal/metal oxide	Density of metal (kg m^{-3})	Density of oxide (kg m^{-3})
Na/Na_2O	966	2270

6.14 A nickel foil is oxidised at 1000 °C. The film thickness as a function of time is given in the table. Confirm that the rate is parabolic and calculate the parabolic rate constant.

Heating time (hr)	Film thickness (m)
1	4.74×10^{-7}
2	0.67×10^{-6}
3	0.82×10^{-6}
4	0.95×10^{-6}
5	1.06×10^{-6}
6	1.16×10^{-6}

6.15 When copper is oxidised under low partial pressures of oxygen gas, it forms Cu_2O via a parabolic rate law. The rate constant is 5.38×10^{-10} m^2 s^{-1} at 0.05 atm pressure, and 900 °C.

a. What will the film thickness be after oxidation of copper foil for 10 hours at this temperature?

b. What will the weight of the copper oxide film be?

c. Experimentally, it is easier to measure the weight gain as a function of time rather than film thickness. What will the weight gain of the film be?

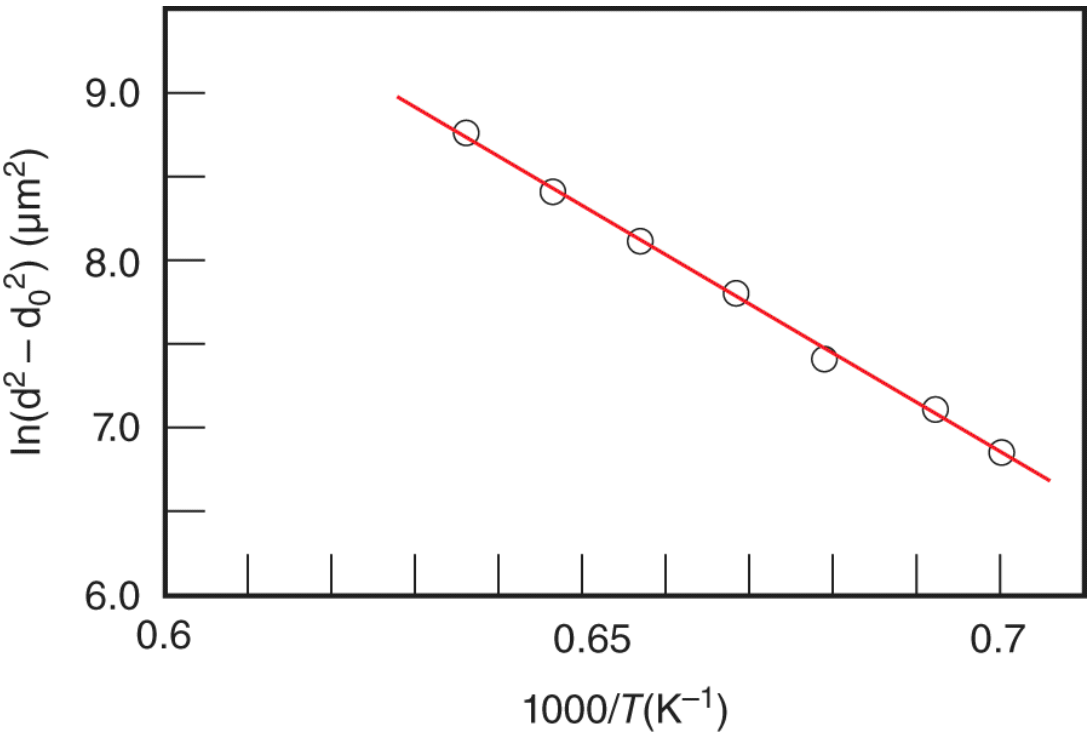
6.16 The thickness of the layer of the spinel $NiAl_2O_4$ formed between NiO and Al_2O_3 when reacted at 1350 °C is given in the table. Check that the reaction is diffusion controlled and calculate the rate constant. The reaction equation is:

$$NiO + Al_2O_3 \rightarrow NiAl_2O_4$$

Layer thickness (μm)	Time (hr)
1.0	20
1.4	40
1.8	60
2.0	80
2.3	100

6.17 The grain diameter of a ceramic sintered at higher temperatures, where normal sintering is predominant, is given in the table. The initial grain size was measured as 5 μm and the sintering time in each case was one hour. Estimate the activation energy for this sintering reaction.

Mean grain diameter (μm)	**Sintering temperature (°C)**
30	1150
35	1175
41	1200
50	1225
58	1250
68	1275
80	1300



7
Oxidation and Reduction

- How does a lithium battery work?
- How do pH meters work?
- What causes a metal to corrode?

Oxidation and reduction involve the transfer of electrons from one component to another. *Oxidation* is equivalent *to electron loss* and *reduction* is equivalent to *electron gain*. Reactions involving oxidation and reduction are called *redox reactions*. An early application of a redox reaction was the construction of the first battery, reported by Volta in 1800. One of the major problems encountered with Volta's device was severe corrosion of the metals used. Thus, it is apparent that batteries and corrosion are closely linked, and indeed, both are oxidation and reduction reactions. Redox reactions are also involved in a wide variety of other processes, including electroplating, metal refining, and in analytical tools such as pH meters. Understanding basic redox chemistry is vital for many modern materials applications. Oxidation and reduction reactions also underpin all life processes, although this aspect is not covered here.

7.1 Galvanic Cells

7.1.1 Cell Basics

A *galvanic cell* is an electrochemical cell that uses a spontaneous chemical reaction to produce an external electric current. Electrochemical cells consist of two *electrodes* (an *anode* and a *cathode*) in contact with an electrolyte that is able to conduct ions but not electrons (Figure 7.1). Galvanic cells are called *batteries* in colloquial speech. In all batteries, oxidation occurs at the *anode* of the cell; the electrode removes electrons from the species in the electrolyte. It is given a negative sign and consists of a relatively easily oxidised metal such as zinc, cadmium, or nickel, sometimes in contact with a *current collector* such a graphite rod. At the *cathode* of a battery reduction occurs; the electrode gives electrons to species in the electrolyte. The cathode is marked as positive and consists of a relatively easily reduced component such as MnO_2 or PbO_2, sometimes in contact with a metallic current collector. *Electrons* flow from anode to cathode (negative to positive) via an *external circuit* and are used to do useful work. *Anions* (negative ions) travel towards the anode through the electrolyte. *Cations* (positive ions) travel towards the cathode through the electrolyte. The driving force for these transformations is the energy of the cell reaction, which stops when the external circuit is broken. Batteries are spent, or need recharging, when one or both of the components have been used up.

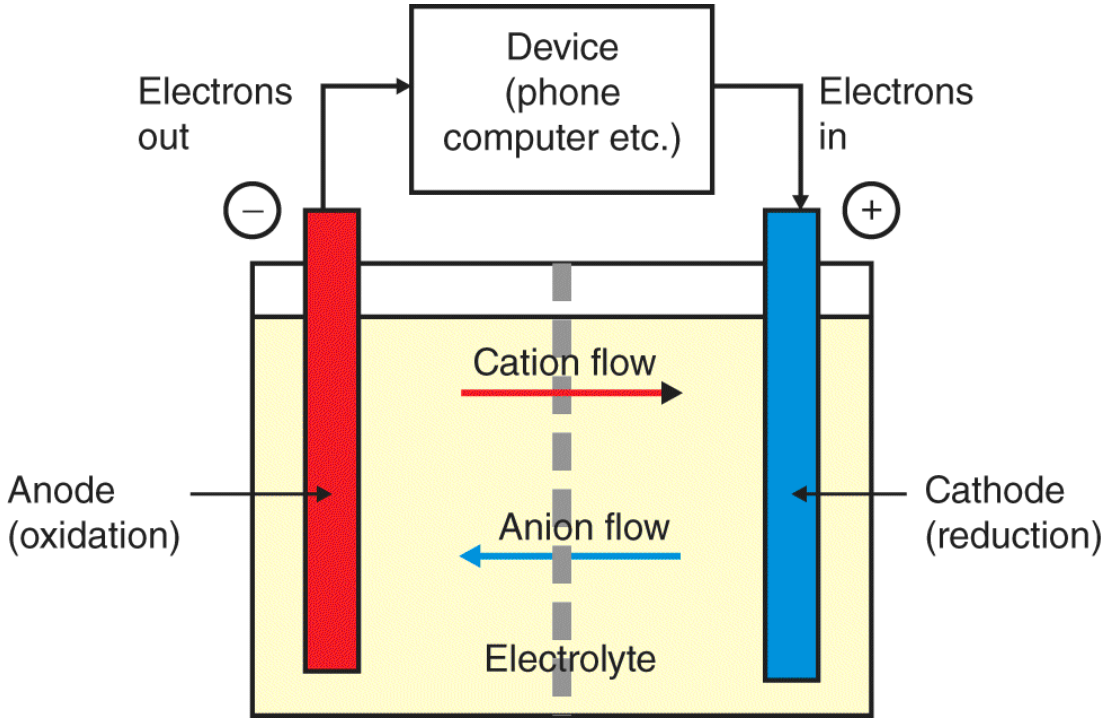


Figure 7.1 The components of a galvanic cell.

The principle behind all galvanic cells can be explained with reference to a simple example, the *Daniell cell*, invented in 1836. It consists of a zinc rod in a solution of zinc sulfate, and a copper rod in a solution of copper sulphate. A porous solid layer allows ions to pass between the sulphate electrolytes. The zinc rod forms the anode, while the copper rod forms the cathode (Figure 7.2). A current of electrons flows from the zinc anode to the copper cathode when the external connection is completed with a metallic conductor.

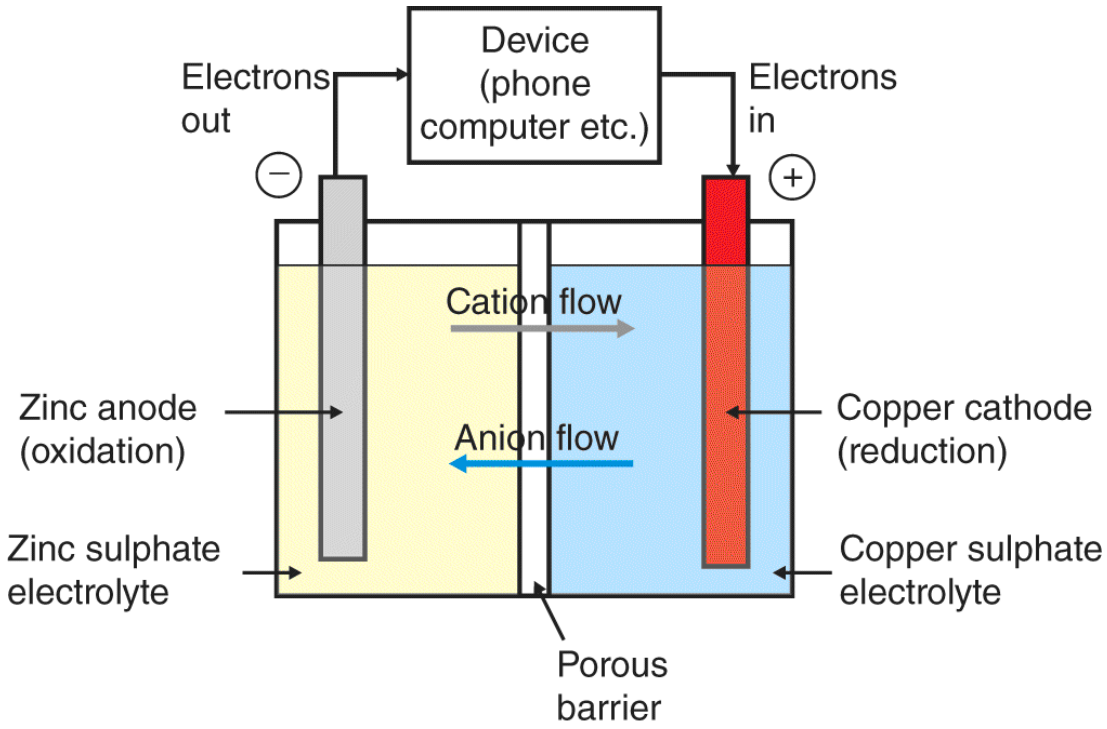


Figure 7.2 Daniell cell, schematic.

The reactions taking place at each electrode, defined by *half-reactions* are:

Anode half-reaction (oxidation):

$$Zn(s) \rightarrow Zn^{2+}(aq) + 2e^{-}$$

Cathode half-reaction (reduction):

$$Cu^{2+}(aq) + 2e^{-} \rightarrow Cu(s)$$

Zinc is oxidised to Zn^{2+} ions on one side of the porous barrier and Cu^{2+} ions are reduced to metallic copper on the other. Electrons pass via the external circuit. Ions pass through the electrolytes to maintain charge balance in the cell. The reduced and oxidised pair of species found in a half-reaction is called a *redox couple*, written oxidised species/reduced species, i.e. Cu^{2+}/Cu. The chemical reaction being used to generate electricity, called the *cell reaction*, is

obtained by adding the anode and cathode half-reactions, ensuring, by appropriate multiplication, that the number of electrons cancels out:

Cell reaction (*redox reaction*):

$$Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s)$$

During use the zinc anode is being dissolved. If left for sufficient time the rod degrades and is noticeably corroded. At the same time the copper rod gains weight and a layer of new copper metal forms on the surface. Electroplating is occurring.

7.1.2 Standard Electrode Potentials

A combination of *any* two dissimilar metallic conductors can be used to construct a galvanic cell. The *cell potential* E_{cell} defines the measure of the energy available in a cell. A high cell potential signifies a vigorous spontaneous redox reaction. Because of the multiplicity of possible cells, it is more convenient to consider the cell potential as being made up from separate voltage contributions from anode and cathode half-reactions. We can then write:

$$E_{cell} = E_c \text{ (cathode; reduction half-reaction)} + E_a \text{ (anode; oxidation half-reaction)}$$

In general, only the reduction half-reaction potentials are listed in tables (Table 7.1). The potential of an oxidation half-reaction is the negative of the value for the reduction half reaction. It is convenient to standardise the concentrations of the components of the cells. If the cell components are in their *standard states*, standard electrode potentials, E^{O}, are recorded.

$$E^0_{cell} = E^0_c + E^0_a$$

Table 7.1 Standard reduction potentials at 25 °C.

Half reaction	E^o**(V)**
$F_2 + 2e^- \rightarrow 2F^-$	+2.87
$Au^+ + e^- \rightarrow Au$	+1.69
$Ce^{4+} + e^- \rightarrow Ce^{3+}$	+1.61
$Cl_2 + 2e^- \rightarrow 2Cl^-$	+1.36
$O_2 + 4H^+ + 4e^- \rightarrow 2H_2O$	+1.23 +0.81 at pH =7
$Br_2 + 2e^- \rightarrow 2Br^-$	+1.09
$Hg^{2+} + 2e^- \rightarrow Hg$	+0.85
$Ag^+ + e^- \rightarrow Ag$	+0.80
$Fe^{3+} + e^- \rightarrow Fe^{2+}$	+0.77
$I_2 + 2e^- \rightarrow 2I^-$	+0.54
$O_2 + 2H_2O + 4e^- \rightarrow 4OH^-$	+0.40 +0.81 at pH = 7
$Cu^{2+} + 2e^- \rightarrow Cu$	+0.34
$AgCl + e^- \rightarrow Ag + Cl^-$	+0.22
$2H^+ + 2e^- \rightarrow H_2$	0 (by definition)
$Fe^{3+} + 3e^- \rightarrow Fe$	−0.04
$O_2 + H_2O + 2e^- \rightarrow OH^- + HO_2^-$	−0.08
$Pb^{2+} + 2e^- \rightarrow Pb$	−0.13
$Sn^{2+} + 2e^- \rightarrow Sn$	−0.14
$Ni^{2+} + 2e^- \rightarrow Ni$	−0.25
$Fe^{2+} + 2e^- \rightarrow Fe$	−0.44
$Fe(OH)_3 + e^- \rightarrow Fe(OH)_2 + OH^-$	−0.56
$Zn^{2+} + 2e^- \rightarrow Zn$	−0.76
$2H_2O + 2e^- \rightarrow 2OH^- + H_2$	−0.83 −0.42 at pH = 7
$Ti^{2+} + 2e^- \rightarrow Ti$	−1.63

Half reaction	E^0(V)
$Al^{3+} + 3e^- \rightarrow Al$	−1.66
$Be^{2+} + 2e^- \rightarrow Be$	−1.85
$Sc^{3+} + 3e^- \rightarrow Sc$	−2.10
$Mg^{2+} + 2e^- \rightarrow Mg$	−2.36
$La^{3+} + 3e^- \rightarrow La$	−2.52
$Na^+ + e^- \rightarrow Na$	−2.71
$K^+ + e^- \rightarrow K$	−2.93
$Li^+ + e^- \rightarrow Li$	⁻3.05

Standard conditions: temperature 25 °C, concentration of each ion is 1 mol l^{-1} and all gases are at 1 atm. pressure.

For cells involving ionic solutions the standard state is a solution of *1 molar concentration*, and for cells involving gases these are at *1 atm. pressure*[1] both at the standard temperature of 298.15 K (25 °C).

It is not possible to measure the voltage generated by half a cell, and it has been agreed that the voltage should be measured with respect to a reference electrode using hydrogen gas, called the *standard hydrogen electrode*. The reference electrode is a mixture of H^+ ions and H_2 gas in the standard state of $H^+(aq)$ at 1 mol l^{-1} and $H_2(g)$ at 1 atm. The standard electrode potential of this half-reaction is *defined* as zero.

$$2H^+(aq) + 2e^- \rightarrow H_2(g) \quad E^0 = 0\,V$$

When the hydrogen electrode is incorporated in a cell, it will form either the anode or the cathode, depending upon the other metal involved. For example, in a cell made with zinc, the zinc electrode is found to be the anode and the hydrogen electrode is the cathode.

Anode half-reaction:

$$Zn(s) \rightarrow Zn^{2+}(aq) + 2e^- \quad E^0 = +0.76\,V$$

(Because zinc is the anode, the value of E^0 is the negative of that given in Table 7.1.)

Cathode half-reaction:

$$2H^+(aq) + 2e^- \rightarrow H_2(g) \quad E^0 = 0\,V$$

Cell reaction:

$$Zn(s) + 2H^+(aq) \rightarrow Zn^{2+}(aq) + H_2(g)$$

$$E^0_{cell} = +0\,V + (+0.76\,V) = 0.76\,V$$

The Zn is oxidised to Zn^{2+} and the $H^+(aq)$ is reduced to $H_2(g)$.

In a cell using copper, the hydrogen electrode is found to be the anode and the copper electrode the cathode.

Anode half-reaction:

$$H_2(g) \rightarrow 2H^+(aq) + 2e^- \quad E^0 = 0\,V$$

Cathode half-reaction:

$$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s) \quad E^0 = +0.34\,V$$

Cell reaction:

$$H_2(g) + Cu^{2+}(aq) \rightarrow 2H^+(aq) + Cu(s)$$

$$E^0_{cell} = +0.34\,V + 0\,V = +0.34\,V$$

When many cells are compared, it is found that the anode is always the material that has the lowest tendency to be reduced. From the two examples earlier, it is seen that the order of the electrodes with respect to this tendency is $Zn^{2+}/Zn < H^+/H_2(g) < Cu^{2+}/Cu$. Thus, in a Daniell cell, Zn forms the anode and Cu the cathode.

Anode half-reaction:

$$Zn(s) \rightarrow Zn^{2+}(aq) + 2e^- \quad E^0 = +0.76\,V$$

Cathode half-reaction:

$$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s) \quad E^0 = +0.34\,V$$

Cell reaction:

$$Zn(s) + Cu^{2+}(aq) \rightarrow Cu(s) + Zn^{2+}(aq)$$

$$E^0_{cell} = +0.34\,V + (+0.76\,V) = 1.10\,V$$

The comparison of each element with a hydrogen electrode in a standard galvanic cell allows the reduction tendency to be ranked. A table of these values is arranged so that the elements with the greatest tendency to be reduced (the most strongly oxidising) are at the top is referred to as the *electrochemical series*. The oxidised species in a redox couple has the ability to oxidise the reduced species in any redox couple below it in the table. Moreover, such a reaction will be spontaneous. For example, fluorine gas, F_2, will have the highest tendency to be reduced, or gain electrons, and lithium metal, Li, has the highest tendency to be oxidised, or lose electrons. Mixing these elements will lead to a spontaneous, and very vigorous, reaction, leading to the production of Li^+ and F^- ions.

When forming a galvanic cell, the couple *higher* in the table forms the *cathode* and the couple *lower* in the table forms the *anode*. This is written using a standard notation

Lower couple (anode) || higher couple (cathode)

The cells described earlier are written:

$$Zn(s) \mid Zn^{2+}(aq) \parallel H^+(aq) \mid H_2(g) \mid Pt \quad E^0 = 0.76\,V$$

$$Pt \mid H_2(g) \mid H^+(aq) \parallel Cu^{2+}(aq) \mid Cu(s) \quad E^0 = 0.34\,V$$

$$Zn(s) \mid Zn^{2+}(aq) \parallel Cu^{2+}(aq) \mid Cu(s) \quad E^0 = 1.10\,V$$

7.1.3 Cell Potential, Gibbs Energy, and Concentration Dependence

The direction of spontaneous change taking place in a galvanic cell is that of decreasing Gibbs energy. The cell potential is related to the Gibbs energy change of the cell reaction ΔG_r by:

$$\Delta G_r = -nE_{cell}F$$

E_{cell} is the cell potential, defined to be positive; F is the Faraday constant and n is the number of moles of electrons that migrate from anode to cathode in the cell reaction. Thus, a galvanic cell is also a Gibbs energy meter.

When the electrodes are in their standard states, the free energy change is called the *standard reaction Gibbs energy*, ΔG_r^0, and the cell voltage is just the standard cell potential, E^0. In this case:

$$\Delta G_r^0 = -nE^0F$$

For example, in the Daniell cell:

$$Zn(s) + Cu^{2+}(aq) \rightarrow Cu(s) + Zn^{2+}(aq)$$

Two electrons are transferred in the cell reaction, hence $n = 2$. (Note that the number of electrons taking part in the reaction is usually clearer from the half-reactions rather than the cell reaction.) Hence:

$$\Delta G_r = -2E_{cell}F$$

$$\Delta G_r = -1.93 \times 10^5 \, E_{cell}$$

When E_{cell} is measured in volts the value of ΔG_r is in joules. When the concentrations of both Cu^{2+} and Zn^{2+} ions are in the standard state:

$$\Delta G_r^0 = -nE^0F$$

The potential generated by a cell is dependent upon the concentration of the components present. The relationship is given by the equation:

$$E_{cell} = E^0 - (RT/nF) \ln Q$$

the *Nernst equation*, where E_{cell} is the cell potential, R is the gas constant, T is the temperature (K), F is the Faraday constant, n is the number of moles of electrons that migrate from anode to cathode in the cell reaction, and Q is the reaction quotient, defined in the following text. Inserting values for the constants, the Nernst equation can be written:

$$E_{cell} = E^0 - (0.02569/n) \ln Q$$

or

$$E_{cell} = E^0 - (0.05916/n) \log Q$$

The reaction quotient, Q, of a reaction:

$$aA + bB \leftrightarrow xX + yY$$

is given by:

$$Q = Q_c = [X]^x[Y]^y/[A]^a[B]^b$$

where [A] denotes the concentration of compound A at any time. For reactions involving gases, the concentration term can be replaced by the partial pressure of the gaseous reactants, to give:

$$Q = Q_p = p_X^x p_Y^y / p_A^a p_B^b$$

Pure liquids or *solids*, or *water in solutions*, appearing in the cell reaction equation are *not* entered into the equations for *Q*. (Note, for most accurate work, the quantity of importance is the *activity* rather than concentration. Activity and concentration are equal in dilute solutions).

7.2 Chemical Analysis Using Galvanic Cells

7.2.1 pH Meters

The measurement of the potential of a galvanic cell can be used to determine the concentration of the ions in a solution via the Nernst equation. The most widespread use of this analytical technique is the measurement of pH (the acidity or the concentration of hydrogen ions) in a liquid.[2] For example, the combination of a standard electrode and a (non-standard) hydrogen electrode can be used to measure the concentration of hydrogen ions present. The standard electrode chosen is often a *calomel electrode*, which is a particularly stable electrode that uses the redox couple $Hg_2Cl_2/Hg, Cl^-$.

The cell is:

$$Pt|H_2(g)|H^+(aq)\| Cl^-(aq)|Hg_2Cl_2|Hg(l)$$

Anode half-reaction:

$$H_2(g) \rightarrow 2H^+(aq) + 2e^- \quad E^0 = 0$$

Cathode half-reaction:

$$Hg_2Cl_2(s) + 2e^- \rightarrow 2Hg(l) + 2Cl^-(aq)$$
$$E^0 = +0.27\,V$$

Cell reaction:

$$Hg_2Cl_2(s) + H_2(g) \rightarrow 2H^+(aq) + 2Cl^-(aq) + 2Hg(l)$$

The reaction quotient for the cell reaction is:

$$Q = [H^+]^2[Cl^-]^2/p_{H_2}$$

The Nernst equation is then:

$$E_{cell} = E^0(0.05916/2) \times \log\{[H^+]^2[Cl^-]^2/p_{H_2}\}$$

To simplify matters, take the hydrogen pressure and $[Cl^-]$ as standard, in which case:

$$E_{cell} = E^0 - 0.05916 \log [H^+]$$

i.e.

$$E_{cell} = E^0 + 0.059\,16\,\text{pH}$$

In cases where the hydrogen pressure and $[Cl^-]$ are not standard, they are incorporated into a *cell constant* E':

$$E_{cell} = E' + 0.059\,16\,\text{pH}$$

After calibration the pH can be read directly on a voltage scale.

Commercial pH meters that are used for measuring the pH of fluids as different as fruit juice or blood are constructed with a glass electrode that has a sensing element made of a thin membrane of a special glass sensitive only to H^+ ions (Figure 7.3a). The potential of the electrode is found to be proportional to the pH of the surrounding solution, and the response of this electrode is similar to that of a hydrogen electrode.

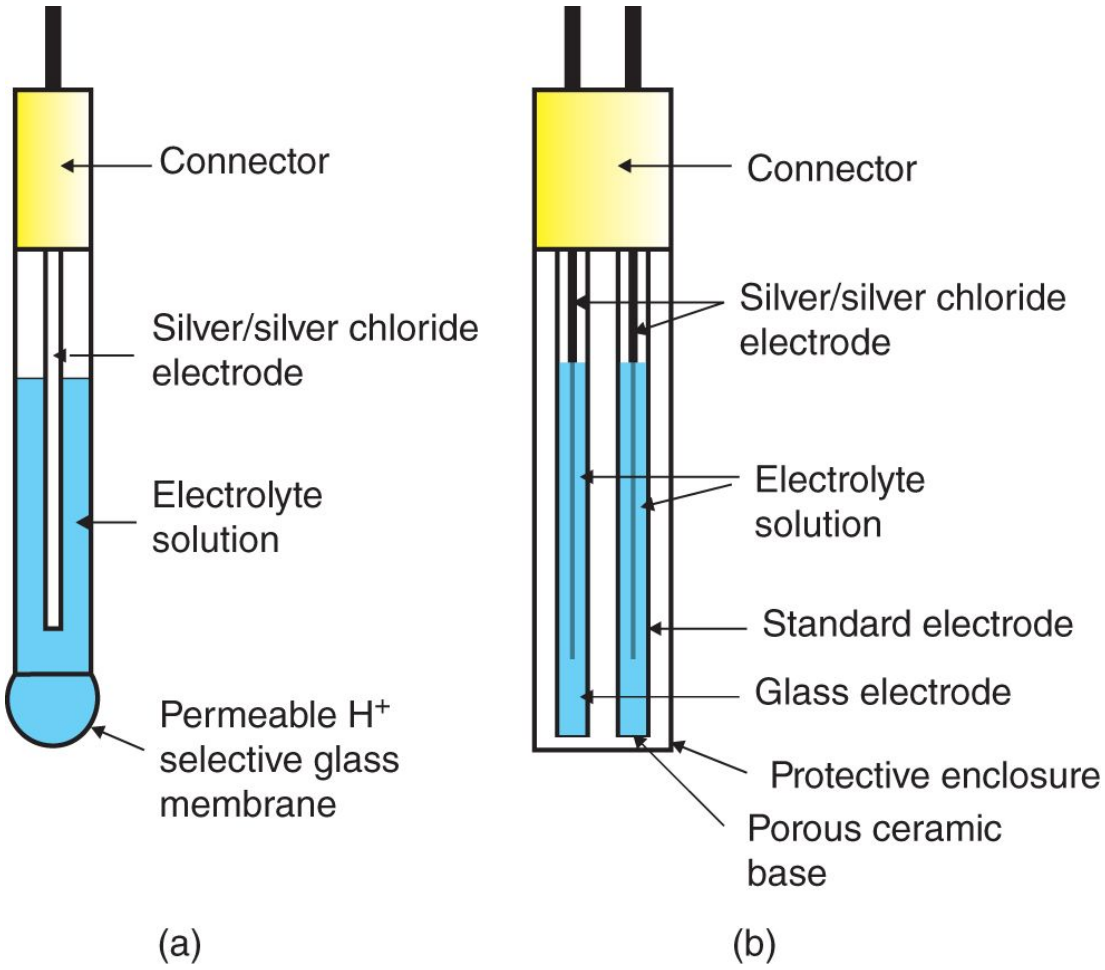


Figure 7.3 (a) A glass electrode for the measurement of pH. (b) Practical arrangement of a glass electrode and standard electrode in a single unit.

Apart from the calomel electrode, one of the more common standard electrodes used in pH meters is a silver/silver chloride electrode. This consists of a silver wire coated with silver chloride immersed in a 4 M solution of potassium chloride (KCl), saturated with silver chloride (AgCl). The half-reaction is:

$$AgCl(s) + e^- \rightarrow Ag(s) + Cl^-$$

$$E^0 = 0.2046\,V \quad \text{at } 25^\circ C.$$

In practice the hydrogen ion selective electrode and the standard electrode are packaged together in a small plastic cylinder, which is easily transported (Figure 7.3b). Measurement is made by dipping the electrode into the solution to be checked, and the voltage is transformed into pH electronically.

7.2.2 Ion Selective Electrodes

The same measurement principle can be used to determine the concentration of other ions in solution. An electrode that is (ideally) sensitive to one ion only, called an *ion selective electrode*, is paired with a standard electrode (Figure 7.4). The potential developed by such a combination is of the general form:

$$E = E' + (2.303\,RT/nF)\,\log\,[C]$$

where E' is a cell constant characteristic of the ion selective electrode and the reference electrode, and $[C]$ is the concentration of the ion. This is a linear dependence, in which the slope is $(2.303\,RT/nF)$. Experimentally, the value of the slope is found to lie between 50 and 60 mV for monovalent ions and 25–30 mV for divalent ions.

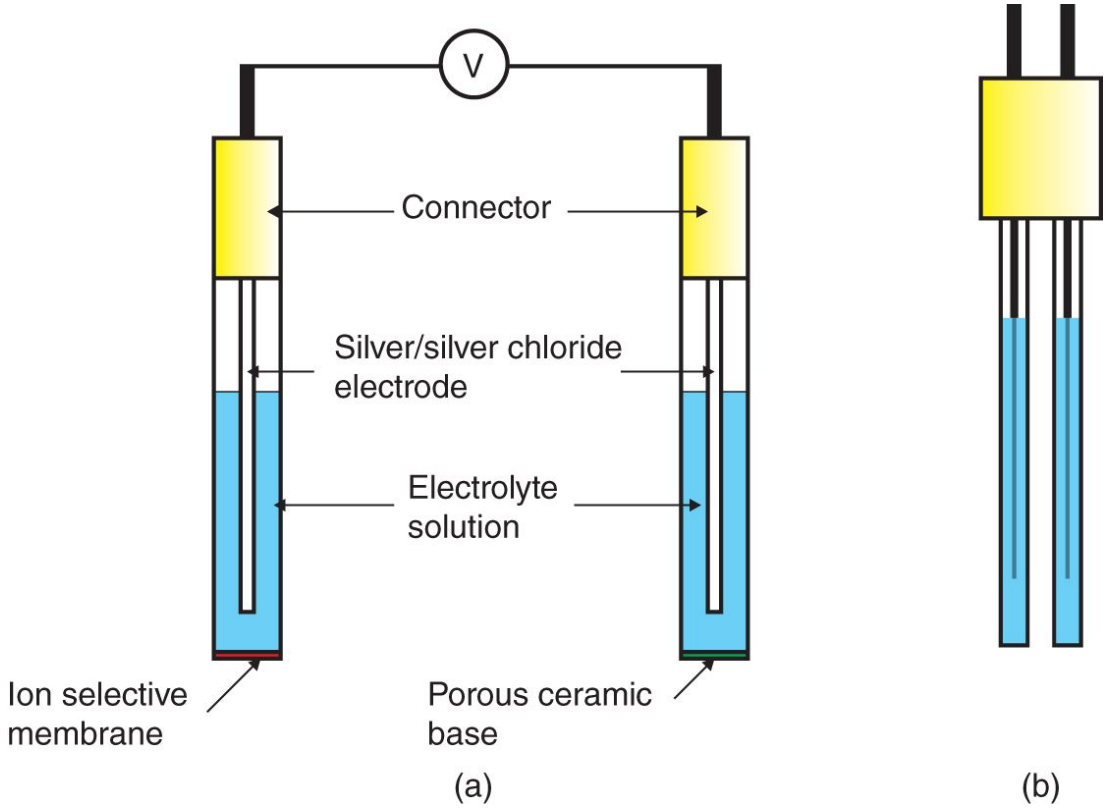


Figure 7.4 Arrangement of an ion selective electrode, schematic. (b) Practical arrangement of electrodes in a single unit.

The critical component of an ion selective electrode is a membrane that acts to pass the selected ions into the interior of the electrode assembly. These are of two principal types. *Crystal* membranes consist of a polycrystalline or single crystal plate. For example, fluoride (F^-) ion sensors are made from single crystal lanthanum trifluoride (LaF_3), doped with europium difluoride (EuF_2). The Eu^{2+} ions substitute for La^{3+} in the LaF_3 matrix, and each substituted ion

is accompanied by a vacancy on the F^- sub-structure, to maintain charge neutrality. The large number of vacancies thus generated increases the diffusion coefficient of F^- in LaF_3 enormously. The membrane has a similar permeability to F^- as the surrounding liquid, and is found to be highly selective for the passage of F^- ions.

The other type of membrane in use consists of a polyvinyl chloride (PVC) disc, impregnated with a large organic molecule that can react with the ion. The binding must be weak enough for the ion to be passed from one molecule to another across the membrane under the driving force of a concentration gradient. For example, K^+ ion selective membranes are made using the antibiotic valinomycin. This has a structure that accommodates the K^+ ions and can pass them on from one molecule to another. The operation of this material mimics the way in which living cells transfer ions across the cell membrane. Unfortunately, such molecules are not usually completely specific for a single ion, and usually also channel chemically similar species. The potassium membrane, for example, can also pass lesser amounts of sodium ions.

7.2.3 Oxygen Sensors

The Nernst equation indicates that a cell potential will develop even when the electrode materials are the same, provided that here is a difference in concentration on each side of the electrolyte. This has been widely utilised as an *oxygen sensor* using calcia-stabilised zirconia as the active electrolyte. Calcia-stabilised zirconia is a non-stoichiometric oxide prepared by reaction of ZrO_2 and CaO containing a high concentration of oxygen vacancies, which means that oxygen ions can diffuse very rapidly through he ceramic (Section 5.13). In its simplest form, the sensor is just a slice of stabilised zirconia separating oxygen gas at two different pressures. The high oxygen ion diffusion coefficient will allow ions to move from the high-pressure side to the low-pressure side to even out the pressure differential. Connecting both sides of the stabilised zirconia via porous platinum electrodes sets up an electrochemical cell in which the voltage is proportional to the difference between the oxygen pressures. The cell can be represented as:

$$\mathrm{Pt},\ \ \mathrm{O_2}(p'_{\mathrm{O_2}})\|\text{stab.zirconia}\|\mathrm{O_2}(p''_{\mathrm{O_2}}),\ \ \mathrm{Pt}$$

The reactions taking place are:

$$\textit{Anode reaction:}$$

$$2\mathrm{O}^{2-} \rightarrow \mathrm{O_2(g)}(p'_{\mathrm{O_2}}) + 4\mathrm{e}^-$$

Cathode reaction:

$$4e^- + O_2(g)(p''_{O_2}) \rightarrow 2O^{2-}$$

Overall cell reaction:

$$O_2(g)(p''_{O_2}) \rightarrow O_2(g)(p'_{O_2})$$

The cell voltage is related to the oxygen pressures by the Nernst equation:

$$E_{cell} = E^0 - (RT/nF)\ \ln\ Q$$

In this case, the number of electrons transferred, n, is 4 and the appropriate reaction quotient is:

$$Q = (p'_{O_2})/(p''_{O_2})$$

Noting that E^o for this cell reaction is zero:

$$E = (RT/4F)\ln[(p'_{O_2})/(p''_{O_2})]$$

$$E = -(RT/4F)\ \ln\ [\text{anode pressure (low)/cathode pressure (high)}]$$

Solving this equation for the oxygen partial pressure gives:

$$(p'_{O_2}) = (p''_{O_2})\exp[-4E/RT]$$

The high pressure p''_{O_2} is taken as a reference pressure so that the unknown pressure p'_{O_2} is readily determined. These equations are often seen in the form:

$$E = +(RT/4F)\ln[(p''_{O_2})/(p'_{O_2})]$$

$$E = -(RT/4F)\ \ln\ [\text{cathode pressure (high)/anode pressure (low)}]$$

The same principles apply for oxygen in solutions as diverse as liquid metals or blood. Because the oxygen is dissolved, the voltage measured depends upon

the activity of the oxygen in the solution. For low concentrations, the activity can be approximated to the concentration, hence:

$$E = -(RT/4F)\ln\{[O_2(\text{solution})]/[p_{O_2}(\text{reference})]\}$$

where [O_2 (solution)] is the concentration of the oxygen molecules, which is assumed to be lower than the reference pressure. If p_{O_2} is taken as 1 atm. the equation becomes:

$$E = -(RT/4F)\ \ln[O_2\ (\text{solution})]$$

These equations consider the oxygen present as molecules in solution. If the oxygen exists as atoms in solution, only two electrons are needed in the cell equation and the potential is given by:

$$E = -(RT/2F)\ \ln[O\ (\text{solution})]$$

The design of the sensor depends upon its ultimate use. For high temperature applications a stabilised zirconia tube coated inside and outside with porous platinum forms a suitable design. The tube can be used directly as an oxygen meter if p''_{O_2} is a standard pressure, such as 1 atm. of oxygen or else the pressure of oxygen in air, which is approximately 0.21 atm. Such a system is utilised to monitor the oxygen content of a car engine exhaust when the coated zirconia tube is arranged to project into the exhaust stream of the engine. Using air as the standard oxygen pressure, the output voltage of the sensor is directly related to the stoichiometry of the air–fuel mixture. The cell voltage may be used to alter the engine input fuel/air mix automatically to optimise engine efficiency.

7.3 Batteries

Batteries fall into one of three main types. A *primary cell* is a battery that cannot be recharged. A *secondary cell* is a battery that can be recharged and reused. A *fuel cell* has a continuous input of chemicals (fuel) to produce a continuous output of current. Batteries are of considerable importance, and the failure of electric vehicles to become established in the early twentieth century was inherently due to inadequate batteries. Even at present the performance of electric vehicles is constrained by battery performance. Batteries are also vital for many sorts of energy storage, including energy from intermittent sources of power, such as wind turbines and solar power generators.

Batteries are simple in concept. It is merely necessary to combine two couples such as those listed in Table 7.1 and link them via a suitable electrolyte, which must be able to transport ions but not electrons. For optimum operation, the electrolyte compartment is frequently divided into two volumes by a porous separator. In a primary cell, during operation (discharge), the couple lower in Table 7.1 forms the negative electrode of the battery, the *anode*. Electrons are transferred from the anode to the positive electrode of the battery, the *cathode*, and in so doing can be used to supply power to the load, which might, for example, be a light bulb (Figure 7.5a). In a secondary cell, the power supply operation is identical. However, the cell reaction can be driven backwards to regenerate the initial configuration. For this, a charging unit replaces the load, and this supplies the power needed to reverse the cell reaction (Figure 7.5b). The 'to and fro' motion of the electrons and cations during charge and discharge is called the *rocking chair* mechanism.

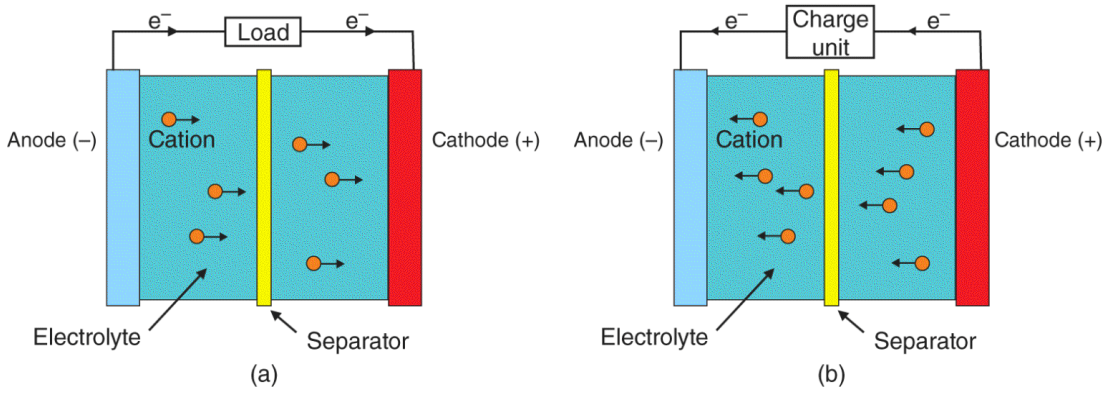


Figure 7.5 The 'rocking chair' operation of a secondary battery, schematic: (a) in use; (b) recharging.

The amount of electricity that can be drawn from a battery depends upon the chemistry of the redox reaction and the overall power depends upon the redox couple involved. For a more powerful battery, combine couples from near the

top and bottom of Table 7.1. The battery should ideally have a high specific energy (energy/weight) and a high energy density (energy/volume).

The nineteenth and early twentieth century saw vast numbers of electrochemical couples investigated for potential batteries, but the majority of these never achieved commercial production. The problems encountered centred upon the difficulty of finding a suitable electrolyte and degradation at the interfaces in the cell. Indeed, the first successful cells, invented in the mid-nineteenth century, are still widely available; the most important recently developed cell – the lithium-ion (Li-ion) battery – is widely used in phones, computers, power tools, and electric vehicles. The commercially available Li-ion batteries still show shortcomings and there is much current research exploring improvements to these cells. These include batteries in which Li is coupled with S, as well as batteries in which the Li is replaced by Na, K, or Mg, and which also are frequently coupled to S.

7.3.1 Primary Batteries

7.3.1.1 'Dry' and Alkaline Primary Batteries

The widely available primary *carbon-zinc battery* (also called a dry cell or Leclanché cell) was invented in 1866 and gives a voltage of about 1.5 V. It is especially used for intermittent applications, such as flashlights. The redox couple used is Zn/MnO_2. The current collector is a graphite rod buried in the positive cathode, a mixture of MnO_2 and carbon. This is kept moist by the electrolyte, aqueous ammonium chloride. The negative anode is the container itself, made from zinc (Figure 7.6).

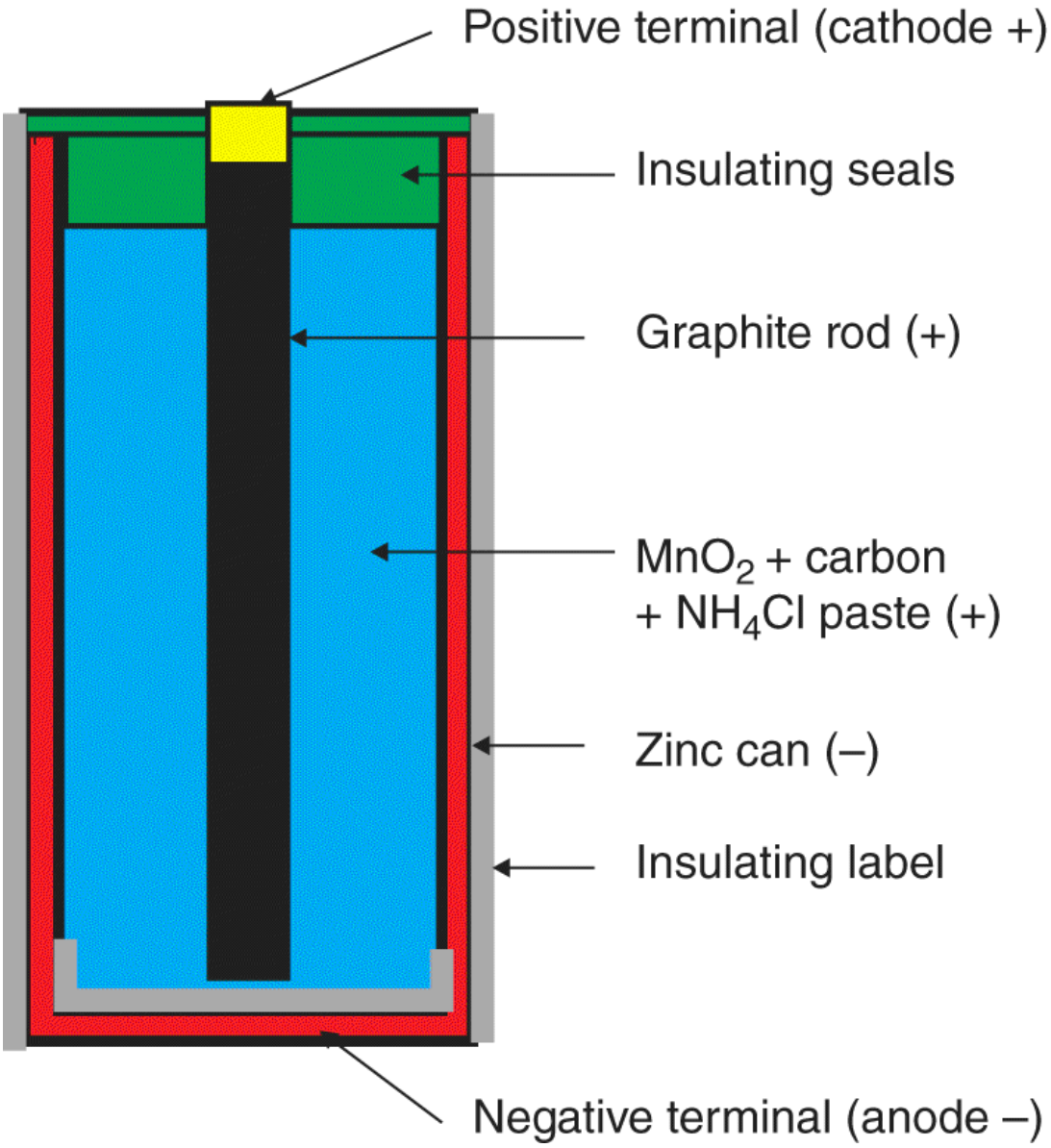


Figure 7.6 Section through a zinc-carbon battery, schematic.

As with all commercial batteries, the real cell reaction is complex, and the reactions are approximated by:

$$\textit{Anode reaction:}\ \ Zn(s) \rightarrow Zn^{2+}(aq) + 2e^-$$

$$\textit{Cathode reaction:}\ \ MnO_2(s) + H_2O(l) + e^- \rightarrow MnO(OH)(s) + OH^-(aq)$$

$$\textit{Cell reaction}:\ Zn(s) + 2MnO_2(s) + 2H_2O(l) \rightarrow 2MnO(OH)(s) + 2OH^-(aq)$$

The main problem encountered with this cell is the build-up of Zn^{2+} and OH^- at the respective electrodes, which is why the battery is best used intermittently. When the battery is not being used, the concentrations of these reaction products fall again. This is because the OH^- ions migrate to the Zn anode where they form ammonia with NH_4^+ ions in the electrolyte:

$$NH_4^+(aq) + OH^-(aq) \rightarrow H_2O(l) + NH_3(aq)$$

The concentration of the Zn^{2+} ions in the vicinity of the anode subsequently drops due to reaction with the NH_3:

$$Zn^{2+}(aq) + 4NH_3(aq) \rightarrow Zn(NH_3)_4^{2+}(aq)$$

The cell is spent when ionic conduction is no longer possible due to build-up of $Zn(NH_3)_4Cl_2$.

Alkaline cells use the same zinc–manganese dioxide couple as Leclanché cells. However, the ammonium chloride electrolyte is replaced with a solution of about 30 wt% potassium hydroxide. The cell reactions are identical to those above, but the battery construction is rather different (Figure 7.7). The negative material is zinc powder, and the anode is a brass pin. The positive component is a mixture of MnO_2 and carbon powder that surrounds the anode. A porous cylindrical barrier separates these components. The positive terminal (cathode) is the container, which is a nickel-plated steel can.

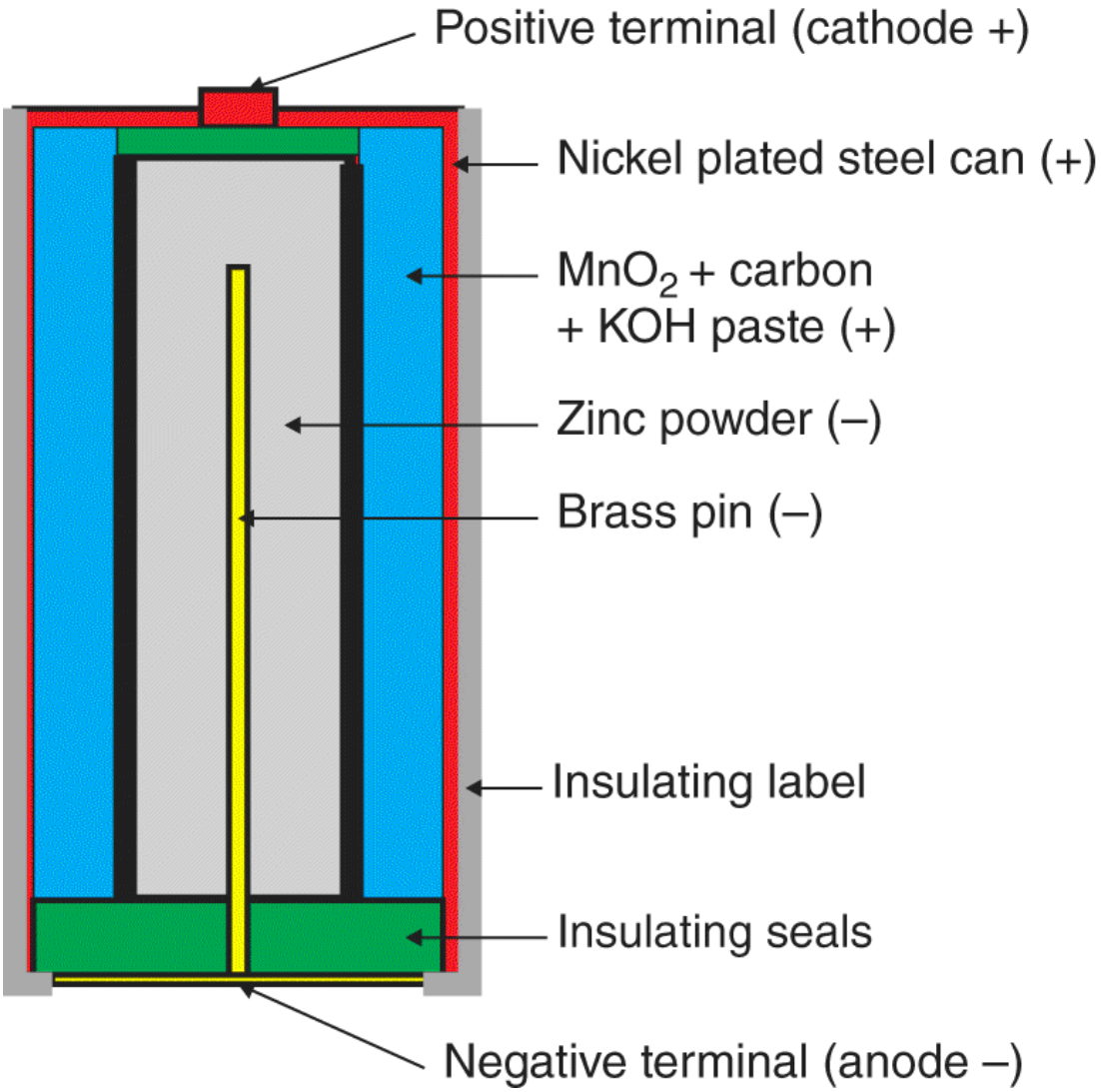


Figure 7.7 Section through an alkaline battery, schematic.

7.3.1.2 Lithium-Ion Primary Batteries

Lithium has a number of advantages over other materials for battery manufacture. It is the lightest true metal, and it also has a high electrochemical reduction potential (Table 7.1). There is one disadvantage in using lithium, in that it is very reactive, a feature that poses problems not only in manufacture but also in the selection of the other battery components. Despite this, there are a large number of lithium based primary cells available, both in traditional cylindrical form and as button and flat coin cells. The lithium forms the anode in such cells and a variety of compounds form the anode. The most usual of these is manganese dioxide, MnO_2, giving a working potential of approximately 3 V. Because lithium reacts vigorously with water, the electrolyte must be non-aqueous. It is frequently a solution of lithium salts in a

polar organic liquid, typically lithium perchlorate, $LiClO_4$, dissolved in dimethoxyethane. The conductivity of such solutions is low compared with that of aqueous solutions of hydroxides, which means that the design of a cell is constrained by needing large electrode areas separated by a thin electrolyte. The coin cell is a natural result of such considerations (Figure 7.8).

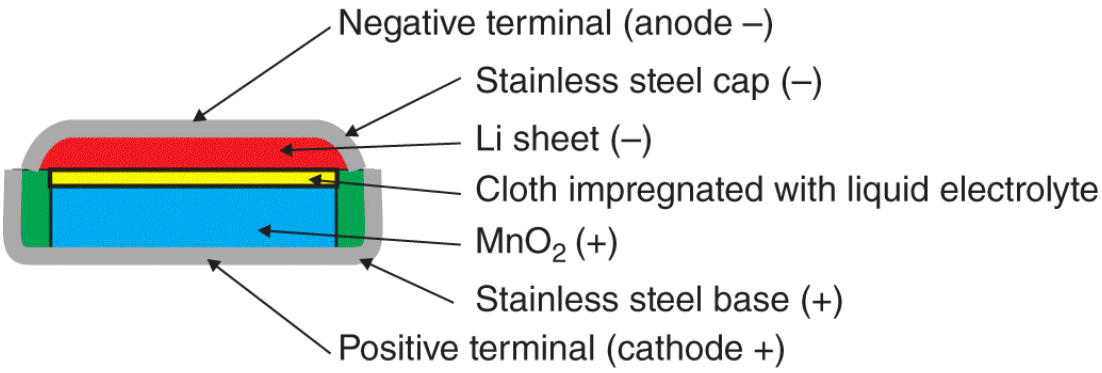


Figure 7.8 Section through a lithium coin cell primary battery, schematic.

The cell reactions are poorly understood, but can be written schematically as:

$$\textit{Anode reaction:} \quad Li(s) \rightarrow Li^+(aq) + e^-$$

$$\textit{Cathode reaction:} \quad MnO_2(s) + e^- \rightarrow MnO_2^-(s)$$

$$\textit{Cell reaction:} \quad Li(s) + MnO_2(s) \rightarrow LiMnO_2(s)$$
$$E^0 = 3.2\,V$$

7.3.1.3 Lithium–Air Batteries

Lithium–air and lithium–O_2 batteries are still not fully developed, mainly due to electrolyte problems and the fact that the cathode cannot be exposed to ordinary air, because water vapour will react with the reactant Li_2O_2 (formed during the cathode reaction, as below) to produce lithium hydroxide, and carbon dioxide will then react with this to form lithium carbonate. This means that the cathode must be enclosed in a protective membrane that allows the free passage of O_2 but excludes H_2O and CO_2. There are large numbers of aqueous and non-aqueous electrolytes that are being explored including the liquid electrolyte $LiPF_6$ in polypropylene carbonate and solid electrolytes related to NASICON (Section 5.13.4), as well as garnet structure and perovskite structure oxides. A typical cell assembly employs lithium metal as the anode and air as the cathode, often reaching the electrolyte surface via a porous carbon layer (Figure 7.9).

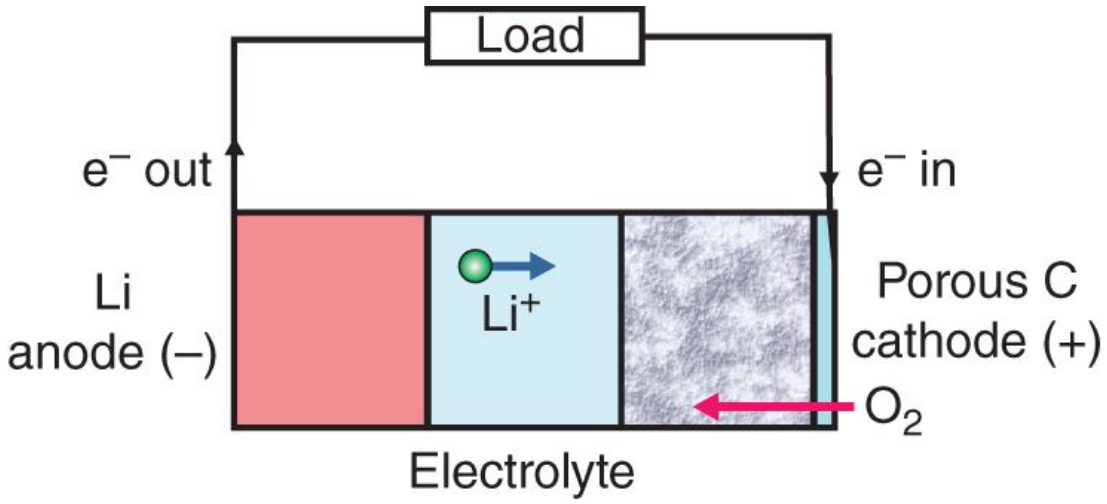


Figure 7.9 Li-air battery, schematic

Details of the cell reactions depend upon the electrolyte, and not all are known in detail. For a non-aqueous electrolyte, the reactions may be approximated as:

Anode reaction:

$$Li \rightarrow Li^+ + e^-$$

Cathode reaction: (in solution)

$$2Li^+(sol) + O_2^-(sol) \rightarrow Li_2O_2$$

$$2Li^+(sol) + 2O_2^-(sol) \rightarrow Li_2O_2 + O_2$$

Cathode reaction: (on Li_2O_2 surface)

$$Li^+ + O_2(sol) + e^- \rightarrow LiO_2 \text{ (on } Li_2O_2)$$

$$Li^+ + LiO_2 \text{ (on } Li_2O_2) + e^- \rightarrow 2 \text{ } (Li_2O_2) \text{ (on } Li_2O_2)$$

$$2 \text{ } (2LiO_2) \text{ (on } Li_2O_2) \rightarrow 2 \text{ } (Li_2O_2) \text{ (on } Li_2O_2) + O_2$$

Cell reaction: (approximate)

$$2Li + O_2 \rightarrow Li_2O_2$$

7.3.2 Fuel Cells

Batteries have a fixed amount of reactant present, stored in the battery casing. Fuel cells are primary cells with a continuous input of chemicals and a continuous output of power. The reactants are stored separately from the electrodes and electrolyte and can be replenished when necessary. (The Li-air cell is, in fact, a form of fuel cell with only one consumable chemical introduced.) There is much research at present on fuel cells as a source of clean electricity. The favoured reaction is the production of water from hydrogen gas and oxygen gas giving a cell voltage of about 1.2 V. The concept is simple (Figure 7.10), and early fuel cells, containing an alkaline solution, typically potassium hydroxide solution, as electrolyte, demonstrate the principle involved.

Anode reaction:

$$H_2(g) + 2OH^-(aq) \rightarrow 2H_2O(l) + 2e^- \quad E^0 = -0.83\,V$$

Cathode reaction:

$$O_2(g) + 2H_2O(l) + 4e^- \rightarrow 4OH^-(aq) \quad E^0 = 0.40\,V$$

Cell discharge reaction:

$$2H_2(g) + O_2(g) \rightarrow 2H_2O(l)$$

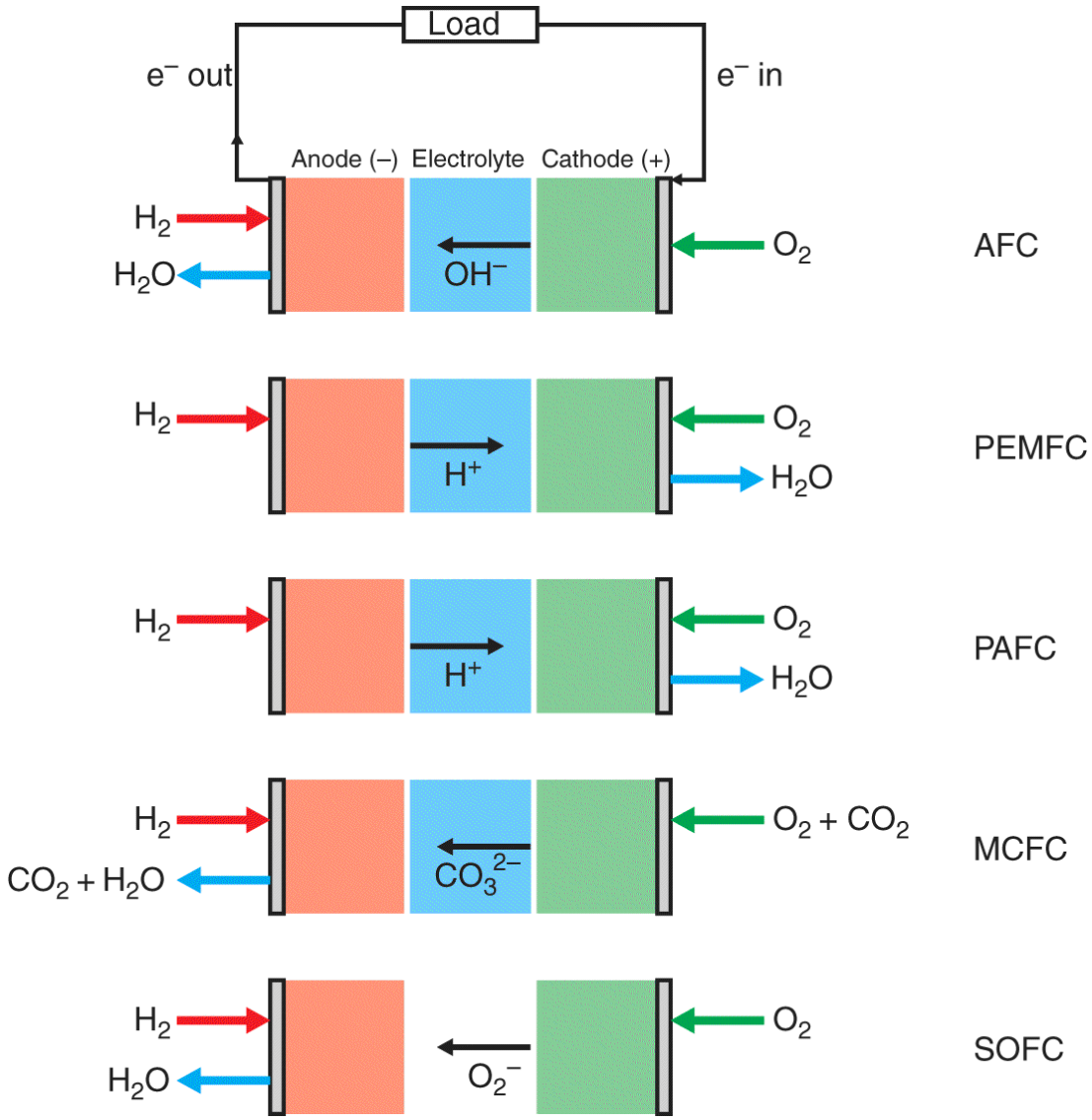


Figure 7.10 Fuel cells, schematic: AFC, alkaline fuel cell; PEMFC, proton exchange membrane fuel cell; PAFC, phosphoric acid fuel cell; MCFC, molten carbonate fuel cell; SOFC, solid oxide fuel cell.

Fuel cells are described in terms of the electrolyte that is used. The principle types are *alkaline fuel cell* (AFC), described earlier, with aqueous KOH as electrolyte transferring OH^- ions; *proton exchange membrane fuel cell* (PEMFC), using a solid polymer electrolyte that conducts H^+ ions; *phosphoric acid fuel cell* (PAFC), using phosphoric acid as electrolyte in a matrix of silicon carbide, to transfer H^+ ions; *molten carbonate fuel cell* (MCFC), with molten Li_2CO_3 as electrolyte in an $LiAlO_2$ matrix, transporting CO_3^{2-} ions; and *solid oxide fuel cell* (SOFC), with solid yttria-stabilised zirconia electrolytes that allow oxide ion, O^{2-}, transport. The majority of cells use a catalyst to speed up the reactions taking place. These are mostly platinum based, and are deposited

upon the electrodes as small highly dispersed particles, although the molten carbonate cell has an advantage in that it uses a cheaper nickel catalyst.

As the cell reaction is between hydrogen and oxygen, the continuous supply of these gases is vital. Oxygen poses no problem, as it is freely available from the air. The best way to provide hydrogen fuel has not yet been resolved. Hydrogen reservoirs containing pressurised gas, or as liquid, have been considered. One promising avenue seems to be the use of non-stoichiometric metal hydrides, which can reversibly store hydrogen. In addition to this there is much work on using methane, methanol, petrol (gasoline), and natural gas as fuel. Except for natural gas, these are converted catalytically into mixtures of H_2 and CO before entering to cell although *direct methanol conversion fuel cells* (DMCFC) are also under investigation.

Of the cells available, SOFCs are suitable for large-scale electricity generation although these suffer from the problem that diffusion of oxygen ions in the electrolyte is slow at room temperature and satisfactory cell operation is only accomplished when the electrolyte is held at higher temperatures of 700–800 °C. A variety of designs are being explored, including planar, and tubular geometries (Figure 7.11). In all cases, single cells are linked to give a fuel-cell stack by an interconnect material. The electrolyte in these cells is the oxygen ion conductor yttria-stabilised zirconia (YSZ). The defect chemistry of this material is similar to calcia-stabilised zirconia (Section 5.13.3). Zirconia is doped with Y_2O_3 (yttria) to form a phase with Y^{3+} substituted for Zr^{4+}. As the dopant ions have a lower charge than the Zr^{4+} ions the crystal compensates by introducing anion vacancies, resulting in a high oxygen ion diffusion coefficient. Oxygen in the form of air is supplied to the cathode, often called the *air electrode*. The oxygen gas is ionised and oxygen ion transport across the electrolyte ensues. Hydrogen fuel is supplied to the anode or *fuel electrode*. Here it reacts with the oxide ions to form water.

Anode reaction:

$$H_2(g) + O^{2-}(s) \rightarrow H_2O(l) + 2e^-$$

Cathode reaction:

$$\tfrac{1}{2}\, O_2(g) + 2e^- \rightarrow O^{2-}(s)$$

Cell discharge reaction:

$$H_2(g) + \tfrac{1}{2}\, O_2(g) \rightarrow H_2O(l)$$

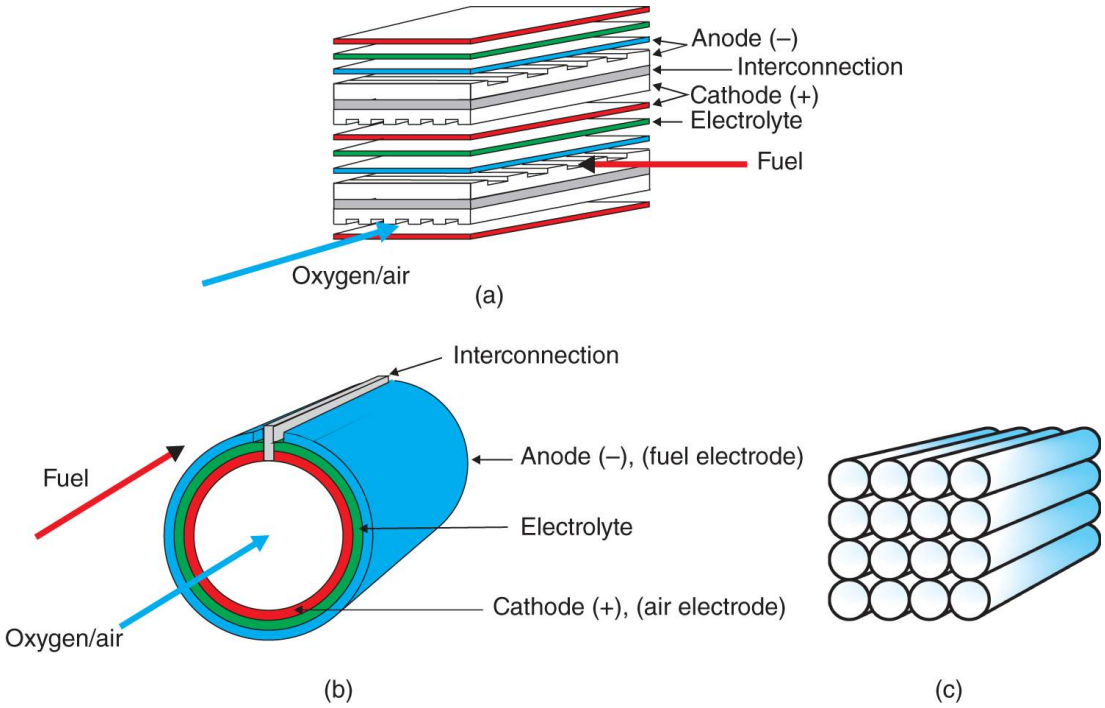


Figure 7.11 Solid oxide fuel cells: (a) planar-design; (b) tubular design; (c) an array of tubular fuel cells.

The other widely used fuel cell, especially for low power generation, is the PEMFC cell. In this type of cell, the fuel, typically hydrogen, is oxidised at the anode to give electrons, which leave via the external circuit, and H^+ cations, which are transported across the membrane to the cathode. Here they are reduced on the presence of oxygen to form water:

Anode reaction:

$$H_2(g) \rightarrow 2H^+(aq) + 2e^-$$

Cathode reaction:

$$\tfrac{1}{2}\, O_2(g) + 2H^+(aq) + 2e^- \rightarrow H_2O(l)$$

Cell discharge reaction:

$$H_2(g) + \tfrac{1}{2}\, O_2(g) \rightarrow H_2O(l)$$

The membrane in use has the trade name Nafion, which is readily available in a variety of forms. Although the chemical details may vary slightly from one product to another, this material is essentially a solid polymer of tetrafluoroethylene, CF_2CF_2. The polymer backbone consists of $-[-(CF_2-CF_2-)_x-(-CF_2-CF-)_y-]-$ chains. Each of the variable number of CF groups

along the chain is linked to a side-chain that ends in an acid $-SO_3Na$ or $-SO_3H$ termination. The polymeric CF backbone is extremely stable and can tolerate operating conditions of up to 190 °C. The terminations allow for the transport of Na^+ or H^+ across the polymer membrane.

7.3.3 Secondary Batteries

7.3.3.1 The Lead-Acid Battery

Planté discovered the basic technology of the rechargeable *lead-acid battery* in 1859. Since then there have been many refinements to the materials used, but the operating principles remain the same and this is the car battery still in use to this day. The anode is lead, the cathode lead dioxide, and the electrolyte dilute sulphuric acid (Figure 7.12). As with all batteries, the chemical reactions taking place are complex, but schematically the processes are:

Anode reaction:

$$Pb(s) + HSO_4^-(aq) \rightarrow PbSO_4(s) + H^+(aq) + 2e^-$$

Cathode reaction:

$$PbO_2(s) + 3H^+(aq) + HSO_4^-(aq) + 2e^- \rightarrow PbSO_4(s) + 2H_2O$$

Cell discharge reaction:

$$PbO_2(s) + Pb(s) + 2H^+(aq) + HSO_4^-(aq) \rightarrow 2PbSO_4(s) + 2H_2O$$

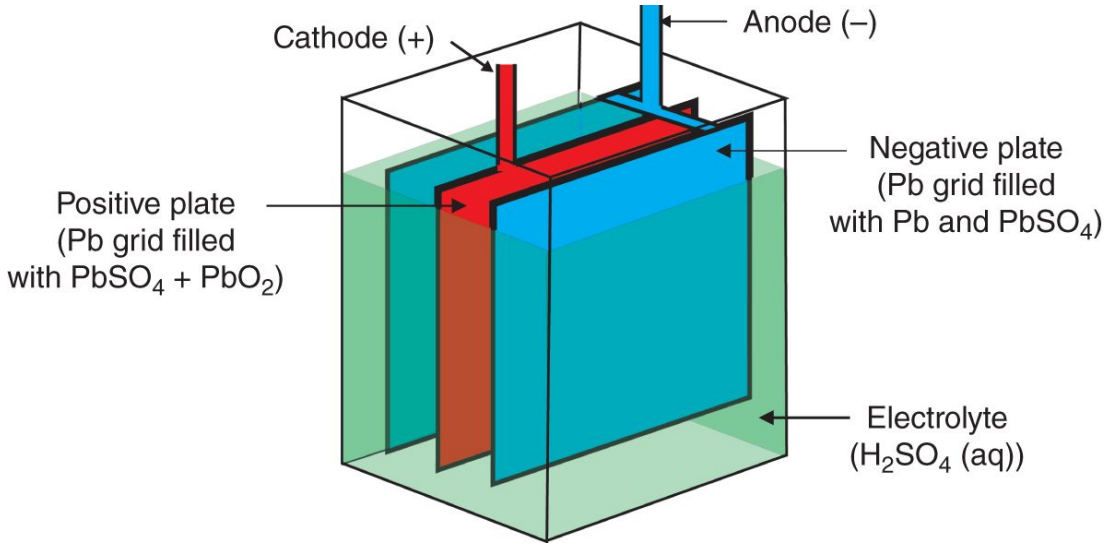


Figure 7.12 A single cell in a lead-acid battery, schematic.

The cell potential is about 2 V. A car battery consists of (usually) six cells in series (a *battery* of cells), to give 12 V. Sulphuric acid is used up during operation, so the state of charge of the battery can be estimated by measuring the concentration of the acid, usually via density.

The cell charging reaction is the reverse of the discharge reaction, and to charge a battery the cell reaction is simply driven backwards by an imposed external voltage.

7.3.3.2 Lithium-Ion Batteries

The advantages of lithium primary cells extend to secondary cells. In particular, the high power available and their lightness make them ideal for portable electronic devices. The original commercial cell design (Figure 7.13) used Li metal incorporated into graphite as the anode (–) and the non-stoichiometric oxide Li_xCoO_2 as cathode. The electrolyte for these cells is usually a solution of $LiPF_6$ dissolved in a 1 : 1 mixture of ethylene carbonate and diethylene carbonate (EC/DEC).

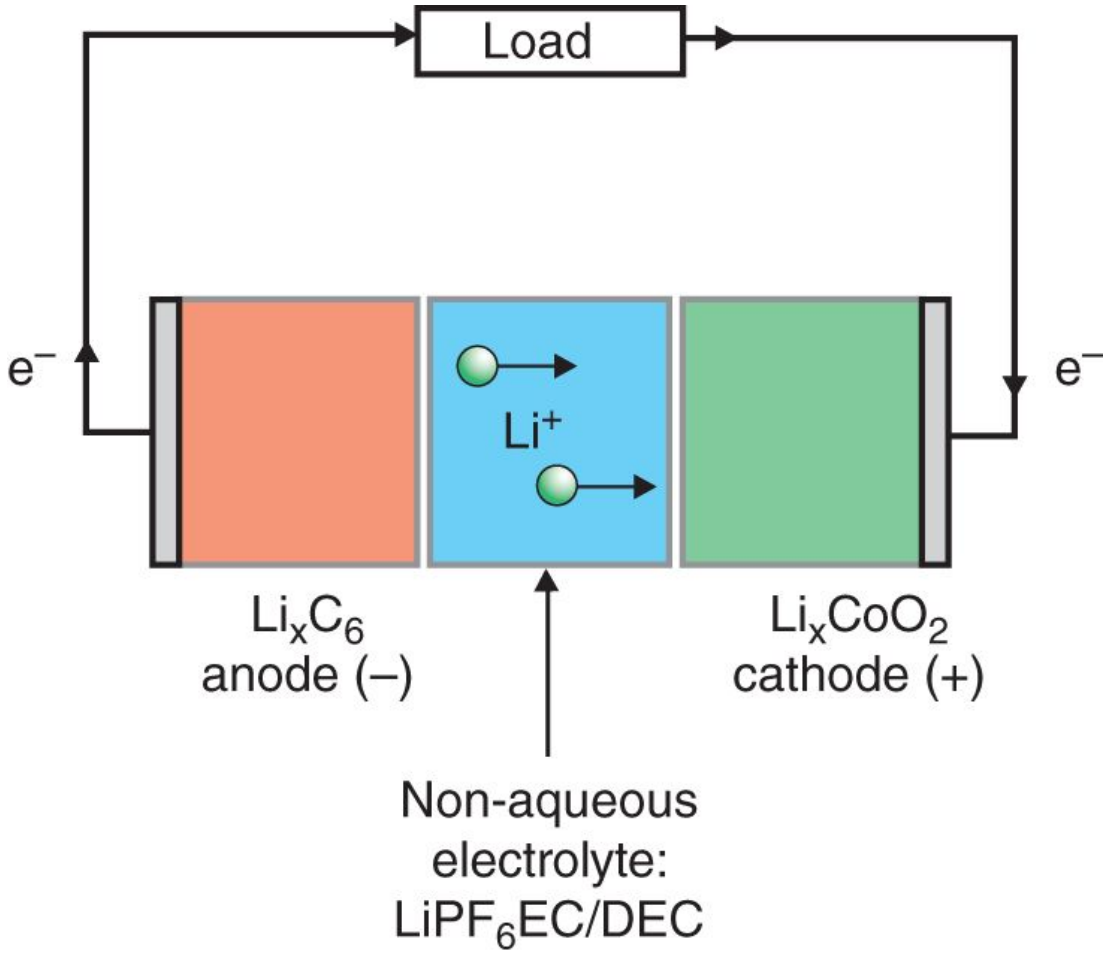


Figure 7.13 The original configuration of a Li-ion cell in discharge operation.

The nominal composition of the lithium-graphite anode is Li_xC_6, with x varying from 0 to approaching 1.0. The stacking of the hexagonal carbon layers is staggered in the pure compound, but in the lithium-containing phase they are directly over each other (Figure 7.14a,b). The stacking, therefore, alters with the degree of lithium incorporation, limiting the performance of the electrode. The cathode material, Li_xCoO_2, has the Li^+ ions inserted between layers of composition CoO_2 (Figure 7.14c,d). In theory, the composition range of the cathode material is from $LiCoO_2$ to CoO_2, i.e. x varying from 1 to 0, but in battery operation the degree of non-stoichiometry is restricted, and x generally takes values from 0.5 to about 0.9. As with graphite, intercalation of lithium changes the layer stacking. In this case, CoO_2 is built of hexagonal close-packed (...AB...) layers of oxygen atoms, while $LiCoO_2$ has cubic (...ABC...) close packing. This change degrades the cathode material during use and is the major reason why the useful composition range is limited.

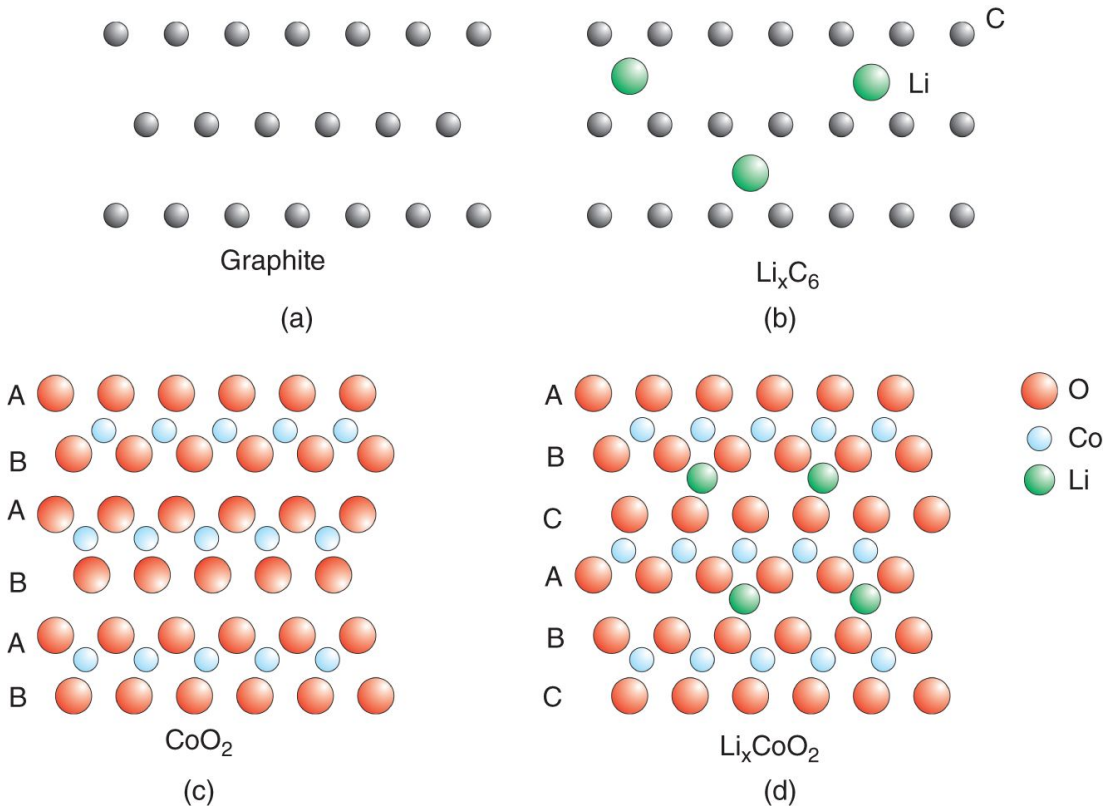


Figure 7.14 Li-ion cell electrode reactions: (a, b) anode: the C layer stacking changes from staggered to regular; (c, d) cathode: the hexagonal close packing of O^{2-} reverts to cubic close packing.

The cell reactions are similar to those utilised in the lithium primary cell. During discharge Li^+ ions are transported from the anode to the cathode via the following reactions:

Anode reaction:

$$Li_xC_6(s) \rightarrow 6C(s) + x\,Li^+(s) + x\,e^-$$

Cathode reaction:

$$Li_{0.55}CoO_2(s) + x\,Li^+(s) + x\,e^- \rightarrow Li_{0.55+x}CoO_2(s)$$

Cell discharge reaction:

$$Li_xC_6(s) + Li_{0.55}CoO_2(s) \rightarrow 6C(s) + Li_{0.55+x}CoO_2(s)$$

These reactions can also be written as:

Anode reaction:

$$Li(s) \rightarrow Li^{+}(s) + e^{-}$$

Cathode reaction:

$$Co^{4+}(s) + e^{-} \rightarrow Co^{3+}(s)$$

Cell discharge reaction:

$$Li(s) + Co^{4+}(s) \rightarrow Li^{+}(s) + Co^{3+}(s)$$

The charging reaction is the reverse of the discharge reaction, driven by an external voltage.

From the inception of this battery, there continues to be much research work on improving the materials in these cells. Cobalt, for instance, is costly and not easy to obtain. Currently a variety of materials with a similar structure and electrochemistry are used, including oxides typified by $LiNi_{0.80}Co_{0.15}Al_{0.05}O_2$, $Li_{0.33}Co_{0.33}Mn_{0.33}O_2$, and more complex phases with compositions $LiNi_{1-x-y}Co_xMn_yO_2$. In addition, there is considerable effort being put into developing an all solid-state cell using a number of Li containing compounds including Lithium analogues of NASICON (LISICON) (Section 5.13.4) and materials with the garnet structure, including $Li_7La_3Zr_2O_{12}$ and doped derivatives.

7.3.3.3 Dual-Ion Batteries

The secondary cells described have all operated with one ion – a cation – present in the electrolyte. During charge and discharge the cations travel to and fro in a *rocking-chair* fashion (Figure 7.5). In dual-ion batteries, both anions and cations are involved in battery function, and during the charge–discharge cycle these ions move in an *advance and retreat* fashion (Figure 7.15). To achieve this mode of operation, the electrolyte is divided into two compartments, one containing cations, typically Li^+, Na^+, and K^+, and the other containing anions, typically PF_6^-, $AlCl_4^-$, and ClO_4^-. These active ions are generated by the electrolytes and so the electrolyte is no longer a passive component, but has an active role to play in the cell operation. The separators also are thicker and exhibit a high porosity in order for the cell to function efficiently. Both the anions and the cations are intercalated into the respective electrodes, and for both species, graphite can be used. This is retained as the anode in the majority of cells constructed, but a number of alternative cathode materials have been tried, including polytriphenylamine.

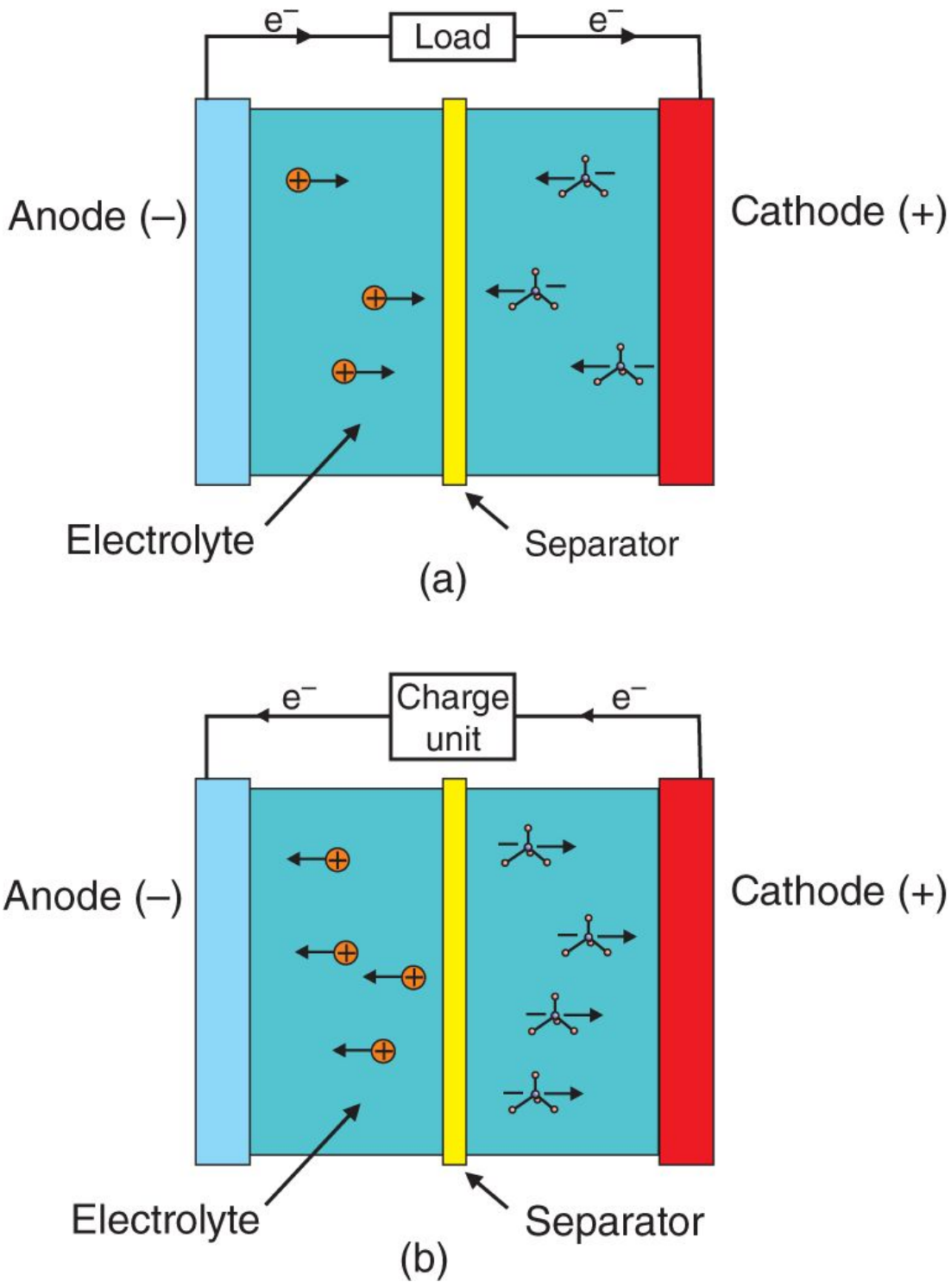


Figure 7.15 The advance and retreat operation of a dual-ion secondary battery, schematic: (a) in use; (b) recharging.

During charging, the cations (Li^+, K^+ etc.) are intercalated into the anode following the schematic mechanism:

$$K^+ + xC + e^- \rightarrow KC_x$$

The anions release an electron and are intercalated as a metastable species:

$$PF_6^- + xC \rightarrow (PF_6)C_x + e^-$$

The cathode material polytriphenylaniline (PTPAn) operates slightly differently in that this material releases an electron to form N^+ ions at sites along the polymer chain. These react with the anions that encounter the cathode to form $N^+PF_6^-$ units on the available surfaces:

$$PF_6^- + PTPAn \rightarrow [N^+(PF_6)^-] + e^-$$

In use, the discharge operation reverses these processes, so that intercalated metal atoms liberate an electron, generating cations (Li^+, Na^+, K^+), which return to the electrolyte:

$$KC_x \rightarrow K^+ + xC + e^-$$

Similarly, the intercalated anions gain an electron and are released into the electrolyte:

$$(PF_6)C_x + e^- \rightarrow PF_6^- + xC$$

$$[N^+(PF_6)^-] + e^- \rightarrow PF_6^- + PTPAn$$

There is a great deal of current research aimed at the improved performance of dual-ion batteries.

7.4 Corrosion

Corrosion refers to the degradation of a metal by *electrochemical reaction* with the environment. At room temperature, the most important corrosion reactions involve water, and the process is known as *aqueous corrosion*. (Corrosion at high temperatures in dry air, called oxidation, tarnishing, or direct corrosion, is described in Section 6.7.) Aqueous corrosion involves a set of complex electrochemical reactions in which the metal reverts to a more stable condition. The processes that take place depend upon the acidity of the water and any impurities present, so that the reactions that can take place are many and varied. For example, carbon steel reinforcing bars in concrete corrode more severely in acidic conditions and in the presence of chloride ions, a process called *electrochemical attack*. On the other hand, alkaline conditions inhibit the rate of corrosion.

From a practical point of view, corrosion may lead to loss of metal, leaving pits, or holes, or else the deposition of stable reaction products such as oxides or hydroxides, which may be crystalline or non-crystalline. Such corrosion products may be voluminous. In this case, they may force overlying protective layers away from the metal, and so allow corrosion to proceed unchecked, which exacerbates the damage.

In principle, corrosion is prevented by one of two methods, modification of the environment, which includes coating the metal, or by replacing the corrodible metal with a corrosion resistant metal. However, these simple remedies are not always possible and corrosion remains of economic importance across the world.

7.4.1 The Reaction of Metals with Water and Aqueous Acids

The reaction of a metal with an aqueous acid to yield hydrogen, a severe form of corrosion, involves oxidation of the metal and reduction of hydrogen ions in solution to H_2 gas, and so can be thought of in terms of an electrochemical cell.

The tendency for a reaction to occur follows the order of the electrochemical series, Table 7.1. Metals below H_2 in the electrochemical series:

$$2H^+ + 2e^- \rightarrow H_2 \quad E^0 = 0$$

that is, those with a negative standard reduction potential will react with aqueous acids to release hydrogen gas. Those above it will not react with acid. Thus, zinc will dissolve in acid to give hydrogen, while copper will not.

Such reactions are generally written in terms of the overall reaction:

$$Zn(s) + 2HCl(aq) \rightarrow ZnCl_2 + H_2(g)$$

However, more insight into the process is given by using the half-reactions:

$$2HCl(aq) \rightarrow 2H^+(aq) + 2Cl^-(aq)$$

$$Zn(s) + 2H^+(aq) \rightarrow Zn^{2+} + H_2(g)$$

$$Zn^{2+} + 2Cl^-(aq) \rightarrow ZnCl_2$$

The product, $ZnCl_2$, may be deposited as a poorly crystalline/amorphous solid.

The same information can also be used to predict which metals may dissolve in acid rain. For example, both lead and tin can enter the water supply in acid rain areas, and, for the same reason, acidic water will react with lead water pipes.

The reaction of metals with water can be examined using identical principles. Two reactions are important, oxidation, and reduction. When water acts as an oxidising agent, it is reduced to H_2. This is similar to the oxidation of a metal by an acid (H^+) to give H_2. Metals in the electrochemical series *below* the couple:

$$2H_2O(l) + 2e^- \rightarrow H_2(g) + 2OH^-(aq)$$
$$E^0 = -0.83\ V \text{ at pH } 14$$

will react with water and produce hydrogen gas. The metals Al, Mg, Na, K, and Li are the most reactive. For example, the reaction of magnesium with water is written:

$$Mg(s) + 2H_2O(l) \rightarrow Mg(OH)_2 + H_2(g)$$

and in terms of half reactions:

$$2H_2O(l) + 2e^- \rightarrow H_2(g) + 2OH^-(aq)$$

$$Mg(s) \rightarrow Mg^{2+}(aq) + 2e^-$$

$$Mg(s) + 2H_2O(l) \rightarrow H_2(g) + 2OH^-(aq) + Mg^{2+}(aq)$$

In practice, the reaction products from these low temperature processes are invariably ill-defined amorphous materials consisting of poorly soluble oxy-hydroxides.

Under the standard conditions ($E^0 = -0.83$ V), the concentration of the OH^- (aq) ions is 1 M and the pH is 14, very alkaline conditions indeed. In order to determine whether neutral water, at pH 7, will react, it is necessary to use the Nernst equation to redefine the reaction voltage.

$$2H_2O(l) + 2e^- \rightarrow H_2(g) + 2OH^-(aq)$$

$$E = -0.42\ \text{V} \quad \text{at pH} = 7$$

The revised reduction potential is −0.42 V. It is of interest to compare this with the reduction potential for iron. The Fe^{2+}/Fe couple has $E^0 = -0.44$ V, which is almost equal to that of neutral water. Thus, surprisingly, iron has no (or at best only a slight) tendency to be corroded by pure water. Corrosion of iron only takes place in water containing dissolved oxygen, discussed in the following text.

When water acts as a reducing agent, it is oxidised to O_2. The relevant equation is:

$$2H_2O(l) \rightarrow 4H^+(aq) + O_2(g) + 4e^-$$

This is the reverse of the reduction half reaction:

$$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$$

$$E^0 = +1.23\ \text{V} \quad \text{at pH} = 0$$

The presence of the H^+(aq) ions indicates that the water is acidic. The concentration of H^+(aq) in the standard state is 1 M, and the pH will be 0, equivalent to very acidic conditions. In this environment, water will be able to reduce redox couples above this reaction in the electrochemical series and liberate O_2. For example, the couple (Co^{3+}/Co^{2+}) has $E^0 = +1.82$ V so that Co^{3+} is reduced by acidified water to give O_2 thus:

$$Co^{3+}(aq) + e^- \rightarrow Co^{2+}(aq)$$

$$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$$

i.e.

$$4Co^{3+}(aq) + 2H_2O(l) \rightarrow 4Co^{2+}(aq) + O_2(g) + 4H^+(aq)$$

The reduction potential for neutral water can be calculated via the Nernst equation.

$$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$$
$$E = +0.81 \quad \text{at pH} = 7$$

There are a number of half-reactions involving non-metals that lie above this value in the electrochemical series and all will be reduced. In these cases, reaction will produce oxygen gas. For example, the dissolution of chlorine, Cl_2, in water will produce oxygen, although the unstable oxyacid, HOCl, forms as an intermediate. The reaction is:

$$2H_2O(l) \rightarrow 4H^+(aq) + O_2(g) + 4e^-$$

$$2Cl_2(g) + 4e^- \rightarrow 4Cl^-(aq)$$

$$2Cl_2 + 2H_2O \rightarrow 4HCl + O_2(g)$$

7.4.2 Dissimilar Metal Corrosion

Two different metals that are connected and immersed in an electrolyte form an electrochemical cell. If a current is allowed to flow, one metal will be consumed and one will remain the same or be increased in some way. These processes lead to *dissimilar metal corrosion*. In order for dissimilar metal corrosion to occur, it is necessary to have an anode, a cathode, an electrolyte, and a connection from anode to cathode. The anode component corrodes while the cathode remains unattacked. For example, buried steel pipes can be protected from corrosion by connecting them to blocks of metal such as magnesium, called *sacrificial anodes*, which corrode in preference to the pipe. The tendency for such reactions to take place spontaneously can be judged from the electrochemical series. Three examples follow.

Copper and iron or steel in juxtaposition can form a cell in which the copper becomes the cathode and the iron the anode. Several reactions are possible. One of these is:

Anode reaction:

$$Fe(s) \rightarrow Fe^{2+}(aq) + 2e^-$$

Cathode reaction:

$$O_2 + 2H_2O(l) + 4e^- \rightarrow 4(OH^-)$$

Cell reaction:

$$2Fe(s) + O_2 + 2H_2O(l) \rightarrow 4(OH^-) + 2Fe^{2+}(aq)$$

Fe^{2+} is soluble and the iron will gradually dissolve. Other related reactions can produce solid corrosion products that are usually ill-defined oxy-hydroxides of iron (Figure 7.16). The copper is not attacked and serves only to complete the cell. If there is a small anode area, such as an exposed nail head, the attack is more pronounced.

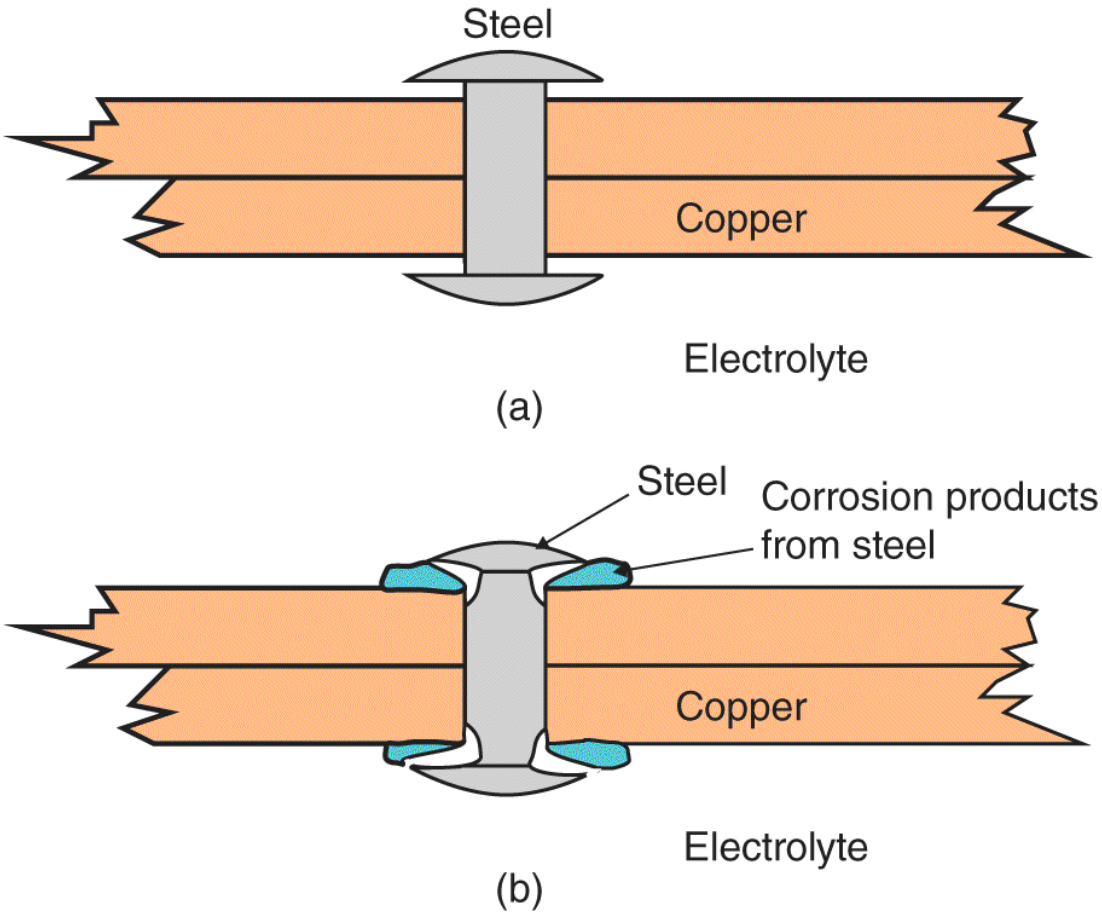


Figure 7.16 Dissimilar metal corrosion of a steel rivet in contact with copper and an electrolyte, schematic.

This form of corrosion has an interesting historical context. Wooden sailing ships were attacked below the water line by wood boring barnacles. Severe infestation could ultimately lead to the destruction of the bottom of the hull and catastrophic loss of the vessel. To prevent this, ships were sheathed in

copper, a practice that gave rise to the expression '*copper-bottomed*', meaning sound or reliable, still in use today. Unfortunately, iron or steel nails were often used to secure the copper sheathing. In the presence of alkaline water and oxygen (i.e. surface seawater, which has a pH of about 8.5 and a high content of dissolved oxygen), the nails corroded and the copper sheathing was often lost.

Coating steel sheet with zinc, a procedure called *galvanizing*, is widely used to prevent corrosion of the steel. The zinc coating does not corrode in air because initial reaction produces a dense layer of zinc oxide that protects the surface from further reaction. Should the zinc coating become penetrated, so that both zinc and steel are exposed to the air, corrosion is inhibited by the formation of a galvanic cell (Figure 7.17). In the presence of water and oxygen, the zinc will become the anode in the cell formed, and will corrode in preference to the exposed iron. Several reactions are possible, including:

Anode reaction:

$$Zn(s) \rightarrow Zn^{2+} + 2e^{-}$$

Cathode reaction:

$$O_2 + 2H_2O(l) + 4e^{-} \rightarrow 4(OH^{-})$$

Cell reaction:

$$2Zn(s) + O_2 + 2H_2O(l) \rightarrow 4(OH^{-}) + 2Zn^{2+}(aq)$$

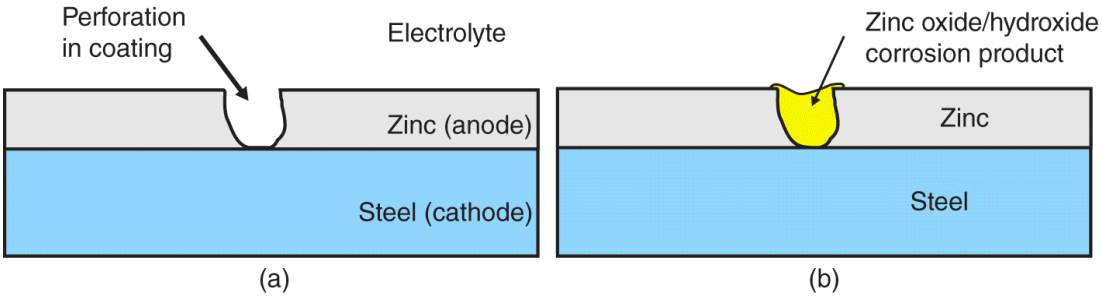


Figure 7.17 Protective galvanizing: (a) a perforation in the zinc coating causes an electrochemical cell to form; (b) the cell action seals the perforation with oxide/hydroxide corrosion products.

The Zn^{2+} ions react in ordinary conditions to produce ZnO or a zinc oxy-hydroxide. These are inert and form insoluble deposits that help to prevent further corrosion. Overall, steel coated with zinc corrodes far more slowly than bare steel.

Steel coated with tin was widely used on food cans, or 'tins', until replaced by aluminium or plastic coatings. Steel coated with tin corrodes faster than steel alone. Unlike the situation with zinc, a scratch in the coating allows an electrochemical cell to form in which the steel forms the anode and corrodes:

Anode reaction:

$$Fe(s) \rightarrow Fe^{2+} + 2e^-$$

Cathode reaction:

$$O_2 + 2H_2O(l) + 4e^- \rightarrow 4(OH^-)$$

Cell reaction:

$$2Fe(s) + O_2 + 2H_2O(l) \rightarrow 4(OH^-) + 2Fe^{2+}(aq)$$

No protective oxide forms and corrosion is enhanced compared with uncoated steel. This accounts for the fact that old tin cans on rubbish tips are always badly corroded.

7.4.3 Single Metal Electrochemical Corrosion

Two subtle corrosion effects can occur when a single metal is in contact with an electrolyte.

Differential aeration can cause corrosion when no obvious galvanic cells are in evidence. This can be illustrated by a cell with a copper anode and cathode. If the concentration of the electrolyte and the temperature of each cell compartment is the same, no potential is generated and no corrosion occurs. However, bubble O_2 into the one compartment, which becomes the cathode compartment, and corrosion will occur in the other, which forms the anode compartment. Electrons will flow from anode to cathode and the anode will corrode:

Anode reaction:

$$Cu(s) \rightarrow Cu^{2+} + 2e^-$$

Cathode reaction:

$$O_2 + 2H_2O + 4e^- \rightarrow 4(OH^-)$$

Cell reaction:

$$2Cu(s) + O_2 + 2H_2O(l) \rightarrow 4(OH^-) + 2Cu^{2+}(aq)$$

Differential aeration is, in fact, a concentration effect, and can be understood by using the Nernst equation.

This type of corrosion can happen within a water drop on steel. The surface contains more dissolved oxygen than the interior of the drop and creates a circular cathode (Figure 7.18). The less aerated centre forms an anode, and corrosion produces a pit at the centre of the drop.

Anode reaction:

$$Fe(s) \rightarrow Fe^{2+} + 2e^-$$

Cathode reaction:

$$O_2(g) + H_2O(e) + 4e^- \rightarrow 4(OH^-)$$

Cell reaction:

$$2Fe(s) + O_2 + 2H_2O(l) \rightarrow 4(OH^-) + 2Fe^{2+}(aq)$$

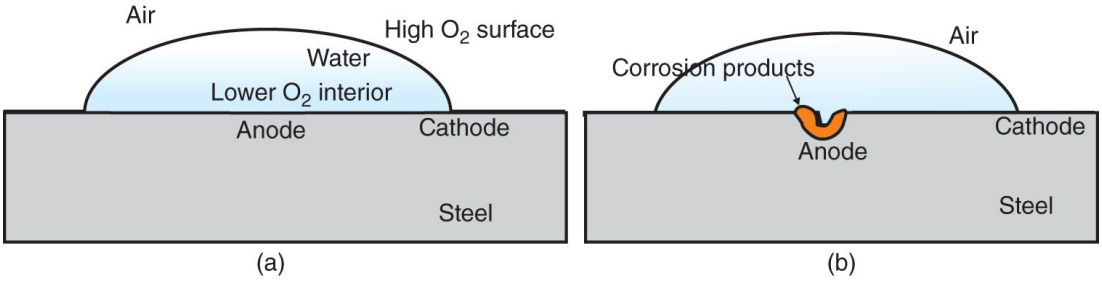


Figure 7.18 Differential aeration leading to pitting in steel: (a) initial situation, a cathode forms at the oxygen-rich circumference of the drop and an anode at the oxygen-poor centre; (b) pitting and corrosion after reaction.

Similar corrosion effects can be seen in narrow fissures, termed *crevice corrosion*. For example, narrow channels between a damp steel rivet and a damp plate can receive less oxygen than the surface of either. The crack becomes anodic and corrosion may occur (Figure 7.19). This problem is often enhanced when the corrosion product has a high volume. The resulting stress may lever the rivet head off.

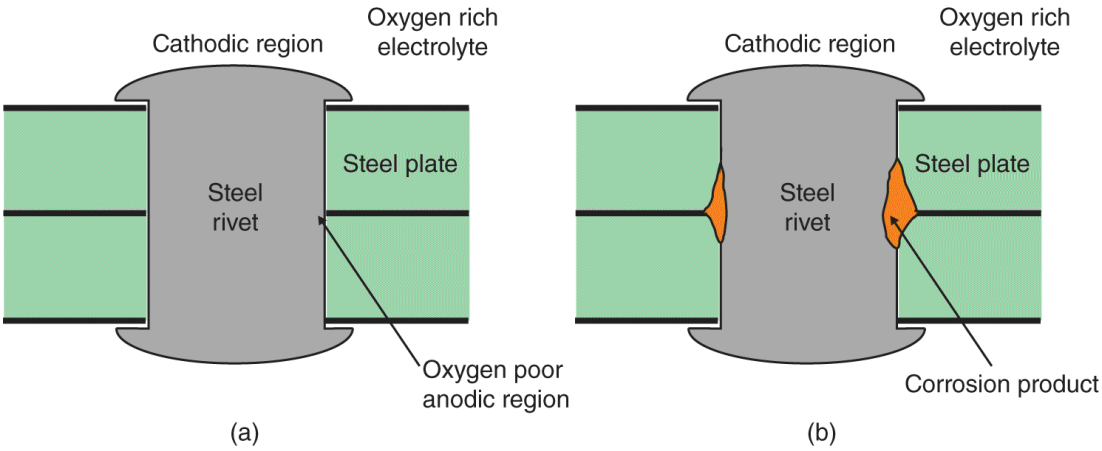


Figure 7.19 Crevice corrosion due to differential aeration: (a) initial situation, a cathode forms at the outer oxygen-rich region rivet bordering the electrolyte and an anode in the oxygen-poor crevice; (b) after corrosion.

Corrosion of a single metal can also occur even in the absence of significant differential aeration. This puzzling occurrence is due to the presence of anodic and cathodic regions on the metal. These can be generated during heat treatment and cold working of metals. For example, the regions of a metal subjected to cold working are often anodic compared with the remainder of the material. In contact with an electrolyte these areas will tend to corrode due to the formation of a galvanic cell, even though no concentration effects exist.

7.5 Electrolysis

While a galvanic cell uses a spontaneous chemical reaction to produce an electric current, an *electrolytic cell* uses an electric current to drive a non-spontaneous chemical reaction. A rechargeable battery thus operates as a galvanic cell when being used and as an electrochemical cell when being charged. The process occurring in an electrochemical cell is called *electrolysis*. Electrolytic cells are widely used in the preparation of chemicals such as magnesium and aluminium, and in electroplating. Recently, the electrolytic production of hydrogen as a component of the non-fossil fuel economy has become increasingly important.

7.5.1 Electrolytic Cells

Electrolytic cells do not need the electrodes in separate compartments and so are often simpler in construction than galvanic cells (Figure 7.20). However, the reactions at the anode, oxidation, and at the cathode, reduction, are identical to those in a galvanic cell. Similarly, during operation, electrons from the external supply enter the cell via the cathode, and leave it via the anode, as in a galvanic cell. Cations in the electrolyte move away from the anode and towards the cathode while anions in the electrolyte move away from the cathode and towards the anode. The anode of an electrolytic cell is labelled + and the cathode –, whereas in a galvanic cell the reverse is true, and the anode is labelled – and the cathode +.

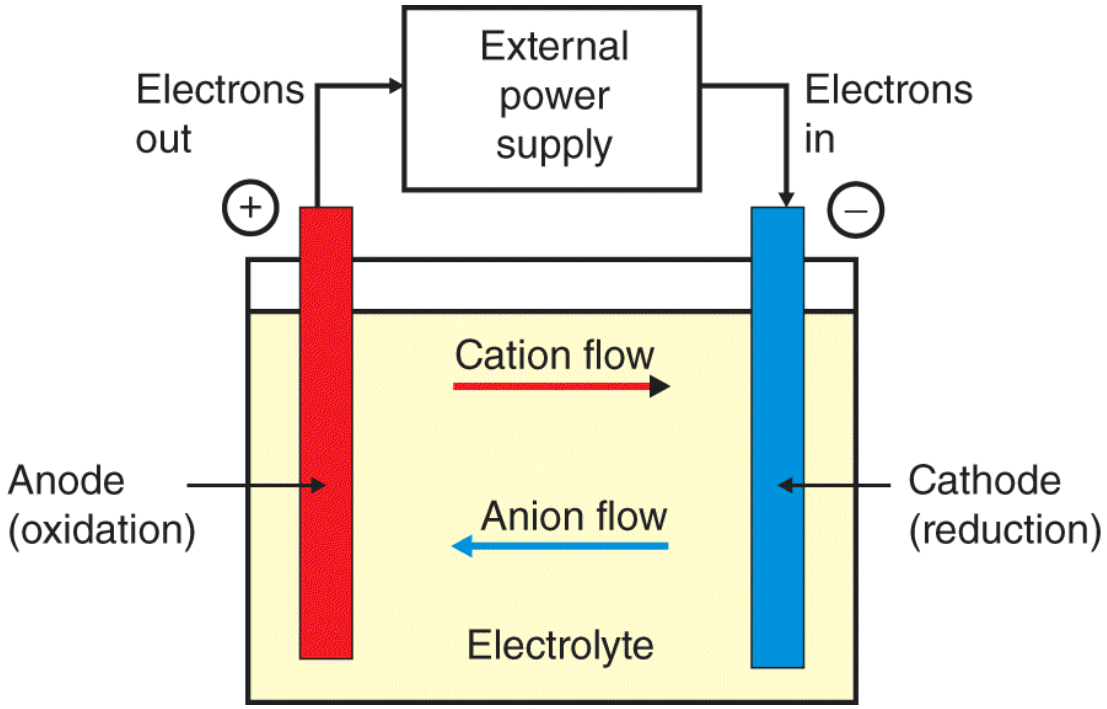


Figure 7.20 The components of an electrolytic cell.

The potential that has to be supplied to make the non-spontaneous reactions occur must be (in principle) the reverse of the potential generated by the spontaneous reaction, that is, equal and opposite to *E*. In practice, this is a minimum that generally has to be exceeded in a working cell. The actual amount of extra potential to be supplied, the *overpotential*, is a function of the electrode materials and conditions at the electrode surfaces. Moreover, in cells where the electrolyte contains several species that could be oxidised or reduced, those requiring the least input of energy will preferentially react.

The electrolytic cells used to produce hydrogen gas from water, as part of a route to the production of 'clean' energy are generally called electrolysers. These are powered by 'surplus' power from wind, solar energy, tidal power, or similar. The hydrogen so produced is stored and later used to produce electricity, in, for example, a fuel cell. The structure of the electrolyser unit is similar to that of fuel cells, Figure 7.10, and includes polymer electrolyte membrane (PEM) electrolysers, alkaline electrolyte electrolysers, and solid oxide electrolysers. In practice these electrolysers are structurally complex, using anodes and cathodes of a variety of materials which are often coated with catalyst nanoparticles such as platinum, to facilitate reactions.

In a polymer membrane cell, water reacts at the anode (+) to produce oxygen gas:

Anode reaction:

$$2H_2O(l) \rightarrow O_2(g) + 4H^+ + 4e^-$$

The H^+ ions cross the polymer membrane to the cathode, and electrons arrive at the cathode (−) via the external circuit to produce hydrogen gas:

Cathode reaction:

$$4H^+(aq) + 4e^- \rightarrow 2H_2(g)$$

Cell reaction:

$$2H_2O(l) \rightarrow O_2(g) + 2H_2(g)$$

In a solid oxide electrolyte containing cell, containing yttria stabilised zirconia or a similar oxide ion superionic conductor (Section 5.13) water input to the cathode combines with electrons from the external circuit to give hydrogen gas and oxide ions. These are transported across the stabilised zirconia to the anode, where they react to give oxygen gas and electrons that return via the external circuit.

Anode reaction:

$$2O^{2-} \rightarrow O_2(g) + 4e^-$$

Cathode reaction:

$$H_2O(g) + 2e^- \rightarrow H_2(g) + O^{2-}$$

In both types of cell, hydrogen is evolved at the cathode and oxygen at the anode, so that these electrodes are often called the 'hydrogen electrode' and the 'oxygen electrode', respectively.

7.5.2 Electroplating

Electroplating is the deposition of a metallic coating onto a metal object using electrolysis. It is widely carried out both for decorative purposes and for corrosion prevention. Although the principles of electroplating do not differ from those given previously, in practice the production of a high-quality film is critically dependent upon a number of factors, especially the cleanliness of the surface to be plated. Metal ions are reduced in the order of high to low standard reduction potential, Table 7.1. Hence Li^+ will be the first cation in a mixture to be reduced to metal, as the Li/Li^+ couple has the lowest potential, −3.05 V. However, alloys that form between metals with similar standard reduction potentials, say, Li co-doped with Na or K, can be deposited by tuning the concentration of the ions in solution and the addition of selected additives. Plating solutions are ideally water-based, but the alkali metals in particular react with water and these must also employ non-aqueous solvents in the cell.

The form of the electroplated film can be variable and will depend not only on the electrolyte but also on the additives that aid the formation of a coherent film. A coherent solid film that adheres strongly to the substrate is usually desired. Not all metals will deposit in such a controlled fashion. Lithium, for example, causes problems (especially important in the operation of Li-ion batteries) because it tends to plate in a dendritic form, consisting of thin filamentary growths that resemble the structure of a deciduous tree in winter. Co-doping the films with other metals, Na in the case of Li, can alleviate the problem, as can solution agitation and modification of the substrate and current applied. In addition, commercial plating solutions contain organic additives to enhance film adherence and to encourage solid film deposition rather than dendritic growth habits.

The process can be illustrated by the *schematic* description of nickel plating (Figure 7.21). The metal object to be plated is connected to negative input from a direct current source, so as to form the cathode of the cell. The electrolyte is a solution of a soluble nickel salt in water, $NiCl_2$, for example. The anode of the

cell is a rod of nickel metal. The current drives the nickel ions in solution towards the cathode, where they are deposited:

$$Ni^{2+}(aq) + 2e^- \rightarrow Ni(s) \quad E^0 = +0.23\,V$$

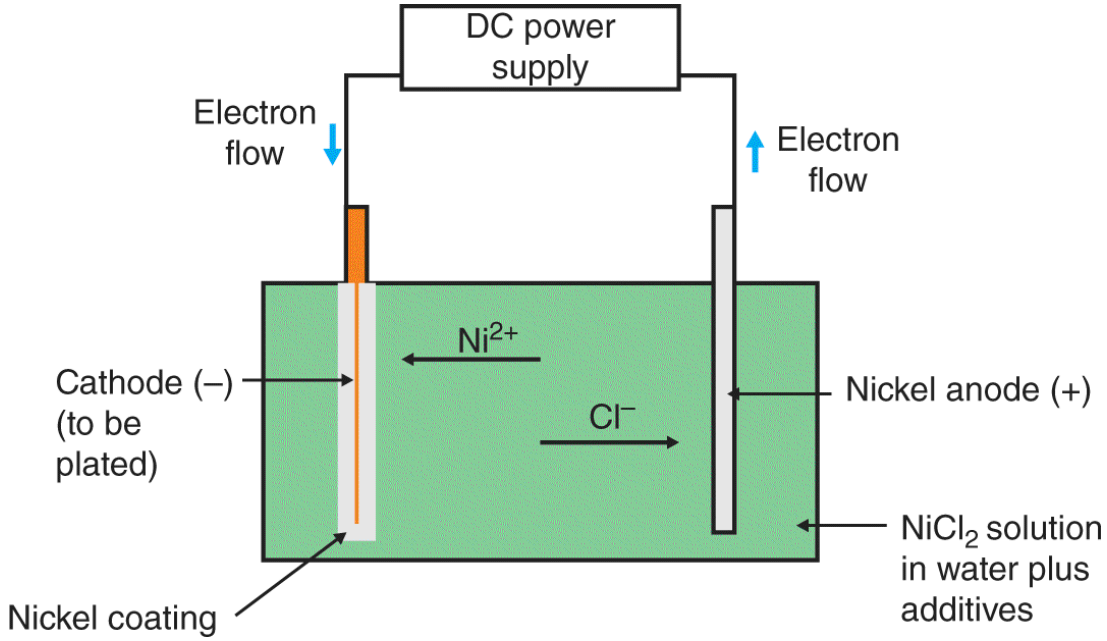


Figure 7.21 Nickel electroplating: the nickel anode is dissolved and transported to the object to be plated, the cathode.

Two moles of electrons must be supplied to deposit 1 M of nickel. At the anode, the chloride ions are discharged, to form chlorine gas:

$$2Cl^-(aq) \rightarrow Cl_2(g) + 2e^- \quad E^0 = +1.36\,V$$

Simultaneously the nickel anode dissolves in the process:

$$Ni(s) \rightarrow Ni^{2+}(aq) + 2e^-$$

The overall result is the transfer of nickel from the anode to the cathode:

$$Ni(s)\ anode \rightarrow Ni(s)\ cathode$$

It is this same process that causes the corrosion of anodes in all batteries. As a first approximation the balanced reactions are:

Anode reaction:

$$Ni(s) + 2Cl^-(aq) \rightarrow Ni^{2+}(aq) + Cl_2(g) + 4e^-$$

Cathode reaction:

$$2Ni^{2+}(aq) + 4e^- \rightarrow 2Ni(s)$$

Plating reaction:

$$2Cl^-(aq) + Ni^{2+}(aq) \rightarrow Ni(s) + Cl_2(g)$$

7.5.3 The Amount of Product Produced During Electrolysis

The *chemical nature* of the product of electrolysis is determined by the reduction potential of the appropriate redox couple. The *amount* of product formed depends only upon the *amount of electricity* that has passed through the cell. This fact was first recognised by Faraday, who formulated what are now known as *Faradays laws of electrolysis*:

1. The mass of substance produced at an electrode is directly proportional to the quantity of electricity that has passed through the cell.
2. The mass of a substance produced by a given quantity of electricity is directly proportional to the molar mass of the substance and inversely proportional to the numbers of electrons transferred per molecule of the substance.

The amount of electricity that is provided in an electrolysis experiment, Q, is given by:

$$Q = I\,t$$

where I is the current and t the time. One mole of electrons has a charge:

$$(1.6022 \times 10^{-19}\ \text{C}) \times (6.0222 \times 10^{23}\ \text{mol}^{-1})$$
$$= 9.6485 \times 10^{4}\ \text{C mol}^{-1}$$

This is called the *Faraday constant*, F. Thus, the quantity of electricity needed to produce one mole of a monovalent element is 9.6485×10^4 C. Double this amount is needed for a divalent element and so on. The number of moles of electrons provided in any electrolysis experiment, Q_m, is given by:

$$Q_m = I\,t/F$$

To obtain the mass of an element that is formed, it is necessary to multiply by the molar mass produced by 1 M of electrons:

$$m = [I\,t/F]\,(M/Z)$$

where m is the mass produced, M is the molar mass of the element, and Z is the charge on the ion involved.

7.5.4 The Electrolytic Preparation of Titanium by the FFC Cambridge Process

The preparation of titanium metal uses the Kroll process, which involves the reduction of titanium tetrachloride with magnesium metal. The method uses corrosive and dangerous chemicals and makes the metal in batches rather than continuously, with the result that the titanium metal is very expensive. Because of this, schemes using electrolysis of molten Ti salts, similar to the production of many other metals, have been widely investigated. Although metal can be produced this way, it is often dendritic in form, consisting of thin many branched whiskers, and very reactive, oxidising on contact with air. A process called the *FFC Cambridge Process* uses a slightly different electrochemical approach. The key to the method, and what distinguishes it from earlier electrolysis attempts, lies in the use of slightly non-stoichiometric titanium dioxide, TiO_{2-x}, as the cathode in an electrochemical cell.

Titanium dioxide itself is an insulator but is able to lose oxygen easily if heated under a low oxygen pressure to give a non-stoichiometric oxide. The material produced is found to be an electronic conductor, like other lower titanium oxides such as Ti_2O_3 and TiO. Even a composition as close to TiO_2 as $TiO_{1.995}$ conducts electricity well.

Thus, the idea behind the electrolysis is not to dissolve the oxide in a flux, and transport titanium ions, but to make the cathode of pellets of slightly reduced rutile, and transport oxygen ions away from it to further increase the degree of reduction of the cathode. The electrolyte is molten $CaCl_2$ and graphite is used as the anode (Figure 7.22). During electrolysis, oxygen is pulled out of the titanium oxide cathode, and transported through the electrolyte to the graphite anode, where some of it reacts to form carbon monoxide and carbon dioxide while the remainder is released as oxygen gas.

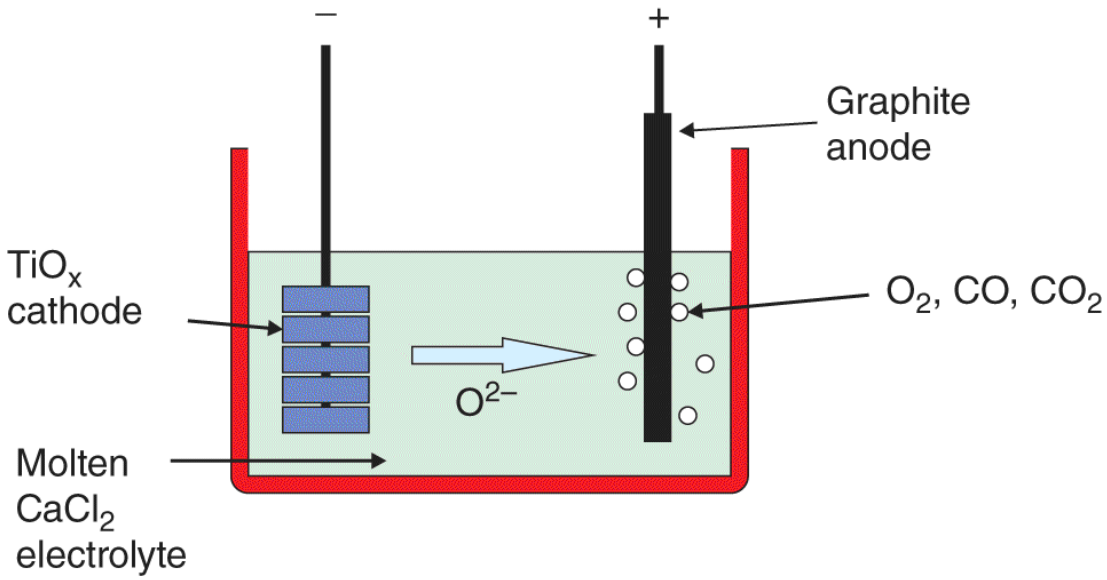


Figure 7.22 The FCC Cambridge cell for the production of titanium, schematic.

The oxide cathode is gradually converted to pellets of titanium metal in a sponge-like form. This material does not oxidise easily and can be melted and turned into ingots with minimum additional processing.

The cell reaction can be written in a simplified form as:

Anode reaction:

$$xC + xO^{2-} \rightarrow xCO(g) + 2xe^{-}$$

Cathode reaction:

$$TiO_x + 2xe^{-} \rightarrow xO^{2-} + Ti$$

Electrolysis reaction:

$$TiO_x + xC \rightarrow Ti + xCO(g)$$

In reality, carbon monoxide, carbon dioxide, and oxygen are produced at the anode. These anode reactions can be approximated by:

$$aC + aO^{2-} \rightarrow aCO(g) + 2ae^{-}$$

$$bC + 2bO^{2-} \rightarrow bCO_2(g) + 2be^{-}$$

$$cO^{2-} \rightarrow (c/2)O_2(g) + 2ce^{-}$$

where $(a + 2b + c)$ is equal to x in TiO_x, that is, the cell reaction is:

$$TiO_{a+2b+c} + (a + b)C(s) \rightarrow Ti + aCO(g) + bCO_2(g) + (c/2)O_2(g)$$

This process works well with a number of other oxides that are difficult to convert into metals by conventional redox methods, including Cr_2O_3, ZrO_2, Nb_2O_5, Ta_2O_5, and WO_3, all of which can be produced in an electronically conducting form by small degrees of reduction. Moreover, if the cathode is made of solid solutions or mixed oxides, alloys can be produced directly. Using this technique, both simple alloys such as $TiAl_3$ and Ni_3Ti, and more complex compounds, such as Ti_6Al_4V, have been synthesised.

7.6 Pourbaix Diagrams

7.6.1 Passivation, Corrosion, and Leaching

Many reactions that occur in water are sensitive to acidity, to electrolyte concentration and to the relative oxidising or reducing conditions at the reaction site. Metallic corrosion does not always occur and under certain combinations of acidity and reduction potential iron, copper, zinc, and other metals are corrosion resistant. This feature is called *passivation*. The dispersal of dangerous metals in the environment is similarly influenced by the same factors. If a nickel–cadmium battery is thrown onto a landfill site, will the toxic cadmium be leached away into streams or rivers, or will it remain in place? The disposal of radioactive wastes raises identical questions – how permanent is the repository with respect to the presence of water?

In order to obtain answers to these questions, it necessary to write down all of the possible half reactions that can be envisaged and then determine how these will vary with acidity, concentration, and oxidation potential. It is tedious to carry out these calculations and the results are often not especially lucid. However, the overall scheme of reactivity can be represented graphically on *Pourbaix diagrams*. Although initially formulated to assist in demarcation of corrosion resistant conditions for metals, they have found applicability in many other areas, including electrochemistry, earth sciences, chemical engineering, and metallurgy.

A Pourbaix diagram uses the oxidising/reducing potential and the acidity of the environment as parameters to quantify the reactivity of the system under consideration, usually for aqueous environments. The oxidising capability is plotted on the ordinate (y-axis) as a voltage. The use of voltage to express oxidation and reduction is simply an adaptation of the electrochemical series, and the voltage used is that of the half-reaction measured against a standard hydrogen electrode. A table of reduction half-reactions and associated voltages is also a table of relative oxidising and reducing capabilities. The acidity is plotted on the abscissa (x-axis) as pH. The area of the diagram is divided up into stability fields, which show where a certain species is stable. As corrosion and leaching require the presence of water the area of the diagram in which water is stable is emphasised.

Pourbaix diagrams are derived from thermodynamic data: standard reduction potentials and reaction equilibrium constants. Because of this, these diagrams are now routinely constructed using thermodynamic software.

7.6.2 The Stability Field of Water

The region over which water is stable, the *stability field of water*, is defined as the range of pH and oxidation/reduction potential over which water is stable to

both oxidation and reduction at 25 °C and 1 atm. pressure (Figure 7.23). Above the upper boundary, water is oxidised to O_2 gas, and below the lower line it is reduced to H_2 gas.

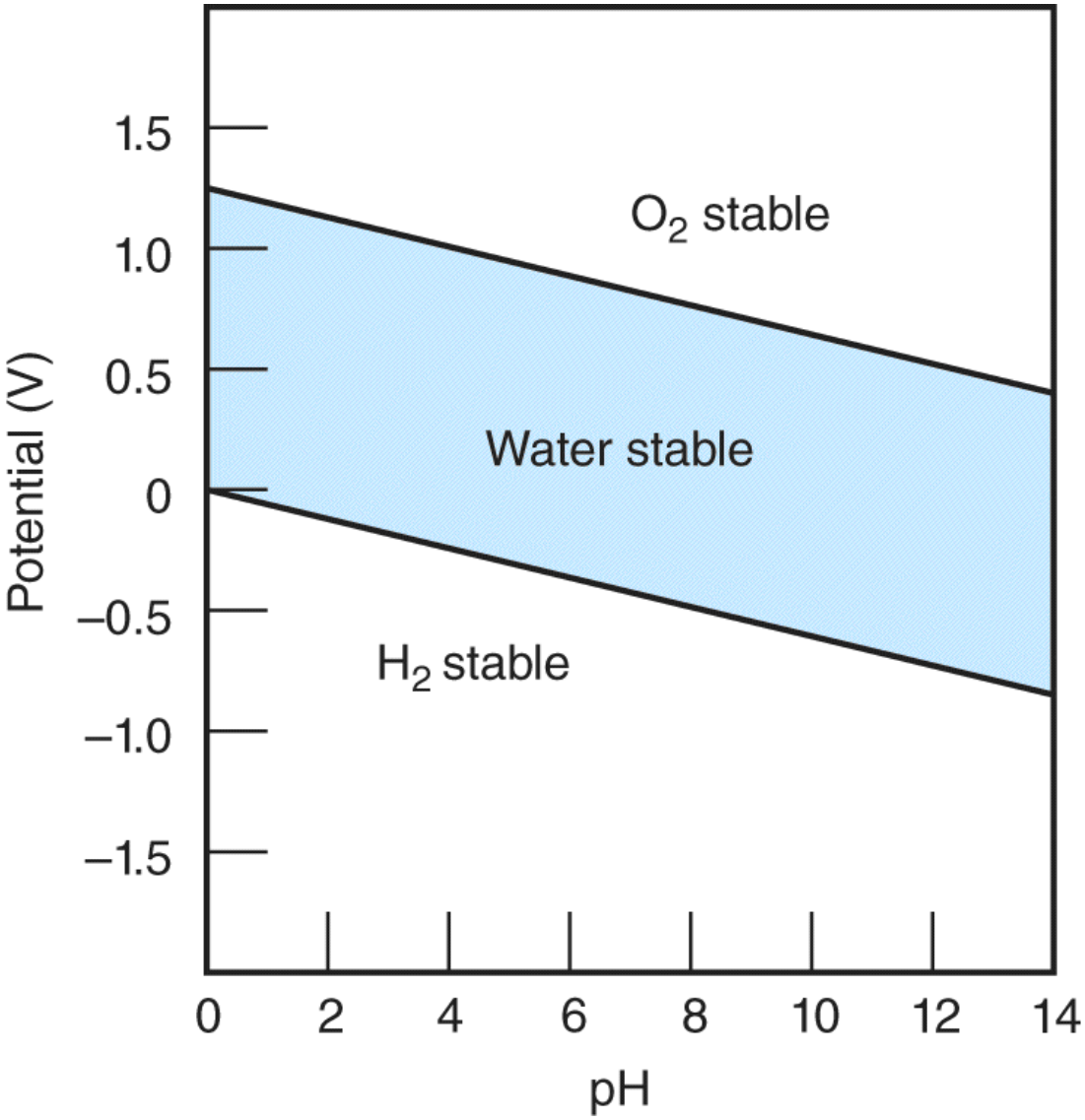


Figure 7.23 Pourbaix diagram showing the stability field of water.

The starting point in plotting the stability field of water is the Nernst equation (Section 7.1.3) in the form:

$$E = E^0 - (0.05916/n)\ \log\ Q \quad \text{at } 25\,^\circ\text{C}.$$

The *oxidation* of water is defined by the half-reaction:

$$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$$

$$E^0 = +1.23\ V$$

In this case, $n = 4$ and $Q = 1/(p_{O_2}[H^+]^4)$

Substituting into the Nernst equation:

$$\begin{aligned} E &= 1.23 - (0.05916/4) \\ &\quad \times \log\ (1/p_{O_2}[H^+]^4) \\ &= 1.23 + (0.014\,79)\ \log(p_{O_2}[H^+]^4) \\ &= 1.23 + (0.014\,79)\ \log\ p_{O_2} \\ &\quad + (0.05916)\ \log[H^+] \end{aligned}$$

For the normal oxygen partial pressure, p_{O_2}, of 0.21 atm, we find:[3]

$$E = 1.22 - 0.05916\ \text{pH (volts)} \tag{7.1}$$

The *reduction* of water is defined by the half-reaction

$$2H_2O(l) + 2e^- \rightarrow H_2(g) + 2OH^-(aq)$$

$$E^0 = -0.83\ V$$

In this case $n = 2$ and $Q = p_{H_2}[OH^-]^2$

$$E = E^0 - (0.059164/2)\ \log\{p_{H_2}[OH^-]^2\}$$

Taking the standard pressure of H_2 (1 atm):

$$E = -0.83 - 0.059164\ \log\,[OH^-]$$

Converting to pH[4]

$$E = -0.05916\ \text{pH} \tag{7.2}$$

Equations (7.1) and (7.2) are plotted on the Pourbaix diagram. The stability field of water is the area between the two lines.

7.6.3 Pourbaix Diagrams for a Metal Showing Two Valence States

Transition metals display several valence states. Generally, the valence state that is stable depends upon the acidity and the oxidation potential in the environment. To illustrate the general usefulness of Pourbaix diagrams in resolving practical problems such as the likelihood of corrosion of bulk metal, or precipitation of a solid phase from solution, as well as the stable valence state of the cations, this and the Sections 7.6.3 and 7.6.4 consider the behaviour of iron and its compounds.

Iron is present in the Earth's core as liquid metal, Fe^0. In the mantle, or in reducing conditions in sediments, Fe is present as Fe^{2+} (Fe(II), ferrous). In oxidising conditions, Fe exists as Fe^{3+} (Fe(III), ferric). The commonest Fe containing minerals, Fe_2O_3 (haematite, Fe^{3+}), Fe_3O_4 (magnetite, lodestone, Fe^{2+}, Fe^{3+}), and $FeCO_3$ (siderite, Fe^{2+}), reflect different formation conditions for the minerals in the Earth's crust. Irrespective of the origin, all iron compounds tend to the stable Fe^{3+} state in air. Iron itself corrodes in moist air and reacts with non-oxidising acids to yield H_2 and Fe(II), which is subsequently oxidised to Fe(III) in air. This is a slow reaction in acidic solution and rapid in a basic solution, when insoluble oxy-hydroxides, typically labelled $Fe(OH)_3$, are precipitated. These confusing relations are most easily understood via the Pourbaix diagram for iron (Figure 7.24).

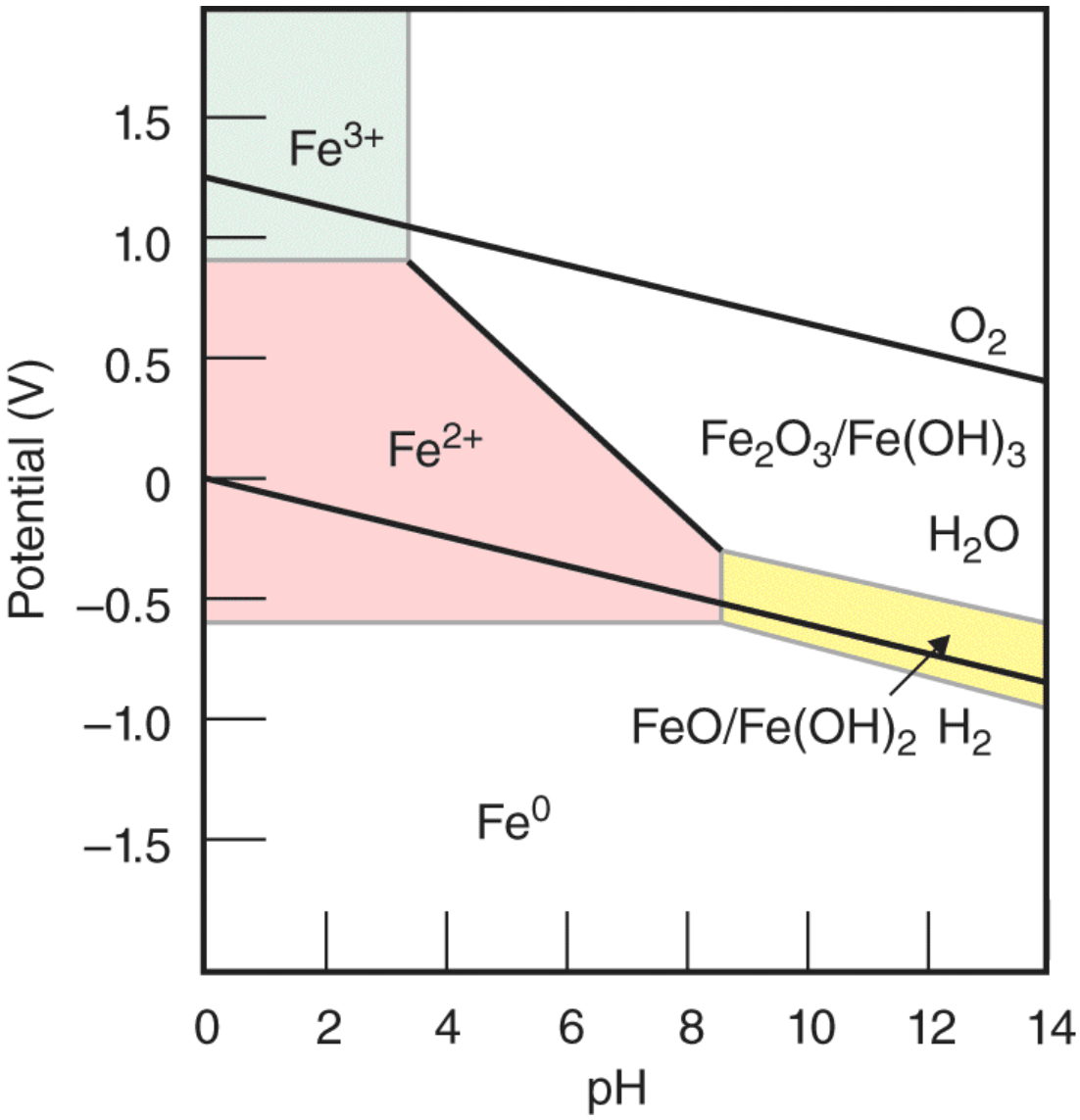


Figure 7.24 Pourbaix diagram showing the stable species in the iron–water–oxygen system.

The upper left-hand corner of the diagram represents conditions that are oxidising and acidic. Under these conditions the higher valence state, Fe^{3+}, is the stable form. The region over which this remains true is called the *stability field* of the Fe^{3+} ions. [Note that in reality, the stable species may be a hydrated ion such as $Fe(H_2O)_6^{3+}(aq)$].

When alkali is added to a solution containing Fe^{3+}, a precipitate of hydroxide is formed. This is represented by an increase in pH and a move towards the right in the diagram. Ultimately a solution of Fe^{3+} is replaced by the oxide, Fe_2O_3, or hydroxide, $Fe(OH)_3$, as the stable species:

$$Fe^{3+}(aq) + 3H_2O(l) \rightarrow Fe(OH)_3(s) + 3H^+(aq)$$

On the diagram this is represented by a new stability field. The change in acidity is revealed by the formation of $H^+(aq)$, but the oxidation state of the Fe^{3+} valence has not changed during this transformation, and the metal is trivalent in both Fe_2O_3 and $Fe(OH)_3$. (Note that in real systems the material that forms in solution is often an ill-defined oxy-hydroxide, and not a simple compound.) Somewhere between the high and low pH regions, a boundary exists that separates the two stability fields. This boundary will be a *vertical line* as these separate stability fields that involve a change in acidity (H^+ or OH^- concentration) and no change in oxidation state (e^- transfer). The position of this line will depend upon the thermodynamic equilibrium constant, *K*, of the reaction given earlier:

$$K = [H^+]^3/[Fe^{3+}]$$

$$\text{so} \quad \log K = 3\log[H^+] - \log[Fe^{3+}]$$
$$= -3\text{pH} - \log[Fe^{3+}]$$

The value of log *K* is available from the thermodynamic literature and so the value of the pH boundary is readily calculated for any specified concentration of Fe^{3+}.

Return to the Fe^{3+} stability field. As the oxidising power of the environment decreases, that is, as the voltage on the ordinate decreases, one moves towards the bottom left of the diagram, representing more reducing conditions. Ultimately the lower valence state, Fe^{2+}, becomes the stable species. The half-reaction is:

$$Fe^{3+} + e^- \rightarrow Fe^{2+} \quad E^0 = +0.77\,\text{V}$$

The half-reaction also confirms that there is no change in acidity. The boundary between the two stability fields will now be *horizontal* as these separate stability fields that involve a change in oxidation state (e^- transfer) and no change in acidity (H^+ or OH^- concentration). Under standard conditions the boundary lies at 0.77 V. As long as the Fe^{3+} concentration is fairly close to the Fe^{2+} concentration, this line will always be close to 0.77 V, but concentrations different from this may be calculated by inserting the true values in the appropriate form of the Nernst equation. No matter where this line is positioned, it will still remain horizontal, of course.

A continued reduction in the oxidising potential, moving further towards the lower left of the diagram, causes the Fe^{2+} ion to be replaced by more stable metal, Fe, which is written Fe^0 when this state needs to be emphasised:

$$Fe^{2+} + 2e^- \rightarrow Fe(s) \quad E^0 = -0.44\,V$$

The boundary between the stability fields for Fe^{2+} and Fe^0 is horizontal, for the reason given above and is at −0.44 V under standard conditions.

Return to the lower part of the Fe^{2+} stability field and consider the consequence of decreasing the acidity (increasing the pH). In this case a precipitate of hydroxide $Fe(OH)_2$ or the oxide FeO will form. That is, the Fe^{2+} stability field gives way to one in which a solid is preferred:

$$Fe^{2+}(aq) + 2H_2O(l) \rightarrow Fe(OH)_2(s) + 2H^+(aq)$$

The boundary between the stability fields is vertical as no change in oxidation state is involved and its position is given by the appropriate equilibrium constant and Fe^{2+} concentration:

$$K = [H^+]^2/[Fe^{2+}]$$

$$\text{so} \quad \log K = 2\log[H^+] - \log[Fe^{2+}]$$
$$= -2pH - \log[Fe^{2+}]$$

Repeat this in the upper part of the Fe^{2+} stability field. The conditions then correspond to a decrease in acidity under oxidising conditions. The Fe^{2+} stability field will now give way to a field in which either oxide, Fe_2O_3, or hydroxide, $Fe(OH)_3$, are preferred.

$$2Fe(OH)_3(s) + 6H^+(aq) + 2e^-$$
$$\rightarrow 2Fe^{2+}(aq) + 6H_2O(l) \quad E^0 = 0.94\,V$$

In this case, the stability field boundary marks a change in *both* oxidation state and acidity. The redox nature of the reaction is revealed by the production of electrons and the acid–base nature of the reaction by the production of H^+. *Sloping boundaries* separate stability fields that involve both processes. Near such boundaries, small changes in pH, or oxidation conditions can make a large difference in whether an ion will stay in solution or transform into a solid.

A more generalised form of the diagram for a metal displaying two valence states, M^{2+}/M^{3+}, will take other potential reactions into account. For example:

$$M^{3+}(OH)_3(s) + H^+(aq) + e^- \rightarrow M^{2+}(OH)_2(s) + H_2O$$

gives the boundary between $M(OH)_3$ and $M(OH)_2$ stability fields, and:

$$M^{2+}(OH)_2(s) + 2H^+(aq) + 2e^- \rightarrow M^0(s) + 2H_2O(l)$$

gives the boundary between the stability fields of $M(OH)_2$ and metal M (Figure 7.25).

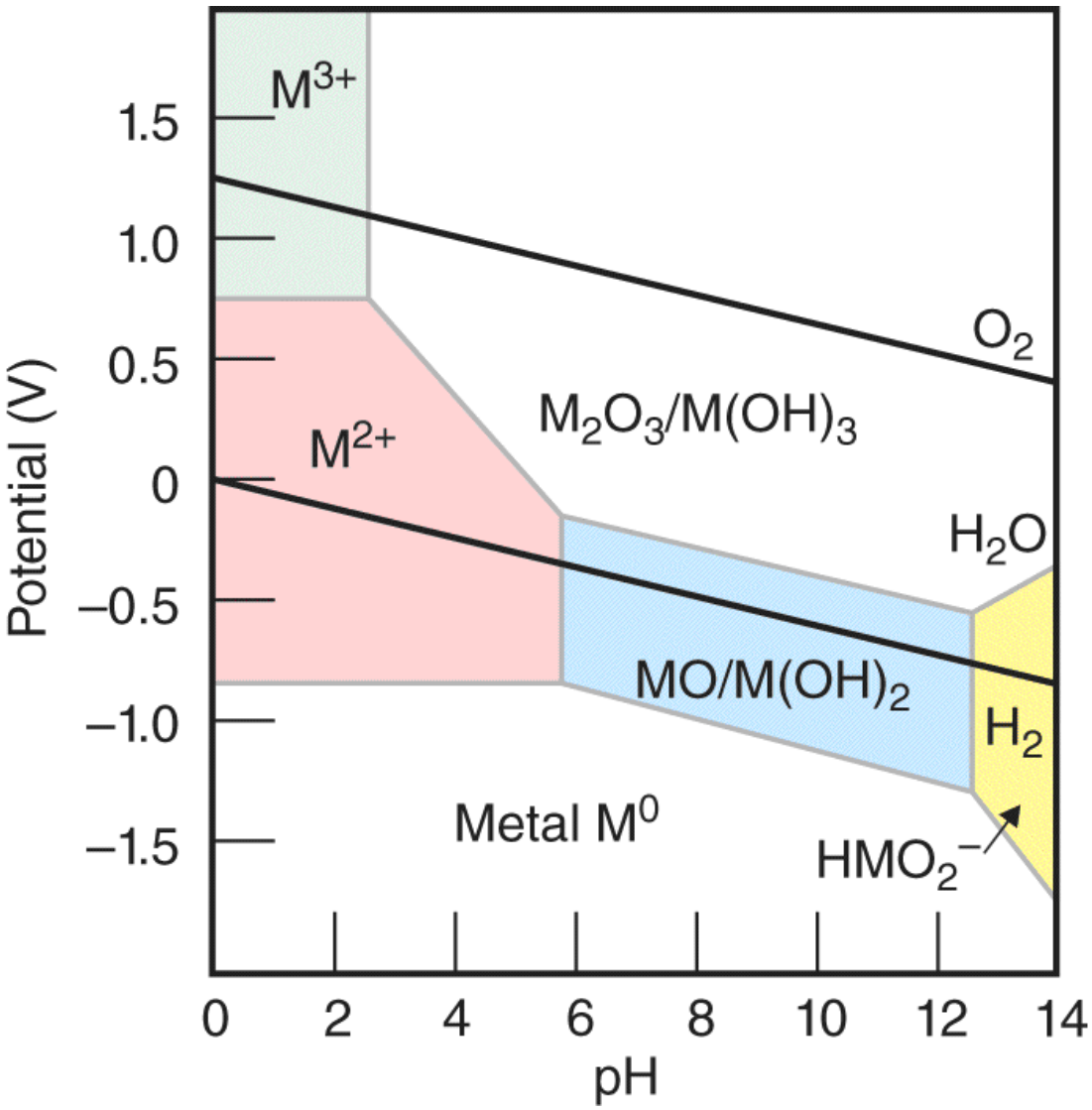


Figure 7.25 Generalised Pourbaix diagram showing the stable species in a system containing a metal capable of two valence states, M^{2+} and M^{3+}.

The Pourbaix diagram includes the stability field of water and so reactions that have importance in the environment are readily distinguished from those that may only occur in more extreme conditions such as deep below the surface of the Earth. With this in mind, recall that any Pourbaix diagram is drawn for specific concentrations of ions. In order to display how the stability fields change with concentration, a third axis, concentration, must be added normal to the *E* and pH axes. The areas then become volumes in this representation.

7.6.4 Pourbaix Diagram Displaying Tendency for Corrosion

Corrosion is likely when the metal is in an environment corresponding to the stability fields in which aqueous ions are stable (Fe^{3+} and Fe^{2+} ions in Figure 7.24). The regions in which a solid occurs are less likely to corrode extensively as the initial formation of a precipitate may prevent further corrosion from taking place. In these stability fields the metal is said to be passivated. Under greatest reducing conditions the metal is stable and not liable to corrode at all. The range of oxidation and pH conditions under which the metal would be immune to corrosion, passivated, or corrode, can be used to label the stability fields on the Pourbaix diagram, if that is of prime importance (Figure 7.26).

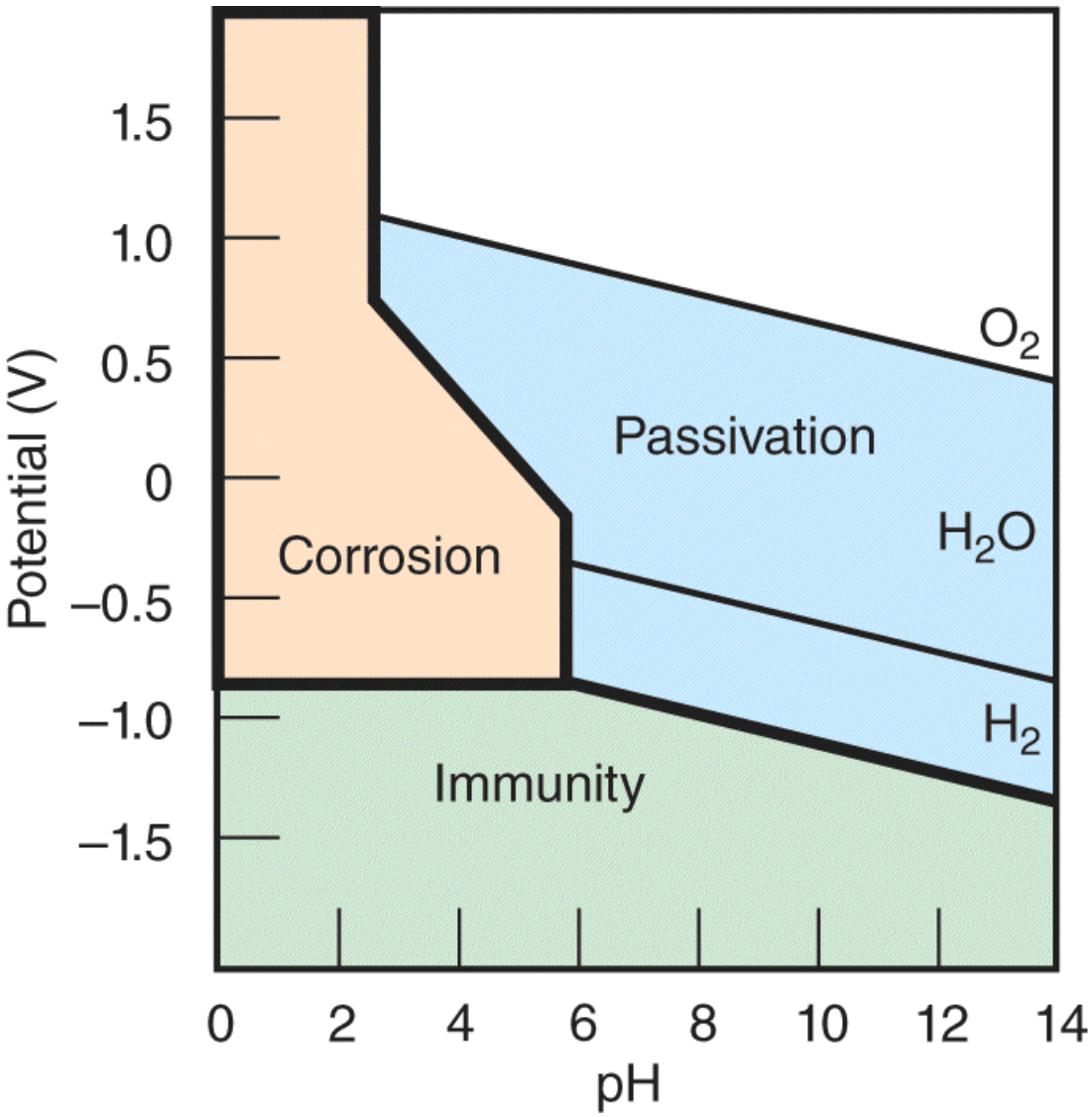


Figure 7.26 Simplified version of a Pourbaix diagram showing the range of conditions over which corrosion, passivation, and immunity are likely to occur for a metal capable of two valence states, M^{2+} and M^{3+}.

These types of representation must, however, be used with caution. The positions of the boundaries to the phase fields are concentration dependent, and published diagrams may not be constructed with concentrations that are relevant to the problems under consideration. Moreover, they are derived using equilibrium thermodynamic data and are only as accurate as the available data. This is of high quality for well-known systems such as iron–water–air, but for some systems involving radioactive materials the data is less accurate. In addition, the diagrams do not consider any kinetic or crystallographic aspects. Thermodynamically unstable reaction products may be of considerable importance in corrosion and kinetic factors such as the flow rate of solutions are not catered for. Similarly, fields displaying passivation may not be such if the nominally protective film is not compact and coherent with the metal but cracks or is porous. The Pilling–Bedworth ratio (Section 6.7.3), is an attempt to determine if a film is likely to fulfil the requirements of passivation.

Reference

Guichelaar, P.J. and Williams, M.W. (1990). A simple demonstration of corrosion cells. *J. Mater. Edu.***12**: 331.

Further Reading

For a General Introduction to Electrochemistry See:

Shriver, D.F., Atkins, P.W., and Langford, C.H. (1994). *Inorganic Chemistry*, 2nde. Oxford: Oxford University Press, Chapter 7.

Structure-property Relations and Defects in Electrode and Electrolyte Solids is Described in:

Tilley, R.J.D. (2008). *Defects in Solids*. Hoboken, NJ: Wiley especially Chapters 6 and 8.

Batteries:

Armand, M. and Tarascon, J.-M. (2008). Building better batteries. *Nature***451**: 652–657.

Dell, R.M. and Rand, D.A.J. (2001). *Understanding Batteries*. Royal Society of Chemistry.

Lai, J., Xing, Y., Chen, N. et al. (2020). Electrolytes for rechargeable lithium-air batteries. *Angew. Chem. Int. Ed.***59**: 2974.

Lopes, P.P. and Stamenkovic, V.R. (2020). Past, present and future of lead-acid batteries. *Science***369**: 923.

Whittingham, M.S. (2004). Lithium batteries and cathode materials. *Chem. Rev.***104**: 4271–4301.

Winter, M. and Brod, R.J. (2004). What are batteries, fuel cells and supercapacitors? *Chem. Rev.***104**: 4245–4269.

Wu, F., Maier, J., and Yu, Y. (2020). Guidelines and trends for next-generation rechargeable lithium and lithium-ion batteries. *Chem. Soc. Rev.***49**: 1569.

Xu, G., Shangguan, X., Dong, S. et al. (2020). Formulation of blended-lithium-salt electrolytes for lithium batteries. *Angew. Chem. Int. Ed.***59**: 3400.

Yoshino, A. (2012). The birth of the lithium-ion battery. *Angew. Chem. Int. Ed.***51**: 5798–5800.

Solid Oxide Fuel Cells:

Anson, C.W. and Stahl, S.S. (2020). Mediated fuel cells: soluble redox mediators and their applications to electrochemical reduction of O_2 and

oxidation of H_2, alcohols, biomass and complex fuels. *Chem. Rev.***120**: 3749.

Huang, L., Li, J., Liu, B. et al. (2020). Electrode design for lithium-sulphur batteries: problems and solutions. *Adv. Funct. Mater.***30**: 1910375.

Jia, X., Liu, C., Neale, Z.G. et al. (2020). Active materials for aqueous zinc ion batteries: synthesis, crystal structure, morphology and electrochemistry. *Chem. Rev.***120**: 7795.

Li, W., Wang, D., Zhang, Y. et al. (2020). Defect engineering for fuel-cell electrocatalysts. *Adv. Mater.***32**: 1907879.

Wang, P. and Buchmeiser, M.R. (2020). Rechargeable magnesium-sulphur battery technology: state of the art and key challenges. *Adv. Funct. Mater.***29**: 1905248.

Wang, C., Fu, K., Kammampata, S.P. et al. (2020). Garnet-type solid state electrolytes: materials, interfaces and batteries. *Chem. Rev.***120**: 4257.

Yang, X., Luo, J., and Sun, X. (2020). Towards high-performance solid-state Li-S batteries: from fundamental understanding to engineering design. *Chem. Soc. Rev.***49**: 2140.

Corrosion:

Roberge, P.R. (2008). *Corrosion Engineering: Principles and Practice*. McGraw-Hill.

Electroplating:

Kanami, N. (2004). *Electroplating: Basic Principles, Processes and Practice*. Oxford: Elsevier.

Sun, X., Zhang, X., Ma, Q. et al. (2020). Revisiting the electroplating process for lithium-metal anodes for lithium-metal batteries. *Angew. Chem. Int. Ed.***59**: 6665.

Problems and Exercises

Quick Quiz

7.1 Oxidation is equivalent to:

a. Electron gain.

b. Electron loss.

c. Electron transfer.

7.2 A redox reaction is one in which:

a. Oxidation *and* reduction occur.

b. Oxidation *or* reduction occurs.

c. Oxygen takes part.

7.3 In a galvanic cell the anode:

a. Is the negative terminal.

b. Is the positive terminal.

c. Links the cell compartments.

7.4 Batteries are examples of:

a. Electrolytic cells.

b. Corrosion cells.

c. Galvanic cells.

7.5 In a battery, oxidation takes place at the:

a. Cathode.

b. Anode.

c. Neither.

7.6 A hydrogen electrode can be:

a. The anode of a cell.

b. The cathode of a cell.

c. Either the cathode or the anode.

7.7 In a battery, the couple *higher* in the electrochemical series:

a. Forms the cathode.

b. Forms the anode.

c. Sometimes forms the cathode and sometimes the anode.

7.8 The cell potential is a measure of:

a. The free energy change of the cell reaction compared with the hydrogen electrode.

b. The free energy of the anode reaction.

c. The free energy of the cell reaction.

7.9 The Nernst equation describes:

a. The free energy of a galvanic cell.

b. The variation of the potential of a galvanic cell with concentration.

c. The reaction equation of a galvanic cell.

7.10 A rechargeable battery is called:

a. A primary cell.

b. A secondary cell.

c. A fuel cell.

7.11 In an alkaline cell the zinc is:

a. Oxidised and forms the anode.

b. Reduced and forms the cathode.

c. Is neutral and forms the container.

7.12 In a lithium primary cell the lithium is:

a. Oxidised and forms the anode.

b. Reduced and forms the cathode.

c. Is neutral and forms the container.

7.13 In a dual-ion battery the ionic transport in the electrolyte is described as:

a. Rocking-chair.

b. Advance and retreat.

c. Steady state.

7.14 When a metal reacts with an acid, it is:

a. Reduced.

b. Oxidised.

c. Neither oxidised or reduced, simply dissolved.

7.15 When a metal reacts with water it is:

a. Oxidised.

b. Reduced.

c. Neither oxidised or reduced, simply corroded.

7.16 During dissimilar metal corrosion, the metal that corrodes is:

a. The anode.

b. The cathode.

c. Neither.

7.17 The amount of chemical produced during electrolysis is governed by:

a. The voltage applied.

b. The concentration of the reactants.

c. The amount of electricity passed.

7.18 A Pourbaix diagram plots:

a. Oxidation potential against free energy.

b. Oxidation potential against temperature.

c. Oxidation potential against pH.

7.19 A Pourbaix diagram does NOT give information about:

a. The corrosion resistance of a metal.

b. The rate of corrosion of a metal.

c. The solubility of a metal.

Calculations and Questions

7.1 Volta's original battery (a Volta pile) was made with six silver and six zinc discs.

a. Which metal forms the anode and which the cathode?

b. Write the anode reaction, cathode reaction, and cell reaction.

c. Determine the voltage of the pile.

7.2 What is the standard reaction free energy for the cell reaction of a Daniel cell?

7.3 Estimate the value of RT/nF for monovalent, divalent, and trivalent ions at 27 °C.

7.4

a. Write the cathode and anode reactions and the overall cell reaction for a cell with Ni and Zn electrodes.

b. Determine the standard cell voltage.

c. Calculate the cell voltage if the concentrations of the ions in solution are Zn^{2+}, 0.016 mol dm^{-3}, Ni^{2+}, 0.087 mol dm^{-3}.

7.5 Determine the voltage of the cell in the previous question if it is operated at 50 °C.

7.6 A cell is constructed with a hydrogen electrode:

$$\mathrm{Pt} \mid \mathrm{H_2(g)} \mid \mathrm{H^+(aq)} \parallel \mathrm{Cu^{2+}(aq)} \mid \mathrm{Cu}$$

$$E^0 = 0.34\ \mathrm{V}$$

a. Write the anode reaction, the cathode reaction and the overall cell reaction.

b. Derive an expression for the variation of the cell voltage with pH of the acid solution and the Cu^{2+} concentration.

c. A cell constructed with a Cu^{2+} concentration of 1 mol dm^{-3} and a hydrogen pressure of 1 atm. has a voltage of 0.855 V. Estimate the pH of the acid solution.

7.7 For the cell with an overall cell reaction:

$$\mathrm{Zn(s) + Fe^{2+}(aq) \rightarrow Zn^{2+}(aq) + Fe(s)}$$

a. Write the anode and cathode reactions.

b. Determine the standard cell potential.

c. Determine the reaction free energy of the cell reaction.

7.8 What will the voltage of the cell in the previous question be:

a. If the Fe^{2+} ion concentration is changed to 0.35 mol dm^{-3}?

b. If the temperature of the cell is subsequently raised to 35 °C?

7.9 A voltammeter connected to a pH meter using a calomel electrode was calibrated with a buffer solution of pH 7.0 and showed a voltage of 0.12 V. What is the pH of a solution that gives a voltage of:

a. 0.195 V.

b. 0.48 V.

7.10 A pH meter of the type in Question 7.9 is made up with a $Cl^-(aq)$ concentration of 0.5 mol dm^{-3}. The cell voltage is 0.48 V. Determine the pH of the $H^+(aq)$ component.

7.11 Three clean steel nails are treated in the following ways:

a. Completely immersed in tap water.

b. Completely immersed in boiled distilled water.

c. Partly immersed in tap water.

d. The results found after several days are:

e. Rust present, especially on the head and point;

f. Little rust present;

g. Dense ring of rust at the waterline and pitting just below the surface of the water. Explain these findings. (see Guichelaar and Williams 1990.)

7.12 A titanium container is to be used to store nuclear waste in mine conditions that are damp and broadly reducing so that corrosion will tend to produce $Ti^{2+}(aq)$. If the container lid is fixed with steel bolts, should these be plated, and if so, what metals might be considered?

7.13 Buried steel pipes can be protected from corrosion by connecting them to blocks of metal such as magnesium, called sacrificial anodes, which corrode in preference to the pipe. A cell made from the redox couples Fe/Fe^{2+} and Mg/Mg^{2+} is a laboratory representation of a sacrificial anode.

a. Draw the cell diagram.

b. Write the cell half-reactions, the cell reaction/

c. Calculate the standard cell voltage.

7.14 Water coming from a waste tip contains (a), Cu^{2+}; (b), Zn^{2+}; (c), Pb^{2+} and (d), Sn^{2+} ions in solution. What will happen when these come into contact with a steel pipe? Write the half-reactions and the overall reactions expected.

7.15 (a) Write an equation for the electrolysis of molten $MgCl_2$ to produce Mg metal and Cl_2 gas. (b) What mass of each of these elements is produced if a current of 30 A is passed for three hours through the cell?

7.16 The simplified equation for the production of aluminium by the Hall–Héroult process is:

$$2Al_2O_3(l) + 3C(s) \rightarrow 4Al(l) + 3CO_2(g)$$

a. What quantity of electricity must be passed to produce 2 M of Al?

b. 2 kg of Al.

c. If a current of 25 A is used, how long will it take to produce these amounts?

7.17 A solution of a metal sulphate, MSO_4, is used to electroplate an object. The plating bath is operated for two hours at a current of 0.5 A. If 1.095 g of metal plates out, determine the molar mass of the metal and identify it.

7.18 On a Pourbaix diagram:

a. What general types of equilibria apply to horizontal lines?

b. What general types of equilibria apply to vertical lines?

c. What general types of equilibria apply to sloping lines?

7.19 The following figure shows the Pourbaix diagram for the Al–H_2O system within the water stability field, for an Al^{3+} concentration of 1×10^{-6} mol dm^{-3}.

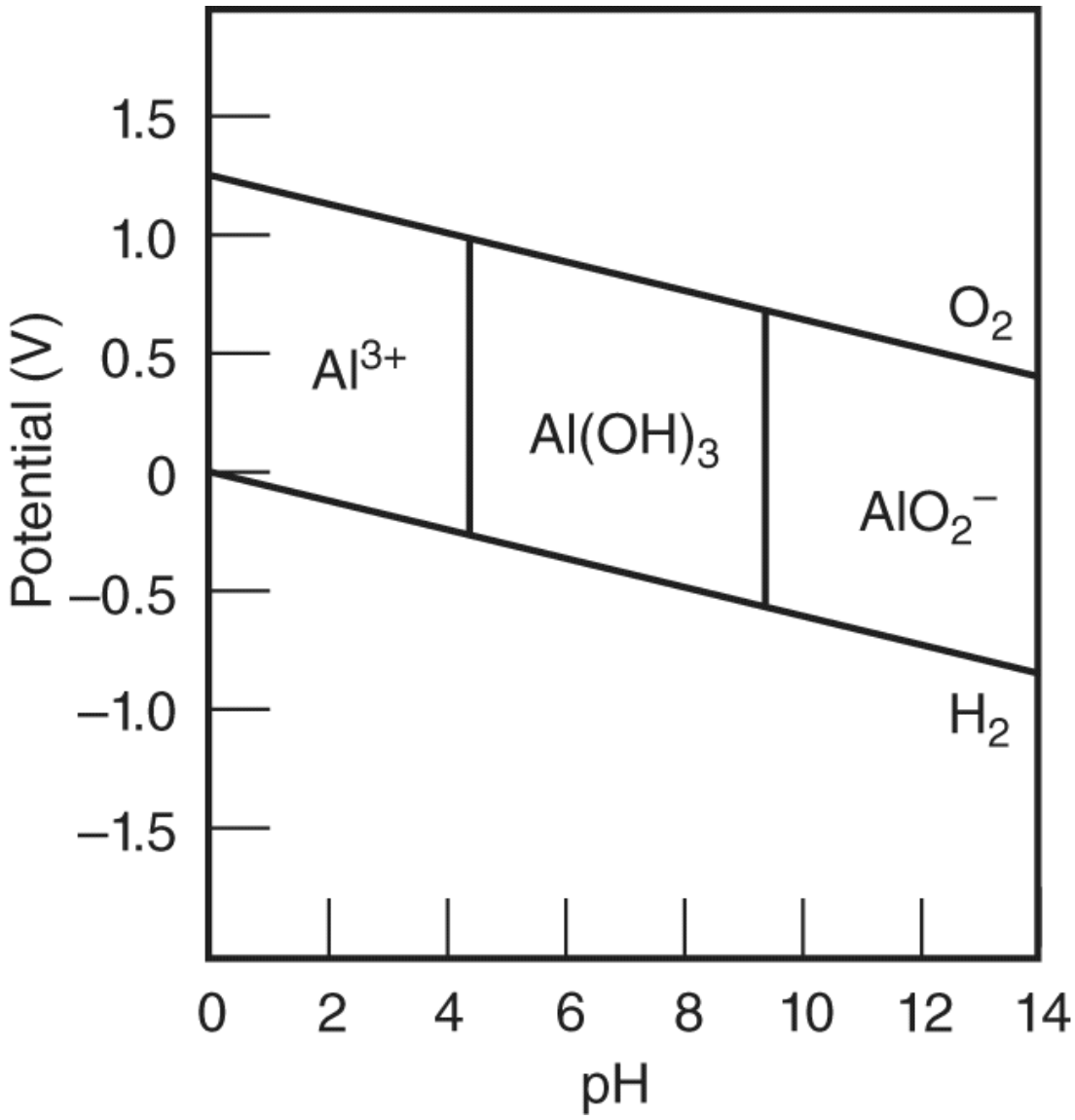
a. What oxidation states of Al exist?

b. Why are there no horizontal boundaries on this diagram?

c. Will Al be soluble in water with an acidity of pH = 3?

d. What species will exist in the surface waters of lakes and streams (pH ~ 7, E ~ 0.6 V)?

e. Over what pH range is Al passivated?



7.20 The following figure shows the Pourbaix diagram for the Cu–H_2O system within the water stability field, for a Cu^{2+} concentration of 1×10^{-6} mol dm^{-3}.

a. What oxidation states of Cu exist?

b. Under what conditions will copper be soluble in water of pH = 3?

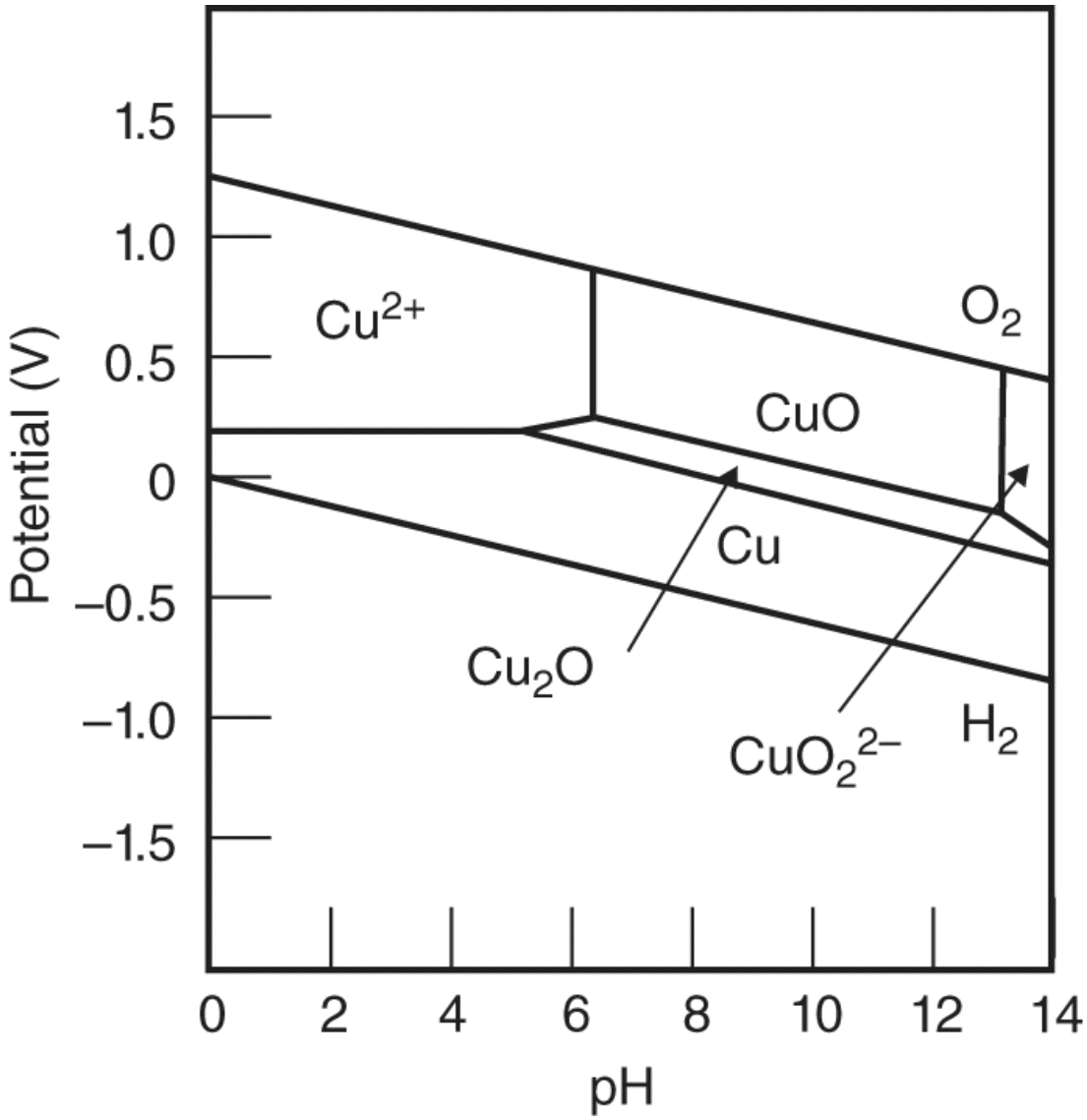
c. Label the diagram to show the regions where copper corrodes, is passivated, and is immune to corrosion.

d. What equilibria do the two vertical lines on the diagram represent?

e. What equilibrium does the horizontal line represent?

f. What equilibrium does the sloping line between the copper and Cu_2O stability fields represent?

g. What equilibrium does the sloping line between the Cu_2O and CuO stability fields represent?



7.21 The water types met with in nature are:

a. Fresh surface waters, pH ~ 7, E ~ 0.6;

b. Organic-rich fresh water, pH ~ 5, E ~ 0;

c. Organic-rich waterlogged soil, pH ~ 4.5, E ~ −0.1 V;

d. Acid bog water, pH ~ 3, E ~ 0.1 V;

e. Fresh water polluted by mine drainage, pH ~ 3, E ~ 0.8 V;

f. Surface ocean water, pH ~ 8, E ~ 0.55 V;

g. Organic-rich ocean water, pH ~ 9, E ~ −0.4.

h. Sketch these onto the stability field of water on the Pourbaix diagram in the previous question. What copper species will be present in each of these water types?

7.22 Silver objects acquire a surface film that dulls the appearance and can lead to a layer of a black film over the surface. One method to remove this is to place the object in contact with a sheet of aluminium foil and immersed in a warm solution of baking soda (sodium bicarbonate). If a voltmeter is connected between the aluminium foil and the silver, will a voltage be registered, and if so, what will this be? Suggest a mechanism for the removal of the surface film.

Notes

1 The recommended standard state pressure is 10^5 Pa, but all tabulated data, including that listed here, refer to a standard state of 1 atm. ($1.013\ 25 \times 10^5$ Pa).

2 The pH of a liquid is defined as pH = $-\log [H_3O^+] = -\log [H^+]$, where $[H_3O^+]$ or $[H^+]$ is equivalent to the concentration of hydrogen ions present.

3 $pH = -\log_{10}[H^+]$.

4 $\log[OH^-] = pH - 14$.

Part III
Physical Properties

8
Mechanical Properties of Solids

- Why are metals ductile and ceramics brittle?
- What is the elastic modulus of a solid?
- What are solid lubricants?

The mechanical properties of materials are controlled by three interacting features, the strength of the chemical bonds in the material, the (crystal) structures of the solids, and the defects present. Materials reveal their mechanical properties when subjected to a force, resulting in a deformation. The amount of deformation will depend upon the magnitude of the force and its direction. In the descriptions in the following text, it will often be assumed that all materials are continuous (not made of atoms) and isotropic, so that forces can be treated as scalars (simple numbers) rather than vectors. However, crystal structure and bonding cannot be ignored when the mechanical properties of single crystals are considered, when the role of defects is described, or when nanoscale properties are investigated.

8.1 Strength and Hardness

8.1.1 Strength

Everyone has a subjective idea of strength, and some materials, such as steel are universally regarded as strong, while others, like plastic films, are considered weak. However, the strength of a material will depend exactly upon how it is evaluated. The strength of a solid may be quantified by measuring the amount of force applied to it and the corresponding deformation that occurs. A material that is stretched is in *tension*, and suffers *tensile forces*. The tensile strength of metals and polymer fibres is usually high. A material that is squeezed is in *compression*, and suffers *compressive forces*. Compressive strength needs to be high in building materials that have to bear heavy loads. A material that is subjected to opposed forces is said to be *sheared*, and suffers *shear forces*. Many polymers behave like very viscous liquids and have very low shear strength. A material that is *twisted* is subjected to a *torsional load*. Torsional strength is important for shafts that transmit rotation. A solid that is *flexed* (bent) is subject to *both* tension and compression, and the *flexural strength* gives a measure of the amount of bending that an object can sustain without fracture. Finally, the *impact strength* of a solid measures the resistance to a sudden blow. Glass has low impact strength, while wooden bats have high impact strength (Figure 8.1). A component may experience all of these forces and resultant deformations at the same time, but here each will be treated separately.

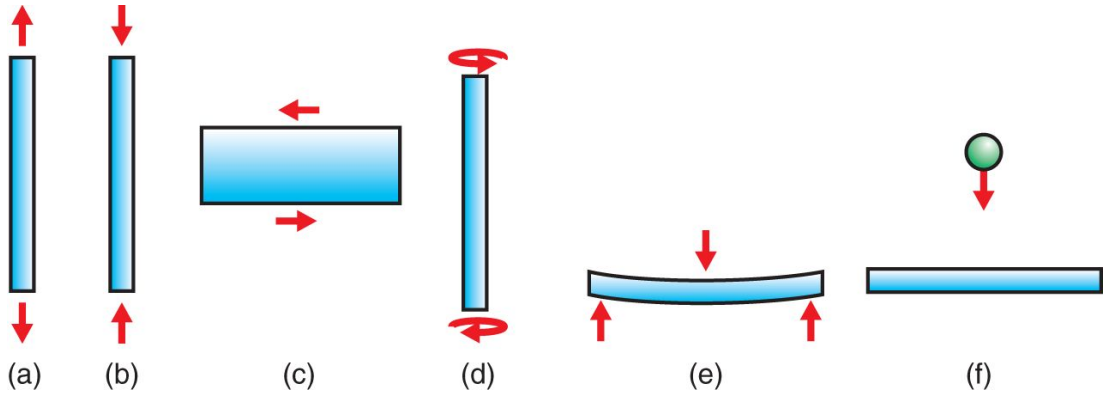


Figure 8.1 Mechanical loading of solids: (a) tension; (b) compression; (c) shear; (d) torsion; (e) flexing; (f) impact.

The interdependence between the forces applied and the deformations that are produced are summarised by a number of *elastic moduli*, of which *Young's modulus*, also called the *elastic modulus*, is the best known (Section 8.2).

8.1.2 Stress and Strain

The *force* (often called the *load*) applied to an object is defined in terms of the *stress* on the object. Stress is measured as the force applied to a unit area of the specimen. The application of a stress results in a dimensional change, which is called the *strain*. For a rod-shaped specimen (Figure 8.2), often used for the evaluation of metal or polymer samples:

$$\sigma = F/A$$

where σ is the (tensile) stress, F is the average force applied to the rod, and A is the cross-sectional area subjected to the force. Stress is measured in pascals, Pa ($N\,m^{-2}$), commonly cited as MPa = 10^6 $N\,m^{-2}$ or GPa = 10^9 $N\,m^{-2}$. For practical purposes, it is often adequate to ignore the continuous change in cross-sectional area that occurs when a force is applied. The stress so defined is called the *engineering* (or *nominal*) stress.

$$\sigma\ (\text{engineering stress}) = F/A_0$$

where A_0 is the initial cross-sectional area of the sample. The elongation of the rod when subjected to a force is the strain. The increment in strain experienced, $\Delta\varepsilon$, when a rod is extended, is defined as the ratio of the increase in length, Δl, to the total length, l:

$$\Delta\varepsilon = \Delta l/l$$

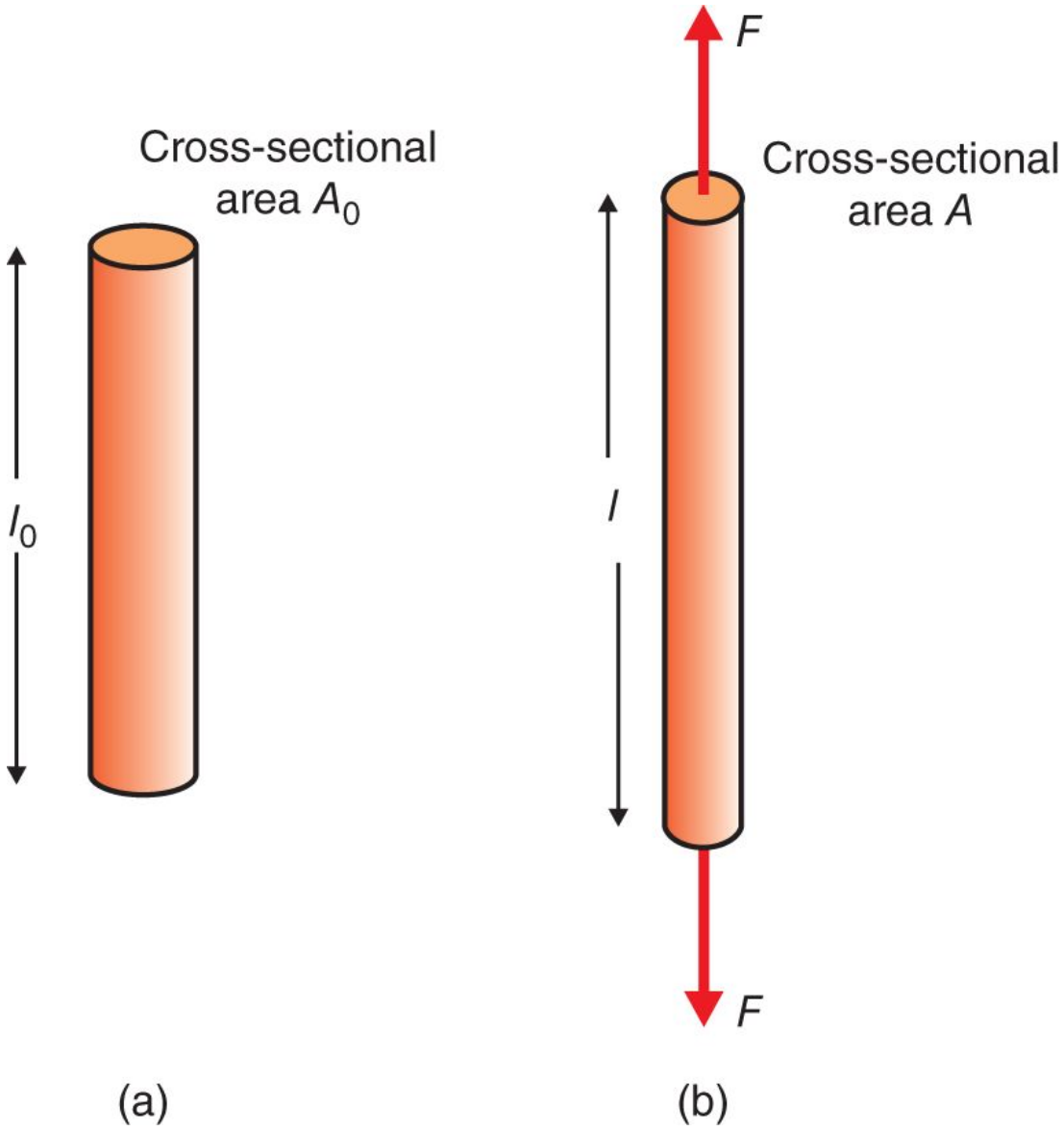


Figure 8.2 A rod in tension: (a) initial state; (b) final state.

The total strain is then given by:

$$\varepsilon = \int_{l_0}^{l} \frac{dl}{l} = \ln [l/l_0]$$

where l is the final length of the specimen and l_0 is the original length. As strain is a ratio, it has no units. If the incremental changes are ignored, the *engineering* (or *nominal*) *strain* is:

$$\varepsilon \text{ (engineering strain)} = (l - l_0)/l_0$$

In the stress–strain diagrams for metals and polymers that follow, engineering stress and engineering strain are plotted.

The stress and strain relationships for a ceramic specimen are more often determined by bending a bar, plate, or cylinder of material. In this procedure, called *a three-point bend test*, the lower part of the ceramic is under tension, and the upper surface is under compression. As ceramic materials are generally much stronger in compression, failure is initiated on the surface under tension. The maximum stress in the upper surface of a deformed sample, σ_m, is given by:

$$\sigma_m = 3\,F\,l/2\,a\,b^2$$

where F is the applied force and l is the arc length of the deformed sample between the supports of a bar with width a normal to the force and thickness b parallel to the force (Figure 8.3). The strain, ε_m, is related to the maximum deflection, δ, at the centre of the bar, is given by:

$$\varepsilon_m = 6\,\delta\,b/l$$

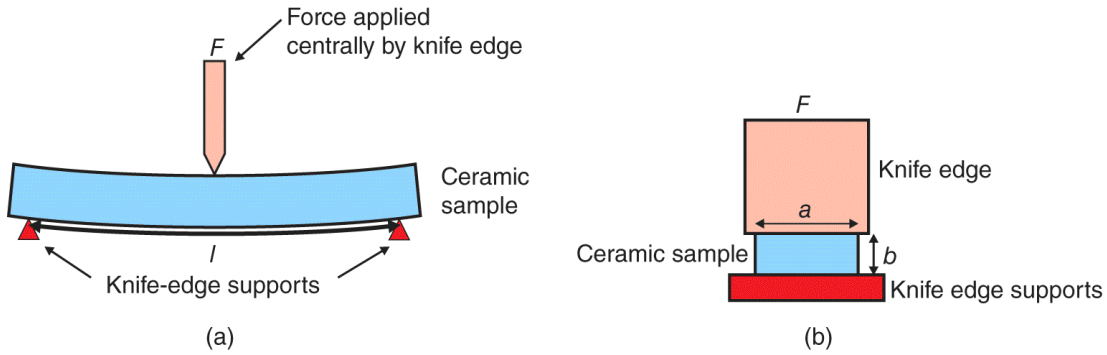


Figure 8.3 Three-point bend testy for ceramic samples: (a) sideways view; (b) end-on view.

The cross-sectional area of the ceramic specimen is not greatly altered during the test, so that the true stress is measured. In testing a ceramic, the force or load is slowly increased until fracture.

Many materials are used under compression rather than tension. At low loads, the compressed material behaves in a similar way to materials tested under tension. In compression tests, the value of the force is taken as negative and hence we have negative values of stress and strain, compared with those obtained in tension.

The most common mechanical test involving metals or polymers is the tensile test, in which a sample of the solid is stretched. The test uses a standard test piece with a shape dependent upon the material to be tested. Metals are

usually cylindrical with a central section, of known *gauge length*, usually a 50 mm, on which the measurements are made. Polymer samples tend to be sections cut from plates and are larger in dimension. During the test, the instrument applies a force to the sample at a constant rate and simultaneously records the change in dimensions of the test piece.

A plot of load against extension, stress against strain, or more commonly, engineering stress against engineering strain, gives a good picture of the mechanical behaviour of the solid in question. The behaviour of *brittle* materials such as ceramics, cast iron, or polymers chilled to well below the glass transition temperature show that the stress is usually directly proportional to the strain over all or most of the range up to fracture. Some solids, termed *slightly ductile*, behave in a similar way, but the stress/strain curve starts to bend at a point called the *yield point*, before fracture occurs (Figure 8.4a). Many metals are typically *ductile* solids, which can be drawn out into wires or permanently deformed under an applied load. These initially show a similar linear stress/strain relationship, but the plot ultimately curves and extensive deformation occurs before fracture (Figure 8.4b). The curves for most polymers are very temperature sensitive. Thermoplastic polymers above the glass transition temperature give rise to a plot that curves in the opposite way to that of a ductile metal (Figure 8.4c). Elastomers (Figure 8.4d) deform at far lower stress levels than other materials. In all of these cases, the linear part of the stress–strain curve is the *elastic region*. Here, removal of the load will allow the solid to return to its original dimensions, quite reversibly.

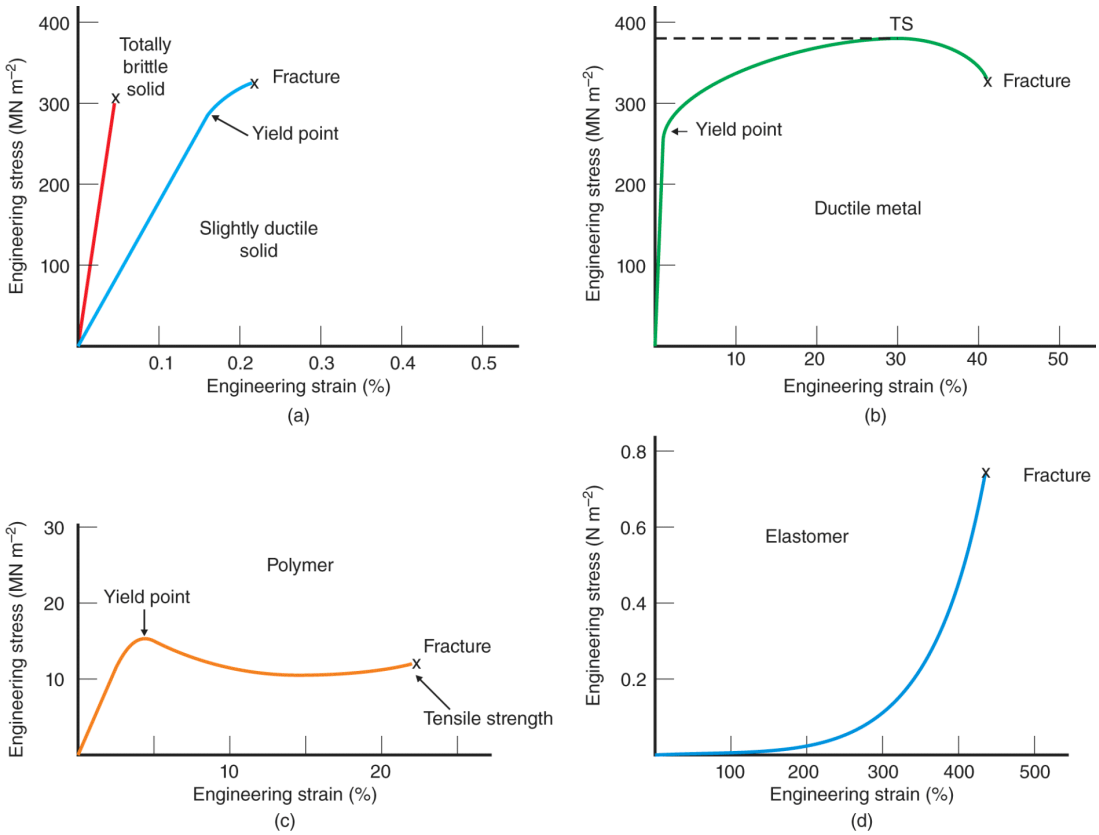


Figure 8.4 Engineering stress-strain curves, schematic: (a) brittle and slightly ductile solids; (b) ductile metals (TS = tensile strength); (c) typical polymer; (d) typical elastomer. Note the different stress scales in (c) and (d).

When the force applied to a material is relatively small and the material is subject only to elastic deformation, the stress is related to the strain by *Hooke's law*:

$$\sigma = E\,\varepsilon$$

The constant of proportionality, E, given by the slope of the stress–strain plot, is the *Young's modulus* of the material (Figure 8.5a). As weight is an important consideration in many applications, the *specific modulus* is often quoted as a parameter:

$$\text{Specific modulus} = \text{Young's modulus}/\text{specific gravity}$$

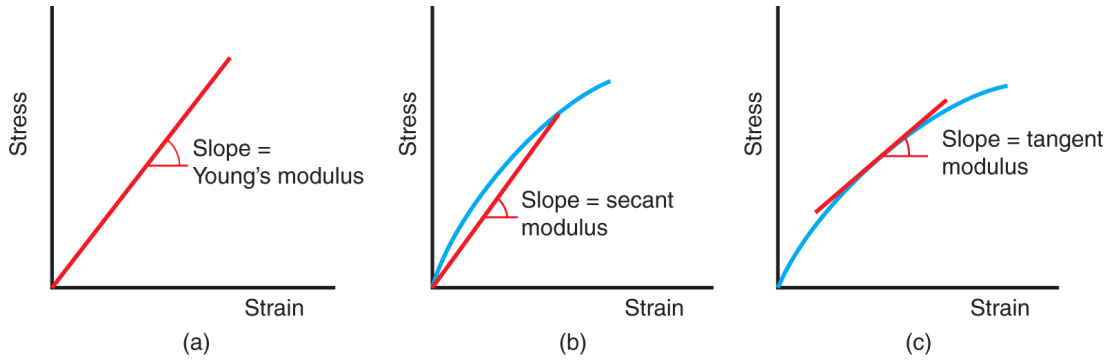


Figure 8.5 (a) Young's modulus; (b) secant modulus; (c) tangent modulus.

In many materials, especially those in which one or more components of the bonding are relatively weak, Hooke's law is not obeyed exactly. Instead, the stress–strain curve shows a distinct curve. This is termed *nonlinear* behaviour. In such cases, a single value for Young's modulus cannot be obtained. As an approximation, it is possible to measure the slope of the line drawn from the origin to intersect the stress–strain curve at a specified value of the strain, the *secant modulus*, or the tangent to the curve at any point, to obtain the *tangent modulus* (Figure 8.5b,c).

For all solids, once the elastic region is passed, the deformation of the solid is not reversed when the stress is removed, and some degree of permanent deformation remains. This is called *plastic deformation*. For metals, the point at which elastic behaviour changes to plastic behaviour is called the *yield point*, which occurs at the *yield stress*. In the case of slightly ductile materials, only a small amount of plastic deformation occurs before the material breaks into two. For a ductile metal, a large amount of deformation is possible before fracture. The maximum load that can be sustained (TS in Figure 8.4b) is called the *tensile strength* (*or ultimate tensile strength*) of the metal. The ultimate tensile strength of a material with respect to its weight is an important engineering parameter, the *specific strength*:

$$\text{Specific strength} = \text{tensile strength}/\text{specific gravity}$$

For a polymer, once the elastic region is passed, almost no increase in stress will bring about a large amount of plastic deformation. Anyone who has carried an overloaded plastic bag for any distance will have practical experience of this. The bag will support the load for a period, and then suddenly start to stretch until it breaks. Elastomers show extensive plastic deformation under any load, but this is always reversible, and so this behaviour differs from the plastic deformation found in the other materials.

The major material properties obtained from the engineering stress–engineering strain tensile curve are Young's modulus, yield strength, ultimate

tensile strength, and amount of elongation at fracture.

8.1.3 Toughness and Stiffness

Toughness is often a much more important material property than strength. Toughness can loosely be defined as the amount of energy absorbed by a material before it fractures. A tough material has a high resistance to the propagation of cracks, and so tough materials are both strong and ductile. The toughness of a solid can be represented as the area under the stress–strain curve (Figure 8.6).

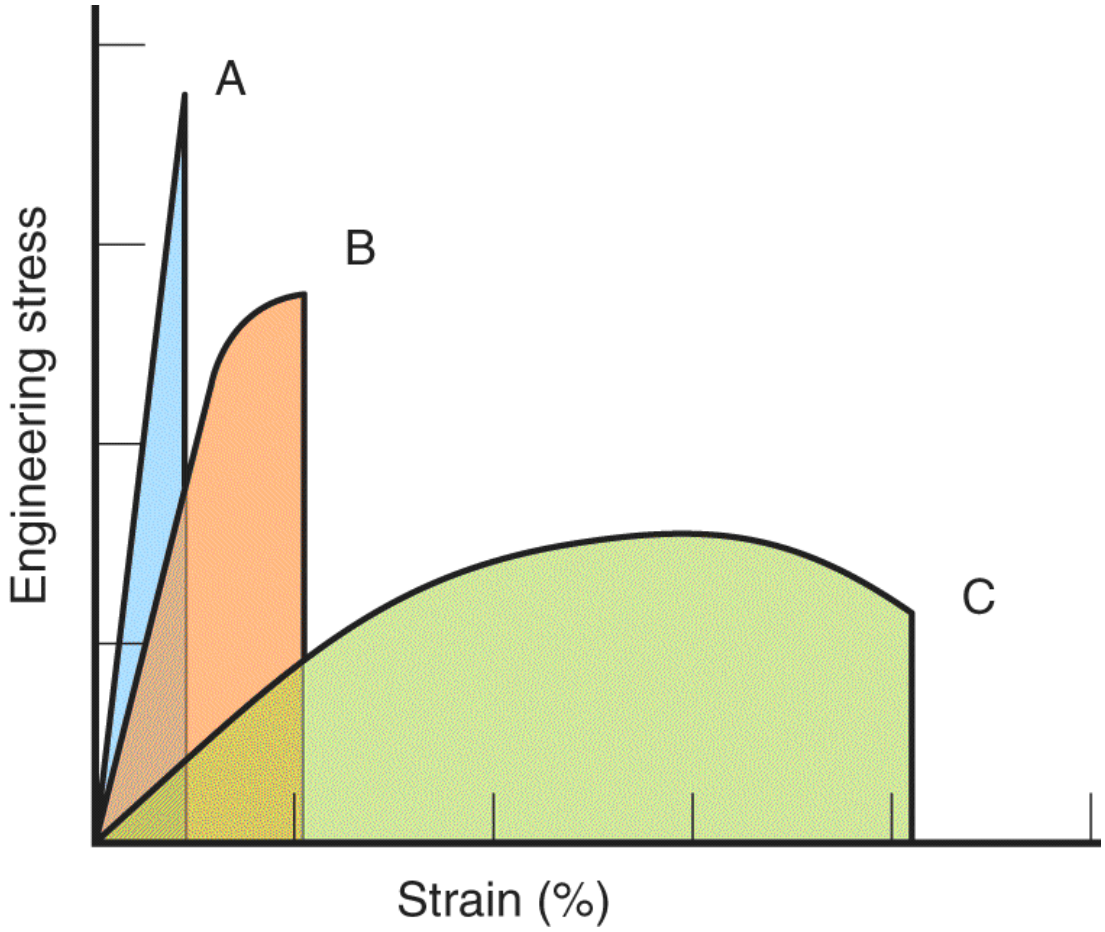


Figure 8.6 Toughness of a solid: curve C represents the toughest solid.

The ease with which a material can be extended, the *stiffness* of the solid, is represented by the slope of the stress–strain curve in the initial, ideally elastic, region. A stiff material shows only a small strain for a given stress. The stiffness is thus equivalent to Young's modulus. However, the material with the highest modulus is not necessarily the toughest. In this case, the material represented by curve C (Figure 8.6) is the toughest of the three, although it has the lowest stiffness. The toughness of a specimen will depend upon specimen geometry and the way in which the stress is applied.

8.1.4 Superelasticity

The stress–strain behaviour of many shape-memory alloys (Section 6.5.4) is different than that just described, and shows a region of extreme deformation called *superelasticity*. On release of the stress the curve usually displays a hysteresis effect (Figure 8.7). In the superelastic region, martensite is forming (even at temperatures above A_f, the austenite finish temperature), under the influence of the applied stress. For example, the martensitic transformation in the shape memory alloy Nitinol is caused by a shear of the atoms in {101} planes (Sections 6.5.3 and 6.5.4). Stress has a similar shearing effect on the structure allowing the martensitic transformation to occur above A_f. This is called a *stress-induced martensitic transformation*. The application of further stress causes more martensite to form. When the transformation is complete the material reverts to normal behaviour. However, above A_f the martensitic form is unstable in the absence of stress. Thus, as the stress is released the martensite plates revert to the parent structure, but normally the start of the reverse transformation is initiated at a lower stress than the forward transformation, resulting in hysteresis. In practice, the degree of hysteresis is found to depend upon the composition of the alloy and prior heat treatment.

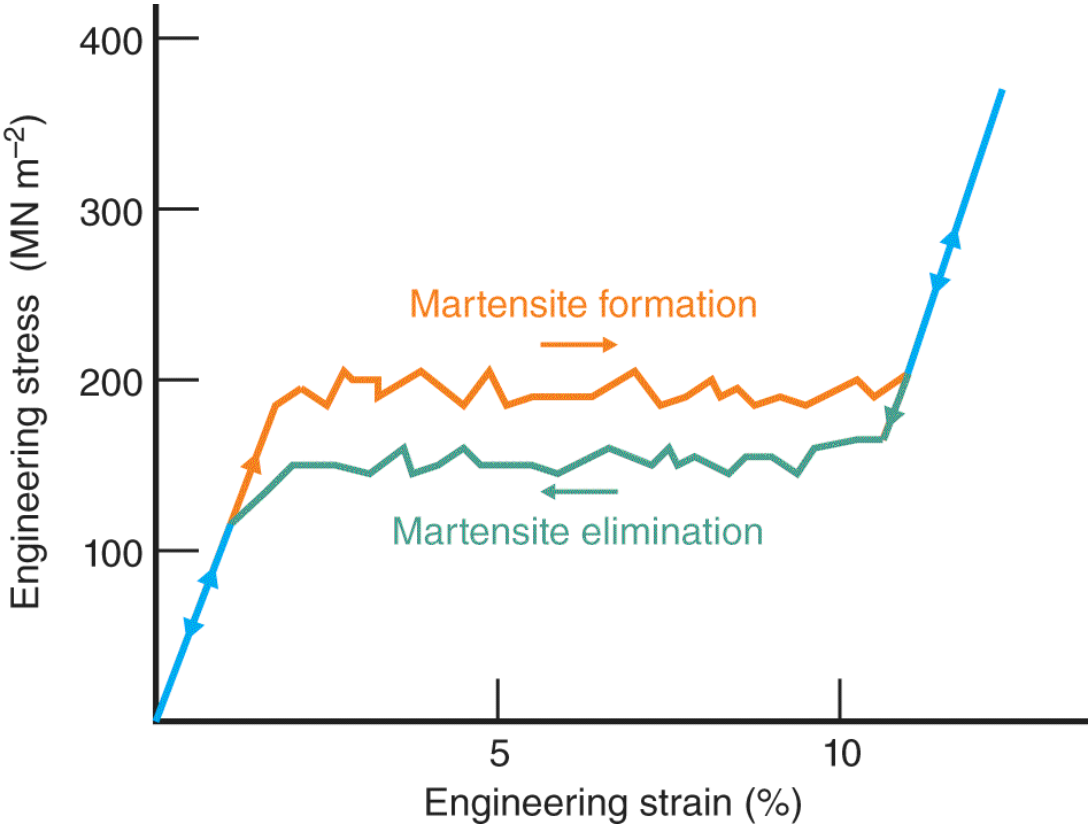


Figure 8.7 Engineering stress–strain curve, schematic, for a Nitinol shape-memory alloy showing superelasticity.

8.1.5 Hardness

Hardness is a property that is intuitively understood, but is difficult to define. It is usually taken to be a measure of the resistance of a material to permanent local deformation and is often measured by forcing a chosen solid into the surface of the material to be tested. Hardness therefore measures a compressive property rather than a tensile property.

The first use of hardness was in the characterisation of minerals. The hardness of a mineral was determined by observation of whether it could scratch or be scratched by another mineral. This is sometimes called the *scratch hardness* of a solid. The *ad hoc* system in use in medieval times was quantified by Mohs, who listed 10 minerals that, as far as possible, were equally spaced on a hardness scale of 1–10. Even today, the commonest description of hardness is still in terms of this scale, called the *Mohs scale* of hardness. The minerals chosen and their hardness are as follows: 1, talc [$Mg_3Si_4O_{10}(OH)_2$]; 2, gypsum ($CaSO_4 \cdot 2H_2O$); 3, calcite ($CaCO_3$); 4, fluorite (CaF_2); 5, apatite ($Ca_5(PO_4)_3(OH)$); 6, orthoclase ($KAlSi_3O_8$); 7, quartz (SiO_2); 8, topaz ($Al_2SiO_4(OH)_2$); 9, corundum (Al_2O_3), and 10, diamond (C). In addition, Mohs suggested the following values: fingernail, less than 2; copper coin, 3; pocket-knife blade, just over 5; window glass, 5.5; and a steel file, 6.5. Each of these materials will scratch the surface of the solid immediately below it in the list. The softest material is talc, which will not scratch any of the others. Diamond is the hardest mineral known.

To measure hardness more precisely, a known load is applied slowly to a hard indenter that is placed onto the smooth surface to be tested and allowing it to remain in position for a standard time before being withdrawn. The resulting indentation size after the indenter is removed is taken as the measure of the *indentation hardness* of the material. Hardness is recorded as a series of internally consistent empirical *hardness numbers*, related to the size of the indentation.

There are five major indentation hardness tests, which differ from each other in the shape of the indenter (Figure 8.8a–e). The first of these, described in 1900, was the Brinell test, using a 10 mm steel ball indenter, giving the *Brinell hardness number* (BHN). This was suitable only for metals softer than steel. In 1920 Rockwell developed a number of tests, including the A, C, and D scales, using a conical diamond indenter with a spherical tip, and the B, E, F, and G scales, in which the indenter is steel. In the Rockwell test the difference in size between the indentations caused by a small and a large load are compared to give the *Rockwell hardness number* (RHN). The Vickers test, introduced in 1924, uses a pyramidal diamond indenter and a standard load time of 15 seconds, to give the *Vickers hardness number*, VHN or HV. The Knoop test, widely used for brittle materials such as minerals and glass, was introduced in

1939. This method uses an elongated pyramidal diamond indenter and gives the *Knoop hardness number*, KHN or HK. Both the Vickers and Knoop tests can use small loads, of the order of grams, and measure what is generally called *micro-hardness*.

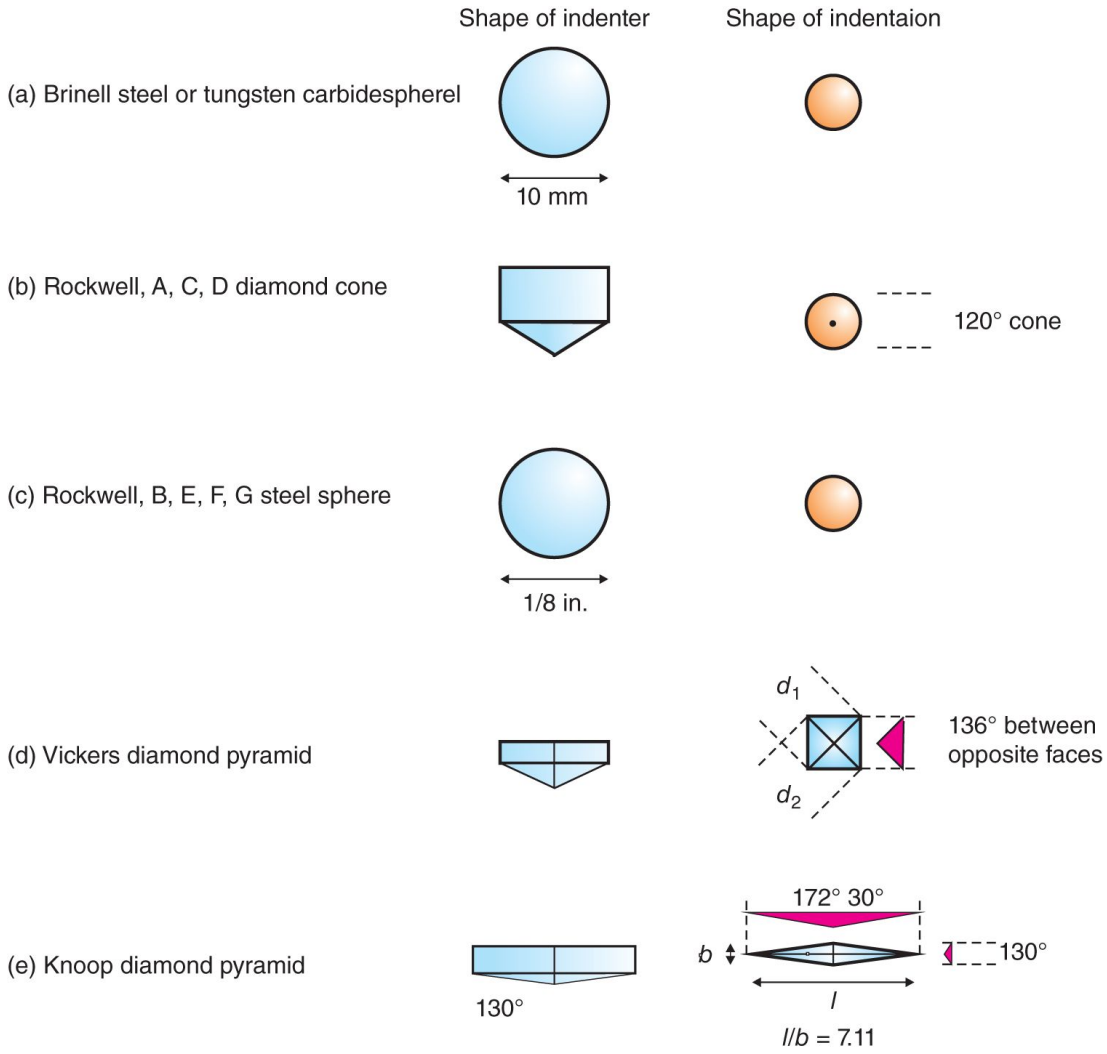


Figure 8.8 Hardness indenters.

Vickers and Knoop hardness numbers are obtained by dividing the load applied to the indenter divided by the area over which the force so generated acts (not the simple cross-sectional area of the indentation). Vickers hardness number is given by:

$$\text{VHN (kg mm}^2) = 2P \sin(136/2)/d^2$$
$$= 1.854\,368\,P/d^2$$

where P is the load d is the average of the indent diagonals d_1 and d_2 (Figure 8.8d). The hardness number is then quoted as, say, 560 HV 3, where the first number specifies the hardness number (kg mm^{-2}), HV is the hardness scale and the last number (3) specifies the load (kg). In addition, Vickers hardness numbers can be converted to Mohs scale with the approximate relation:

$$\text{Mohs hardness} \approx 0.675\,(\text{VHN, kg m}^{-2})^{1/3}$$

KHNs are obtained in a similar way to VHN:

$$\text{KHN (kg mm}^2) = P/c\,l^2$$

where P is the load, l is the long diagonal of the indentation (mm) (Figure 8.8e), and c is a correction factor, ideally 0.070 279. The hardness is quoted as, say, 530 HK 0.5, where the first number specifies the hardness number (kg mm^{-2}), HK is the hardness scale, and the last number (0.5) specifies the load (kg).

Vickers and Knoop hardness can readily be converted into conventional SI units. The applied load (kg) is converted to newtons by multiplying by the acceleration due to gravity, 9.806 65 m s^{-2}, and the dimensions (mm) are converted to metres by dividing by 1000. Multiplying the hardness number (kg mm^{-2}) by 9.806 65 then converts the value to MPa.

The hardness of a material is related to the strength of the chemical bonds within the solid. As with other mechanical properties, however, the direct correlation is masked by the defect structure of the solid. The deformation that acts as a measure of hardness is produced by compression and subsequent deformation. The compression is a temporary elastic displacement operating while the load is imposed. Deformation is due to dislocation slip, initiated by shear stress. It has recently been found there is a good correlation between shear modulus (Section 8.2.4) and hardness (Figure 8.9). This allows the hardness of new materials to be predicted from knowledge of the shear modulus.

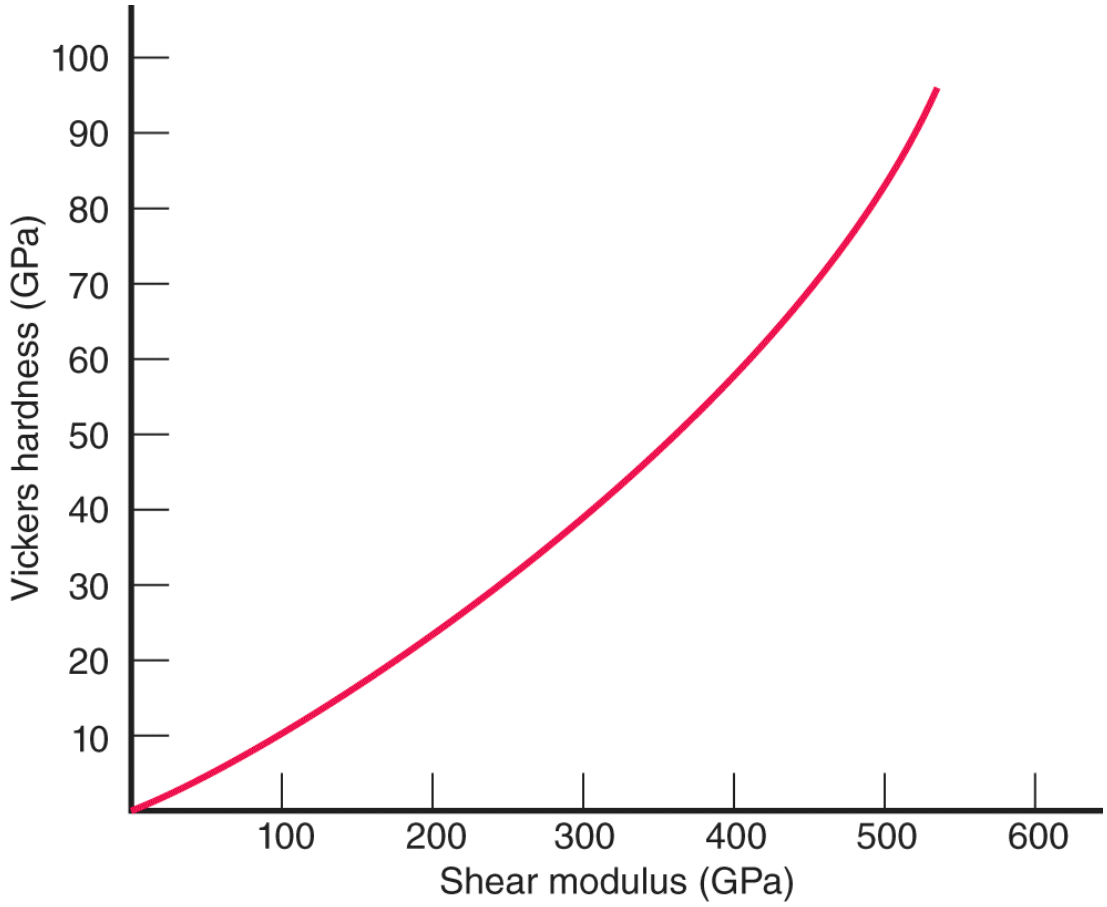


Figure 8.9 Schematic relationship between Vickers hardness and shear modulus. Source: Adapted from Teter, D.M. (1998). Computational alchemy, the search for new superhard materials. *Materials Research Society Bulletin*, 23: 22.

In recent years, efforts have been made to try to make a harder material than diamond. The hardness of diamond is attributed to the strong sp^3-hybrid bonds linking the crystal into one giant molecule, and most attempts at a synthetic alternative have focussed on *iso-electronic crystals*. These are crystals with the same average number of bonding electrons available as in diamond, four per atom. The idea is to force the atoms in the new material to bond via strong sp^3-tetrahedral hybrids. The first such material made was cubic boron nitride, BN. It is seen from the Periodic Table that boron is one place to the left of carbon, and has three valence electrons per atom, while nitrogen, one place to the right, has five. The combination of equal numbers of boron and nitrogen atoms gives an average of four valence electrons per atom. Cubic boron nitride is the second hardest material, after diamond. (A form of boron nitride resembling graphite, with a low hardness, is also known.)

Compounds intermediate in composition between diamond and cubic boron nitride, such as BC_2N, have been synthesised, as well as other iso-electronic

compounds such as B_2CO and B_5NO_2, but none, so far, has the hardness of diamond.

8.2 Elastic Moduli

The simplest description of stress and strain is in terms of isotropic linear elastic materials. Isotropic (homogeneous) materials are those that have no preferred orientation. A sample cut from an isotropic solid in any orientation will behave identically to any other sample. Glasses, some polymers, and cubic crystals are isotropic but crystals of any other symmetry are anisotropic. Many polycrystalline bodies, such as metals and ceramics, behave as isotropic solids because the random orientation of the grains suffices to statistically even out the differences between the different crystal orientations. Composites including wood are usually anisotropic, with properties in one direction different to those other directions. A *linear* elastic material is one in which the strain is proportional to the applied load, and which returns to its original form when the load is removed. Elastic deformation is *reversible*. In particular the strain should not depend upon the rate at which the load is applied or removed, or any previous history of loading. The *elastic moduli* (or *elastic constants*) described in the following sections apply to such homogeneous linear elastic materials. Elastic constants can also be specified for anisotropic materials, but take a more complicated form.

8.2.1 Young's Modulus (The Modulus of Elasticity) (*E* or *Y*)

Young's modulus defines the response of a body, the strain, to a linear stress tending to stretch or compress it (Figure 8.10a,b) defined in terms of *Hooke's law*:

$$\sigma = E\,\varepsilon$$

where σ is the stress on the body and ε is the strain. The constant of proportionality, E, is *Young's modulus*. Representative values of Young's modulus are given in Table 8.1. Young's modulus is a measure of the *stiffness* of the solid. The reciprocal of the stiffness is called the *compliance* of the solid. (Both the stiffness and the compliance are more properly defined in terms of the elastic properties of anisotropic solids, in which the stress and strain are vector quantities.)

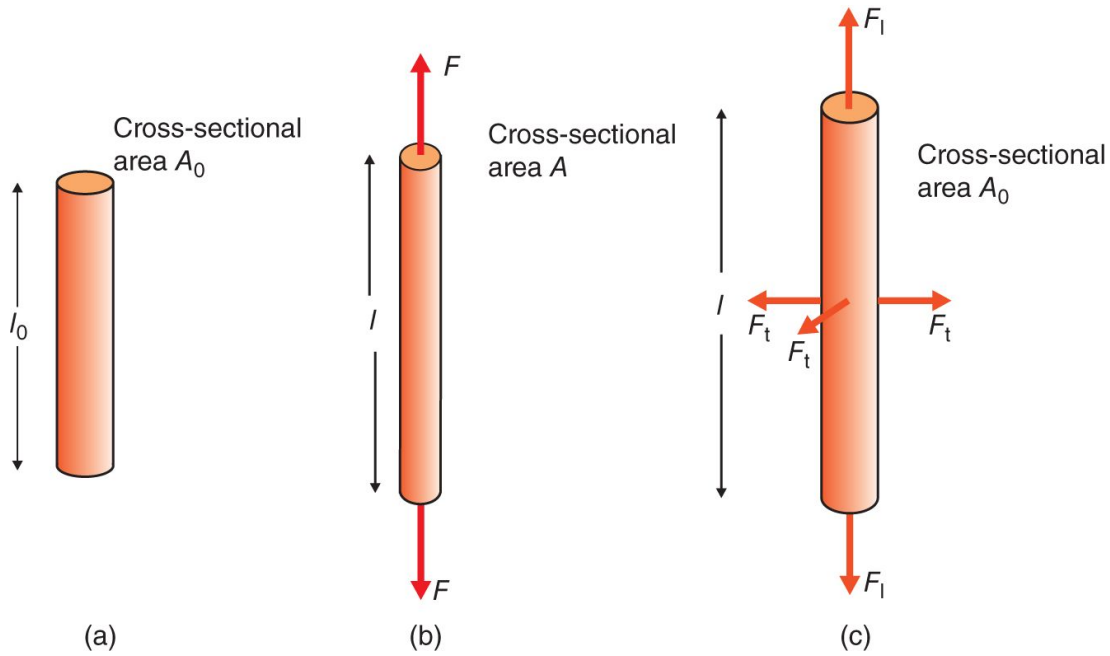


Figure 8.10 Elastic moduli: (a) initial dimensions; (b) specimen after application of a tensile force F; (c) specimen after application of longitudinal (F_l) and transverse (F_t) forces.

Table 8.1 Representative values of Young's modulus and Poisson's ratio.

Material	Young's modulus, E (GPa^{-1})	Poisson's ratio (ν)
Aluminium	70.3	0.345
Copper	129.8	0.343
Iron (cast)	~152	~0.27
Magnesium	44.7	0.291
Nickel	219.2	0.306
Titanium	115.7	0.321
Tungsten	411.0	0.280
Brass	~100	~0.35
Bronze	~105	~0.34
Steel, mild	~212	~0.29
Alumina	379.2	0.22
Magnesium oxide	210.3	0.23
Silicon carbide	468.9	0.17
Silica glass	72.4	0.17
Epoxy resin	~3.2	~0.35
Nylon 6,6	~2.0	~0.39
Polycarbonate	~2.4	~0.36
Polystyrene	~3.5	~0.33

The elastic deformation experienced is a result of pulling atoms apart or pushing atoms together, and is directly related to interatomic bonding. If the energies of the chemical bonds between the atoms can be accurately described, then the amount of deformation that will result from a given applied force can be estimated. At the heart of the calculation is an analytical expression for the potential energy of a crystal in terms of interatomic distances. To give a simple example, suppose that the lattice potential energy U of the solid is given by:

$U = -C_1/r + C_2/r^n$ where C_1, C_2, and n are empirical constants (Section 1.2.2). The first term in the equation represents the attractive energy between the atoms and the second term the repulsive energy (Figure 8.11a). The lattice energy, U_L, corresponds to the minimum in the total energy curve, reached at an inter-atomic separation of r_0. The force between the atoms is defined as F, where:

$$F = -\mathrm{d}U/\mathrm{d}r = -C_1/r^2 + nC_2/r^{n+1}$$

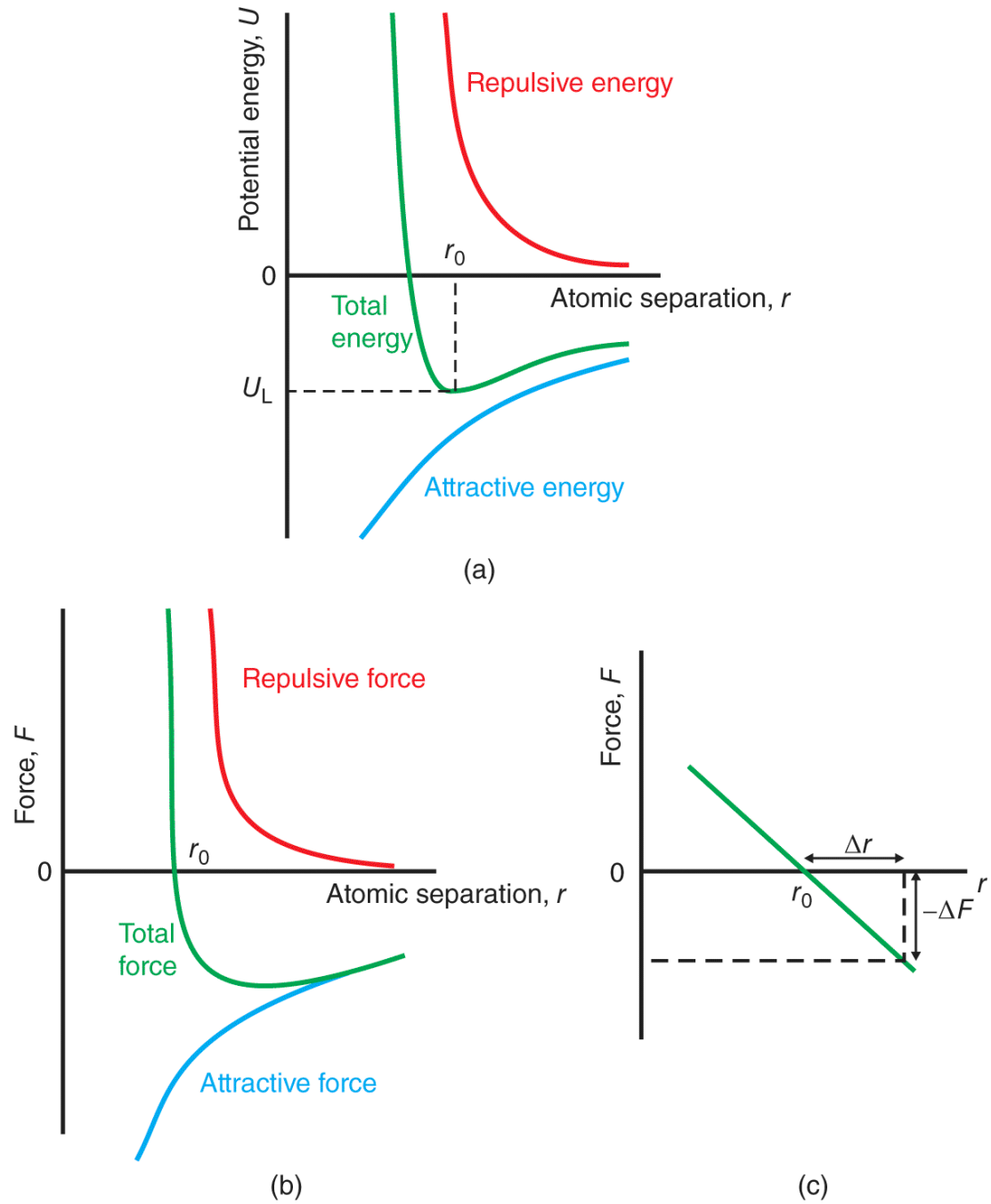


Figure 8.11 (a) The lattice potential energy U as a function of the atomic separation, r; (b) the forces between atoms, F, as a function of atomic separation, r; (c) enlarged potion of the F–r curve close to r_0.

This is also the sum of an attractive and repulsive term (Figure 8.11b). At the equilibrium separation, r_0, the force between the atoms will be zero, hence:

$$0 = -C_1/r_0^2 + nC_2/r_0^{n+1}$$

$$C_2 = C_1 r_0^{n-1}/n$$

For small values of r lying close to r_o, this curve can be approximated as a straight line (Figure 8.11c). When a crystal is subjected to an elastic strain, the force applied, ΔF, causes a displacement Δr so that the resulting stress σ and strain ε are:

$$\sigma \approx \Delta F/r_0^2; \quad \varepsilon \approx \Delta r/r_0$$

Hooke's law gives:

$$E = \sigma/\varepsilon = \Delta F/\Delta r\, r_0$$

$\Delta F/\Delta r$ is simply the slope of the curve (dF/dr) at r_o, hence:

$$(dF/dr)_{r=r_0} = 2C_1/r_0^3 - n(n+1)C_2/r_0^{n+2}$$
$$= C_1(1 - n)/r_0^3$$

Thus:

$$E = C_1(1 - n)/r_0^4$$

In terms of this simple model, Young's modulus depends sensitively upon the interatomic distance in the solid. Atomistic simulations (Section 1.2.3) using accurate potential energy functions between different atoms perform these calculations routinely to give very accurate values of elastic constants.

8.2.2 Poisson's Ratio (ν)

Although measures of engineering stress and engineering strain assume constant cross-sectional area of the rod being stressed, a material deformed elastically longitudinally (in compression or tension) has an accompanying lateral dimensional change, defined by *Poisson's ratio*, ν. If a tensile stress σ_z produces an axial strain $+\varepsilon_z$ and lateral contractions $-\varepsilon_x$ and $-\varepsilon_y$ (in isotropic materials $-\varepsilon_x = -\varepsilon_y$),

$$\nu = -\varepsilon_x/\varepsilon_z$$

The negative sign is to ensure that the numerical value of Poisson's ratio is positive (Table 8.1). For isotropic materials, the theoretical value of ν is ½.

Most metals show values in the range 0.25–0.35. Some materials, counter-intuitively, have a negative Poisson's ratio, and get thicker when under tension (Section 8.5.2).

8.2.3 The Longitudinal or Axial Modulus (*L* or *M*)

The longitudinal modulus is the linear stress required to produce an elongation in a solid without any change in the lateral dimensions of the object (Figure 8.10c). It is equivalent to a Young's modulus for zero transverse strain. That is, a linear stress, F_l must be accompanied by two perpendicular transverse stresses, F_t, to prevent any dimensional change. Thus:

$$\sigma_{\text{long}} = L\,\varepsilon$$

where the strain is defined as:

$$\varepsilon = (l - l_0)/l_0$$

This modulus determines the velocity of ultrasonic stress pulses through a solid.

8.2.4 The Shear Modulus (*G* or *μ*), Bulk Modulus (*K* or *B*), and Lamé Modulus (*λ*)

The shear modulus or modulus of rigidity defines the response of a body to a shear stress tending to distort it. The shear stress and resulting shear strain can be most easily defined for a rectangular block of material (Figure 8.12a,b). The shear stress, τ, is given by the ratio of the shear force (or load) F applied to one face of the block to the area of these faces,

$$\tau = F/A_0$$

where A_0 is the area of each of the opposed faces. When a block is subject to a shear stress, it will become deformed in the direction of the forces. The (shear) strain, γ, is defined as the tangent of the angle of deformation, ϕ:

$$\gamma\ (\text{shear})\ \text{strain} = a/b = \tan\ \varphi$$

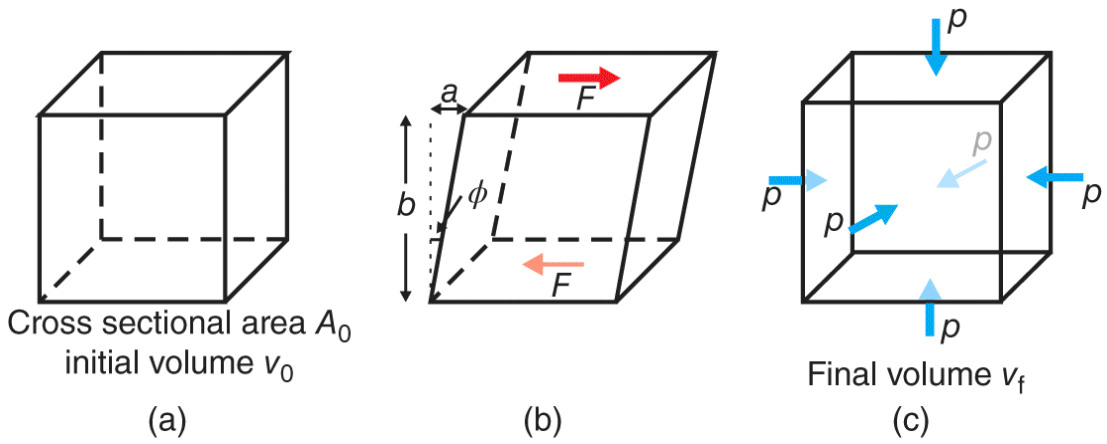


Figure 8.12 Elastic moduli: (a) initial state; (b) after application of shear forces F; (c) after application of a uniform pressure p.

The relationship between the shear strain and shear stress is analogous to Hooke's law:

$$\tau = G\gamma$$

where τ is the shear stress, G is the shear modulus, and γ is the shear strain. The shear modulus is also called *Lamé's second parameter*.

The isothermal bulk modulus, B (or K), relates the change in the volume of a solid, ΔV, to the hydrostatic strain, when subjected to a uniform pressure or hydrostatic stress, p (Figure 8.12c):

$$p = B\,(v_f - v_0)/v_0 = B\,\Delta V$$

where p is the force applied/area of face, v_f is the final volume, and v_0 the initial volume. The hydrostatic strain can be positive or negative, depending upon whether the pressure increases or decreases the volume of the solid. Atomistic simulations (Section 1.2.3) using accurate potential energy functions between different atoms are routinely employed to give values of the bulk modulus. The isothermal compressibility κ (also written as K) is the reciprocal of the bulk modulus. Beware of the possible confusion between the use of K for the bulk modulus and for the compressibility.

The Lamé modulus (λ) is also called the *incompressibility* or *Lamé's first parameter*. The Lamé modulus and the shear modulus are important in the description of waves in a solid and are utilised in seismology for the understanding of waves generated by earthquakes or underground explosions (Section 8.2.6).

8.2.5 Relationships Between the Elastic Moduli

Homogeneous isotropic elastic materials have their elastic properties uniquely determined by any two of the elastic moduli above so if given any two values the others can be calculated. Some of the equivalents are:

$$E = 2G(1 + \nu) = 3B(1 - 2\nu)$$

$$G = E/2(1 + \nu)$$

$$B = E/3(1 - 2\nu) = EG/3(3G - E)$$

$$L = B + (4G/3)$$

$$\lambda = B - 2G/3 = E\nu/(1 + \nu)(1 - 2\nu)$$

8.2.6 Ultrasonic Waves in Elastic Solids

The velocity of an ultrasonic wave travelling through an elastic solid is closely related to the elastic moduli of the solid. There are two wave types to consider, *longitudinal waves* and *transverse waves*. Longitudinal waves are those in which the atom planes making up the solid are displaced in a direction parallel to the direction of propagation of the wave (Figure 8.13a). These waves alternately compress and extend the spacing between the atom planes rather like the compression and extension of the bellows on an accordion or concertina. They are also called *pressure waves*, *compressional waves*, and in the field of seismology, *primary waves* or *P waves*. Transverse waves are those in which the atom planes making up the solid move in a direction perpendicular to the direction of propagation of the wave (Figure 8.13b). These waves move the atoms up and down in the solid like waves travelling along a rope when it is shaken up and down. They are also called *shear waves* and in the field of seismology *secondary waves* or *S waves*. Transverse waves propagate more slowly than longitudinal waves and cannot travel through liquids.

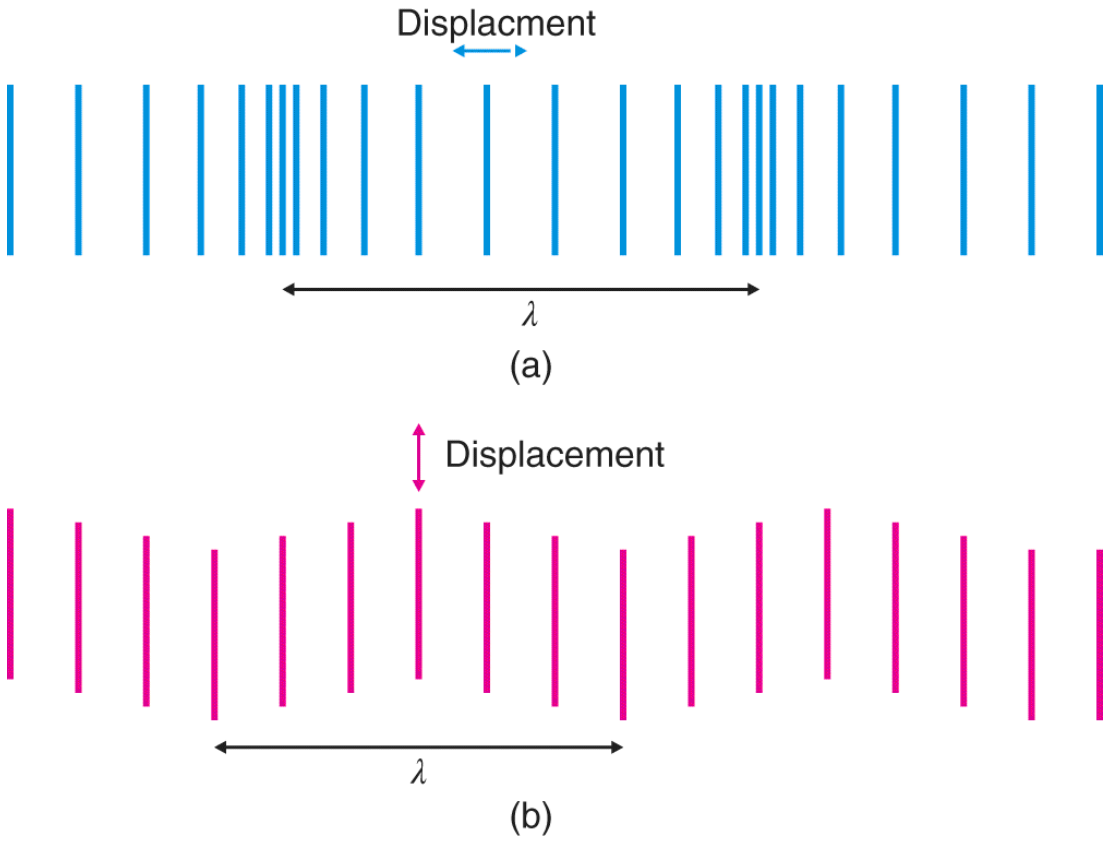


Figure 8.13 Waves in solids, schematic: (a) longitudinal waves; (b) transverse waves.

The measurement of the elastic constants of a material using ultrasound uses the pulse–echo technique employing a transducer. (A transducer is simply a device that converts electrical energy into mechanical energy and vice versa.) The heart of the transducer consists of a slice of quartz single crystal. Quartz is a piezoelectric material (Section 9.2), which means that when a voltage is applied to opposite faces of the crystal slice it will change its shape. The reverse is also true, so that when the crystal slice is forced to change shape a voltage develops across the opposite faces of the slice. To measure the elastic constants of a solid, a transducer is cemented to the sample surface and a short voltage pulse is applied to it (Figure 8.14a). The crystal can respond in one of two ways, depending upon how the slice was cut, either by a change in thickness of the crystal slice perpendicular to the sample surface, which introduces a longitudinal wave pulse into the sample, or else by shearing parallel to the sample surface, which introduces a transverse wave pulse into the sample. The pulses travel through the sample, bouncing from the sample faces to produce a series of echoes. Each time an echo reaches the transducer, it records a voltage spike (Figure 8.14b). The round time taken by the pulse gives the velocity of the ultrasound v in the material.

$$v = 2d/\Delta t$$

where d is the sample thickness and Δt the delay time between the pulse echoes.

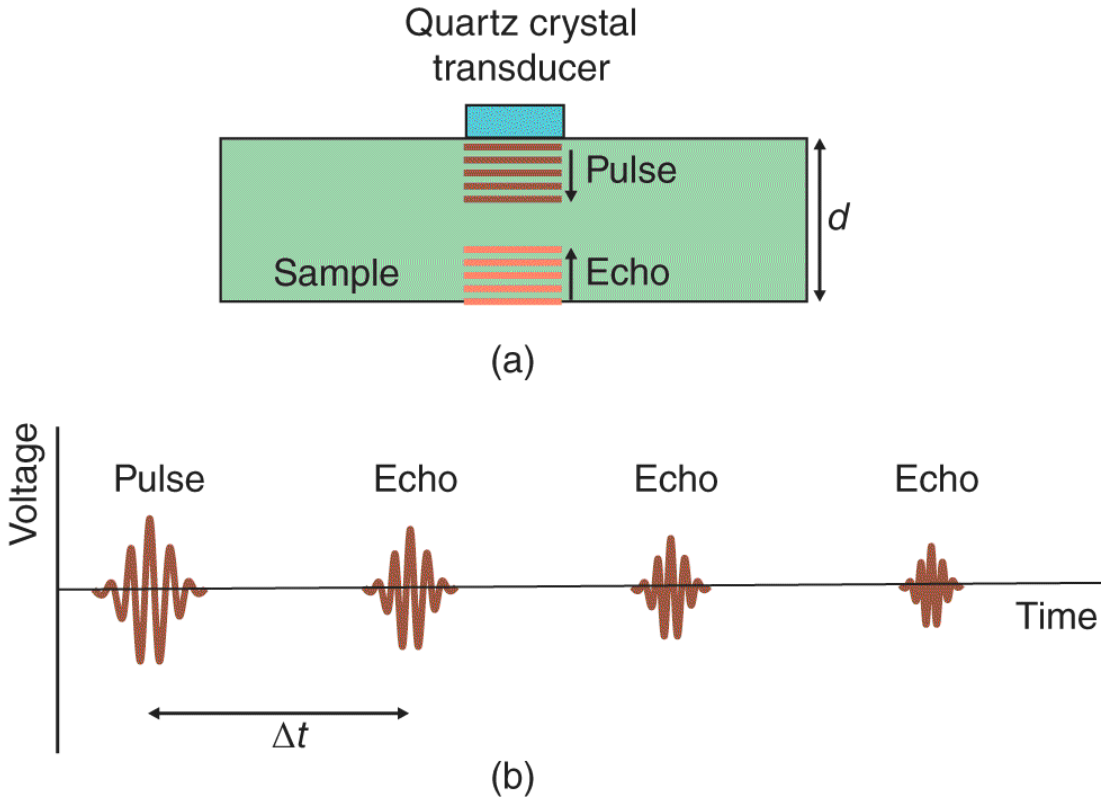


Figure 8.14 The pulse-echo technique for the determination of elastic moduli: (a) schematic of pulse and echo generation; (b) voltage as a function of time.

The pulse velocity of longitudinal waves v_L is given by:

$$v_L = \sqrt{[E\,(1-\nu)/\rho\,(1+\nu)(1-2\nu)]}$$

where the symbols have the same meanings as in Sections 8.2.1, 8.2.2, 8.2.3, 8.2.4, and 8.2.5 and ρ is the sample density. The formula can also be written in terms of other elastic constants, for example:

$$v_L = \sqrt{[(\lambda + 2G)/\rho]} = \sqrt{[B + 4G/3)/\rho]}$$

The pulse velocity of transverse waves v_T is given by:

$$v_T = \sqrt{[E/2\,\rho\,(1+\nu)]} = \sqrt{[G/\rho]}$$

and

$$v_L/v_T = \sqrt{[2\,(1-\nu)/(1-2\nu)]}$$

Taking a value of ν to be 0.3, it is found that v_L/v_T is equal to $\sqrt{3.5}$.

The elastic constants can be written in terms of the v_L and v_T velocities thus:

$$E = 2\,\rho\,v_T\,(1+\nu)$$

$$G = \rho\,v_T$$

$$B = \rho[{v_L}^2 - (4{v_T}^2)/3]$$

$$\begin{aligned}\nu &= \tfrac{1}{2}\,[1 - ({v_T}^2/{v_L}^2 - {v_T}^2)] \\ &= [1 - 2({v_T}^2/{v_L}^2)]/[2 - 2({v_T}^2/{v_L}^2)]\end{aligned}$$

A measurement of the velocities of both the longitudinal and transverse waves then allows all of the elastic moduli to be determined.

8.3 Deformation and Fracture

8.3.1 Brittle Fracture

Many materials are elastic and brittle, especially at lower temperatures, and have a linear stress–strain curve to fracture (Figure 8.4a). Single crystals frequently fracture by cleavage, along crystal planes in which the bonding is relatively weak (Figure 8.15a). A polycrystalline material can fracture in two ways. Fracture across the constituent crystallites is akin to crystal cleavage, and is called *transgranular* or *transcrystalline* fracture. (Single crystals can only fracture in a transgranular fashion.) In some materials, the weakest part is the region between crystallites, and so the fracture surface runs along the boundaries between the constituent crystallites. This is called *intergranular* fracture (Figure 8.15b). Amorphous materials such as glass or polymers that become brittle at low temperatures fracture to produce a smooth surface resembling the inside of a seashell. This is called *conchoidal* fracture (Figure 8.15c). Materials containing voids, such as porcelain, frequently fracture in the neighbourhood of these defects (Figure 8.15d).

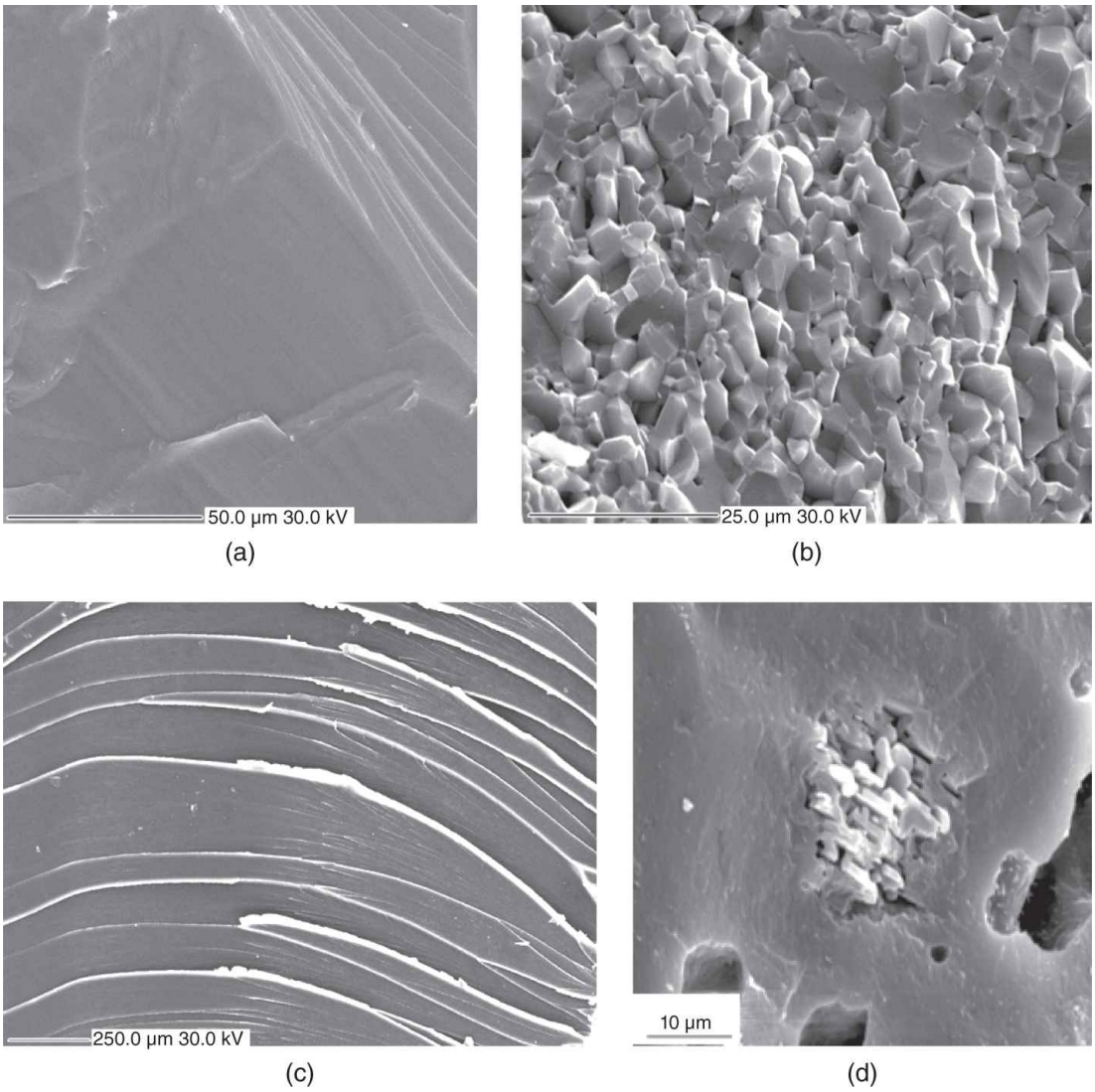


Figure 8.15 Fracture surfaces: (a) cleavage fracture of single crystal silicon; (b) transgranular fracture of a fine-grained polycrystalline alumina ceramic; (c) conchoidal brittle fracture of poly(methyl methacrylate); (d) fracture of porcelain, showing a cluster of crystallites in a glass-like matrix.

Rapid failure occurs when a crack propagates through the solid by the breaking of successive chemical bonds. It is thus logical to conclude that fracture takes place when the tensile stress is greater than the chemical bond strength. The stress on a solid as a function of the interatomic spacing can be estimated by using simple atomistic formulae for the interactions between atoms. The reasoning, published in the first quarter of the twentieth century, runs as follows. The stress is zero at the equilibrium spacing, r_o. When the atoms are stretched, as in a tensile test, the stress will increase rapidly. At a critical value of this stress, σ_c, the applied load will overcome the bond strength and if σ_c is exceeded even slightly, no more force is needed to separate the crystal into two

parts. The work done in separating the two pieces of material is equated to the energy required to form the two new surfaces. This analysis leads to the formula:

$$\sigma_c = (E\gamma/r_0)^{1/2}$$

where E is Young's modulus of the material and γ is the surface energy of the solid. The values for r_0 were just becoming available via the then new technique of X-ray crystallography, and although surface energies were poorly defined, the value of the critical stress calculated from this equation was about $E/6$ for most solids.

Measurements consistently reveal that the real strength of solids is much less than this value, lying somewhere in the region of $E/100$–$E/1000$. In order to resolve the anomaly, it was necessary to consider the distribution of stress in a solid in the region of a crack tip, which could be anticipated to be different from that measured in a tensile test. This refinement was carried out for elliptical cracks; a geometry chosen to make calculations tractable in the period before computers were available. It was found that the stress in the region of a crack tip (or, in fact, at many other objects such as voids, sharp corners or sharp grain boundaries) is much greater than the applied stress. Such flaws are called *stress raisers*. The amount that the stress is increased at a crack tip with an elliptical cross section, the *stress concentration factor*, K_t, is given by:

$$K_t = 1 + 2(a/\rho)^{1/2}$$

where a is the length of a surface crack or half the length of an internal crack and ρ is the radius of curvature of the crack tip (Figure 8.16). For a relatively long crack with a sharp tip:

$$K_t \sim 2(a/\rho)^{1/2}$$

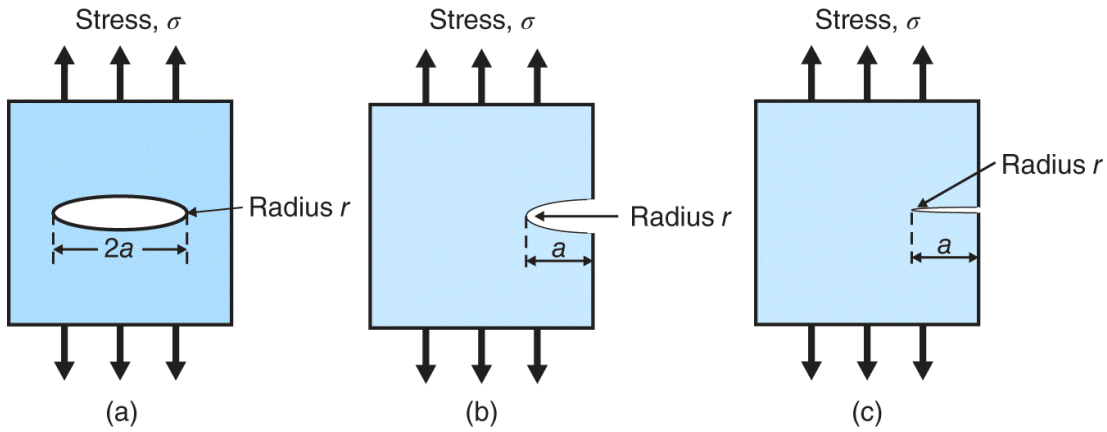


Figure 8.16 Stress in a solid at: (a) an elliptical pore; (b) an elliptical groove; (c) a sharp crack.

Clearly the stress at the tip can be many times that of the nominal applied stress. The critical stress for a material with a crack is then found to be:

$$\sigma_c = (E\gamma/2a)^{1/2}$$

This indicates that the strength of the solid, as measured by the average applied stress, *decreases* as the length of the crack in the material *increases*.

The role of cracks in the process of brittle fracture was first investigated quantitatively by Griffith almost 100 years ago. In this theory, called the *Griffith theory* of brittle fracture, fracture was considered to be due to the presence of microscopic cracks or flaws, now called *Griffith flaws*, distributed throughout the solid. Griffith suggested that failure occurred when the energy introduced into the solid due to the tension was equal to the energy needed to create two new surfaces on either side of a growing crack. (Remember that at this epoch the concept of chemical bonds was still being worked out.) Using the energy approach, the critical stress, σ_c, to fracture the solid was given by the *Griffith equation*:

$$\sigma_c = (2\,\gamma\,E/\pi\,a)^{1/2}$$

where γ is the surface energy of the solid, E is Young's modulus, and a is the length of a surface crack or half the length of an internal crack (Figure 8.16). As before, the longer the crack, the lower is the stress needed to cause fracture. In addition, it allowed the critical stress to be associated with a critical crack radius ρ_c that was defined in terms of the equilibrium atomic separation, r_o:

$$\rho_c \approx 2.5\,r_0$$

Surfaces were considered to have formed when separated by about double the critical radius, $5r_0$, or about 1 nm.

Although these calculations are of historical interest, present-day computer simulation of fracture remains underpinned by the same concepts of bond breaking and surface creation (Section 8.3.7).

8.3.2 Plastic Deformation of Metals

Metals are normally ductile and plastic deformation can be considerable. This is a valuable property of metals and is used, for example, to fabricate dish shapes of one sort or another by impressing a die into a sheet of metal. The property of ductility is revealed by the shape of a tensile test curve (Figure 8.4a,b). In the case of an ordinary metal, once the loading in a tensile test is continued beyond the yield point (at the *yield strength* and *yield stress*, point A, Figure 8.17a), *plastic deformation* occurs and the material will be permanently deformed. As the load is released, the plot of stress vs. strain will follow the line BC, which is parallel to AO. When the stress falls to zero, CD is the amount of elastic strain recovered, and OC represents the *plastic* or permanent *deformation* that has occurred. A rod so stressed will not return to its original thickness, but remains thinner than at the start.

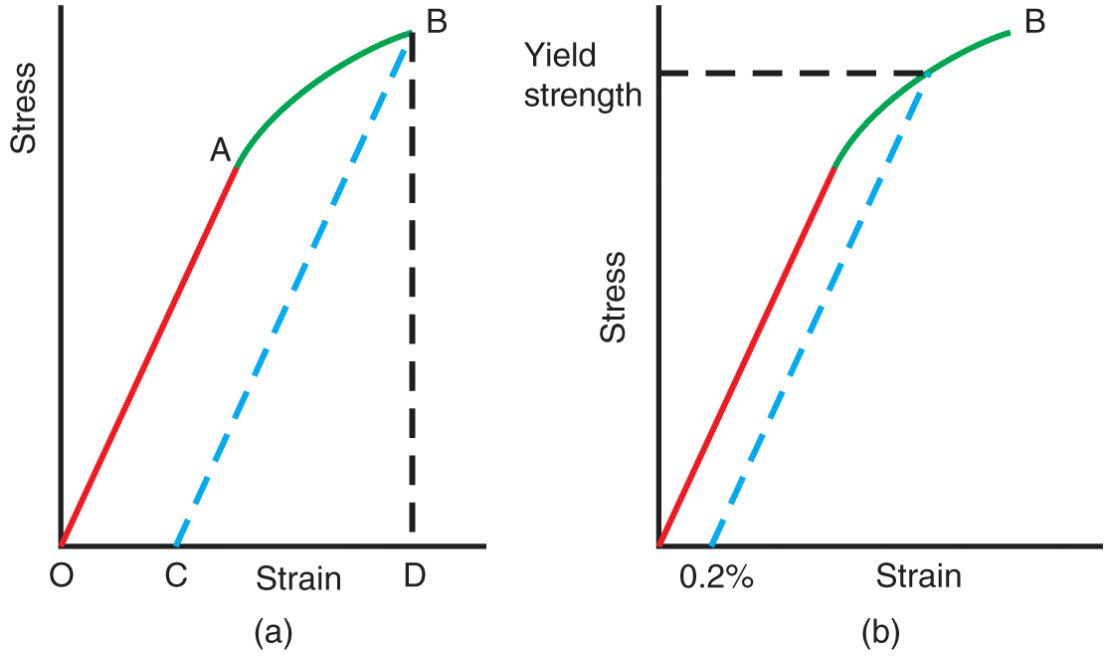


Figure 8.17 Plastic deformation: (a) the permanent distortion of the material, OC; (b) the yield strength, determined by an offset line at 0.2%.

For most materials, the transition from elastic to plastic behaviour is rarely abrupt and a single *point* does not mark the boundary between elastic and plastic deformation. In order to obtain a guide as to when a stress value passes

that required for plastic deformation to occur, it is usual to select a value of the stress that leads to 0.2% *plastic strain* (0.002 strain). This is also called the 0.2% *offset yield strength*. (This value is arbitrary, and any value could be chosen, such as 0.1% offset yield strength.) The yield strength is determined by drawing a line from the 0.2% strain point parallel to the elastic section of the curve, to intersect the stress–strain curve (Figure 8.17b).

The origin of plastic deformation is most readily understood by studying single crystals. The deformation is revealed as a series of steps or lines, where parts of the crystal have moved relative to each other so as to release the applied stress. The process can be one of *slip*, in which atom planes have slid sideways, or one of twinning. Slip is generally easier than twinning in metals, but the reverse is true for many ceramic materials. In polycrystalline solids, both twinning and slip may operate.

During slip one can imagine that a layer of spherical metal atoms rolls over the atoms in the layer below to move a small slice of crystal sideways (Figure 8.18a,b). The slices are usually the order of a few hundred atoms wide, and during slip, a number of these translated slices group together to form a *slip band*, which has the appearance of a line on the crystal face (Figure 8.18c,d). The crystal direction along the deformed rod does not change during slip. For example, when a hexagonal crystal is deformed by slip by imposing a load parallel to the crystallographic **c**-axis, the [001] direction in each part of the sample is unchanged. This differs from twinning as the atom positions are reflected across the twin plane. This means that the crystal direction along the deformed rod is also reflected across the twin boundary. Because of this the distinction between slip and twinning is best determined using X-ray crystallography or transmission electron microscopy, as these techniques are able to reveal the crystallographic relations on each side of a planar boundary.

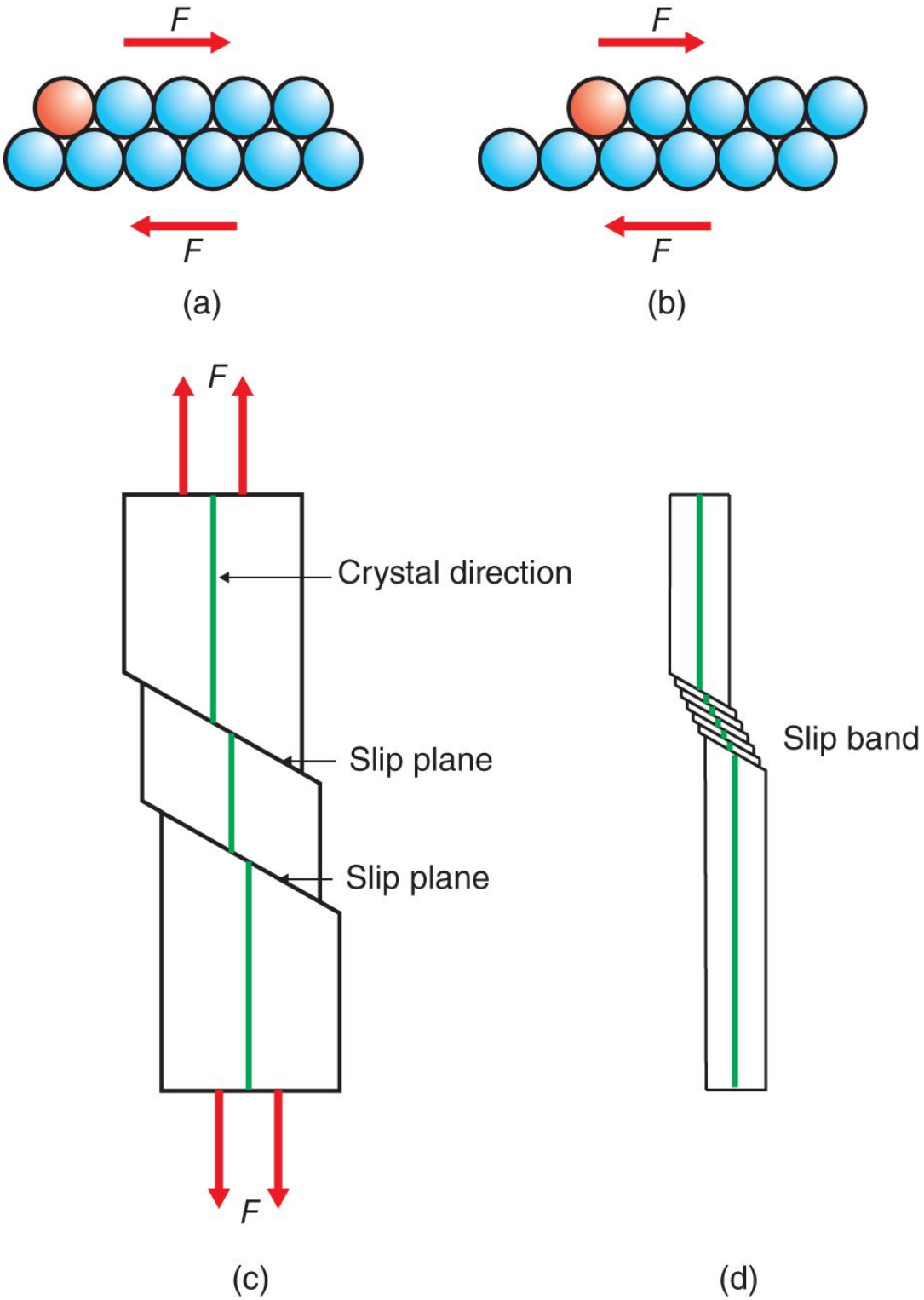


Figure 8.18 Slip: (a, b) slip following the application of a shear stress; (c) slip planes in a rod-shaped specimen; (d) slip planes aggregated into a slip band.

The amount of energy required to move a plane of atoms from one stable position to another can be estimated as the amount of energy needed to break

and reform the chemical bonds across the plane. As with the estimation of brittle fracture strength, it was found that the calculated energy was far greater than that observed to produce slip in practice. As before, the discrepancy was resolved by the introduction of defects – cracks in the case of brittle fracture and edge dislocations in the case of plastic deformation.

When a shear force is applied to the crystal, an edge dislocation *glides* (moves) along a *slip plane* that lies perpendicular to the dislocation line, to reduce the shear. This produces a permanent deformation, recognised as the a step in the crystal profile (Figure 8.19). In essence, only a few bonds are broken each time the dislocation is displaced by one step and the stress required is relatively small. If the same deformation were to be produced in a perfect crystal, large numbers of bonds would have to be broken simultaneously, requiring much greater stress. Slip via dislocation movement is now recognised as the process that successfully explains the deformation properties of metals. The movement of dislocations is constrained by the crystal structure of the metal. Some planes allow movement to take place more easily than others, and form the *slip planes*. Similarly, dislocation movement is easier in some directions than others and are called *slip directions*. The combination of a slip plane and a slip direction is called a *slip system*.

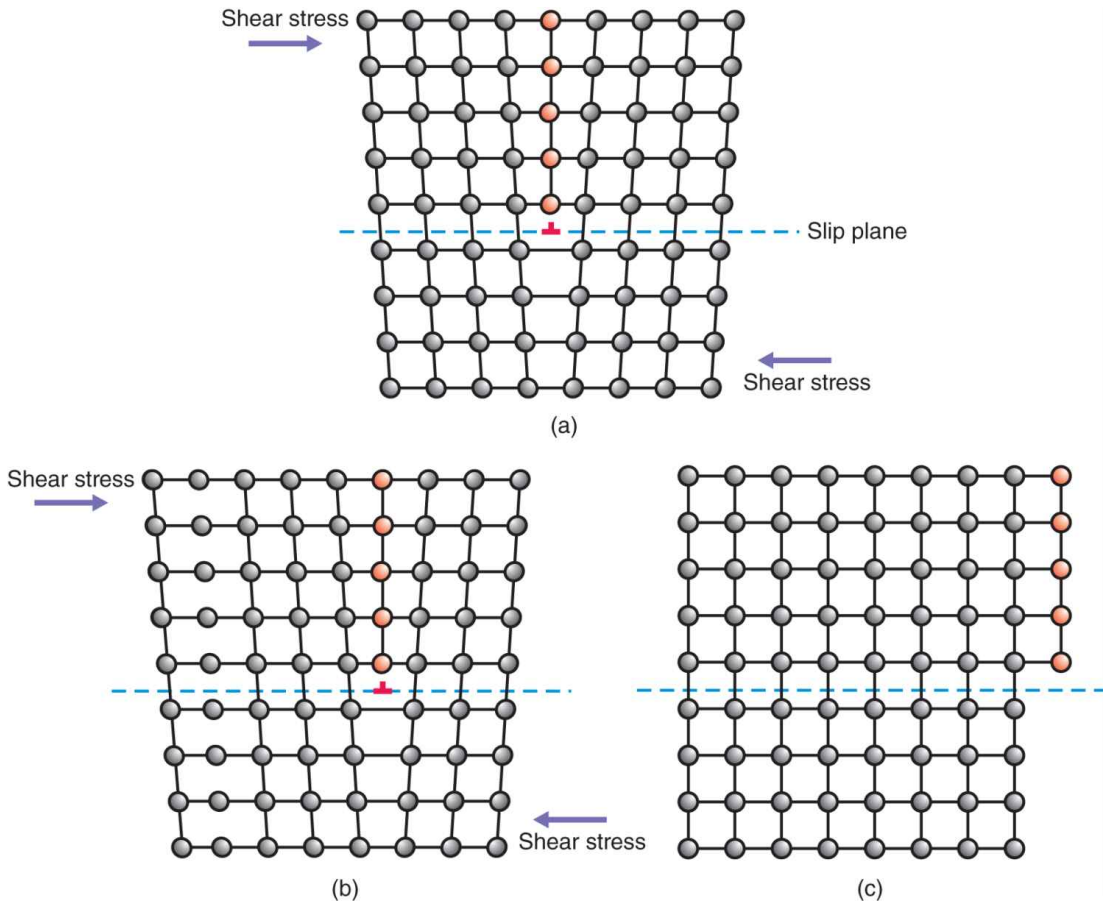


Figure 8.19 Application of shear stress to a crystal containing an edge dislocation (a, b) can allow the dislocation to move out of the crystal to leave a surface step (c).

In general, slip planes are planes with the highest area density of atoms and slip directions tend to be directions corresponding to the direction in the slip plane with the highest linear density of atoms. For example, the planes containing most atoms in metals that adopt the face-centred cubic structure are {111} planes (Figure 8.20). In these planes, the greatest linear density of atoms occurs along the ⟨110⟩ directions. In such a metal there are four different {111} planes and three different ⟨110⟩ directions. The number of slip systems is thus equal to 12. The possibilities are conveniently written as $\{111\}\langle 1\bar{1}0\rangle$. Slip in hexagonal metal crystals mainly occurs parallel to the basal plane of the unit cell, normal to the **c**-axis. The slip systems can be described as $\{0001\}\langle 11\bar{2}0\rangle$, of which there are three. Body-centred cubic metals have slip described by $\{110\}\langle\bar{1}11\rangle$, giving 12 combinations in all. Other slip systems also occur in metals, but those described operate at lowest energies.

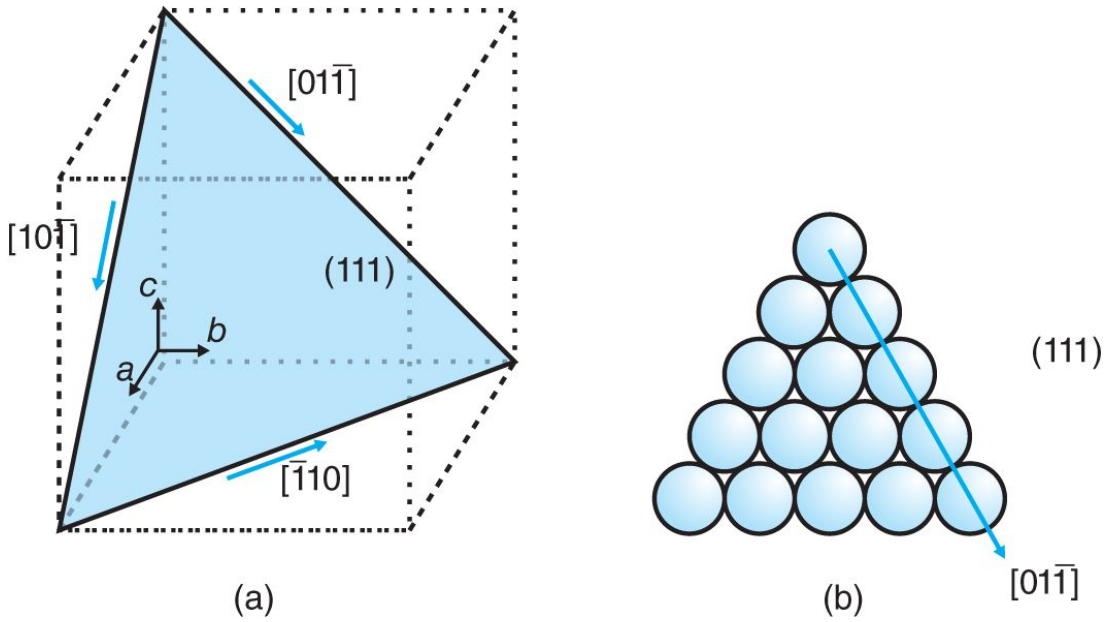


Figure 8.20 (a) The (111) plane in a crystal of a metal with the face-centred cubic structure; (b) the same (111) plane represented as atom packing. One direction, $[01\bar{1}]$, is marked.

8.3.3 Brittle and Ductile Materials

Many materials, such as ceramics, which are regarded as being predominantly brittle, show ductility at high temperatures. Similarly, many metals become brittle and loose ductility at low temperatures. It is generally found that materials are brittle at low temperatures and ductile at high temperatures, although the definition of *high* and *low* is very sensitive to the material under consideration. (For an elastomer such as rubber, room temperature is already high.) These *ductile to brittle transitions* are of great importance in engineering.

An example is provided by tungsten, a particularly brittle metal when pure. It is the best material for use in incandescent light bulbs, the normal light bulbs used throughout the most of the twentieth century and the outset of the twenty-first century, because of its high melting point. However, the brittle metal could not be drawn into wires and initially filaments were made from sintered powder, which remained brittle. The Coolidge process for making ductile tungsten, used in light bulbs after 1911, involved the addition of alloying elements, notably potassium, and drawing the wire at high temperatures. It was this invention that was directly responsible for the widespread use of inexpensive incandescent light bulbs during the twentieth century.

It is of importance to be able to roughly estimate the temperature at which a material will become ductile. The glass transition temperature, T_g, gives a good

indication of the boundary between the brittle and plastic regions of polymeric materials. In the case of ceramics, ductility begins to become important above the *Tamman temperature*, which is about half of the melting temperature, and in this regime, ceramics can deform via slip. However, dislocation motion is much harder in ceramics than in metals, partly due to the stronger bonding but also because of electrostatic repulsion between similarly charged ions, and twinning is often the preferred mechanism of releasing stress. The reason why ceramics tend to be brittle and metals ductile is therefore not due to the presence of dislocations in metals and their absence in ceramics, but because of the greater difficultly of slip in ceramics at normal temperatures. The preferred slip plane in ionic crystals with the halite (NaCl) structure such as NaCl or LiF is {110} and the slip direction used is $\langle 1\bar{1}0 \rangle$ (Figure 8.21). For the more metallic halite structure solids such as TiC the slip system is similar to that in face-centred cubic metals, {111}$\langle 1\bar{1}0 \rangle$.

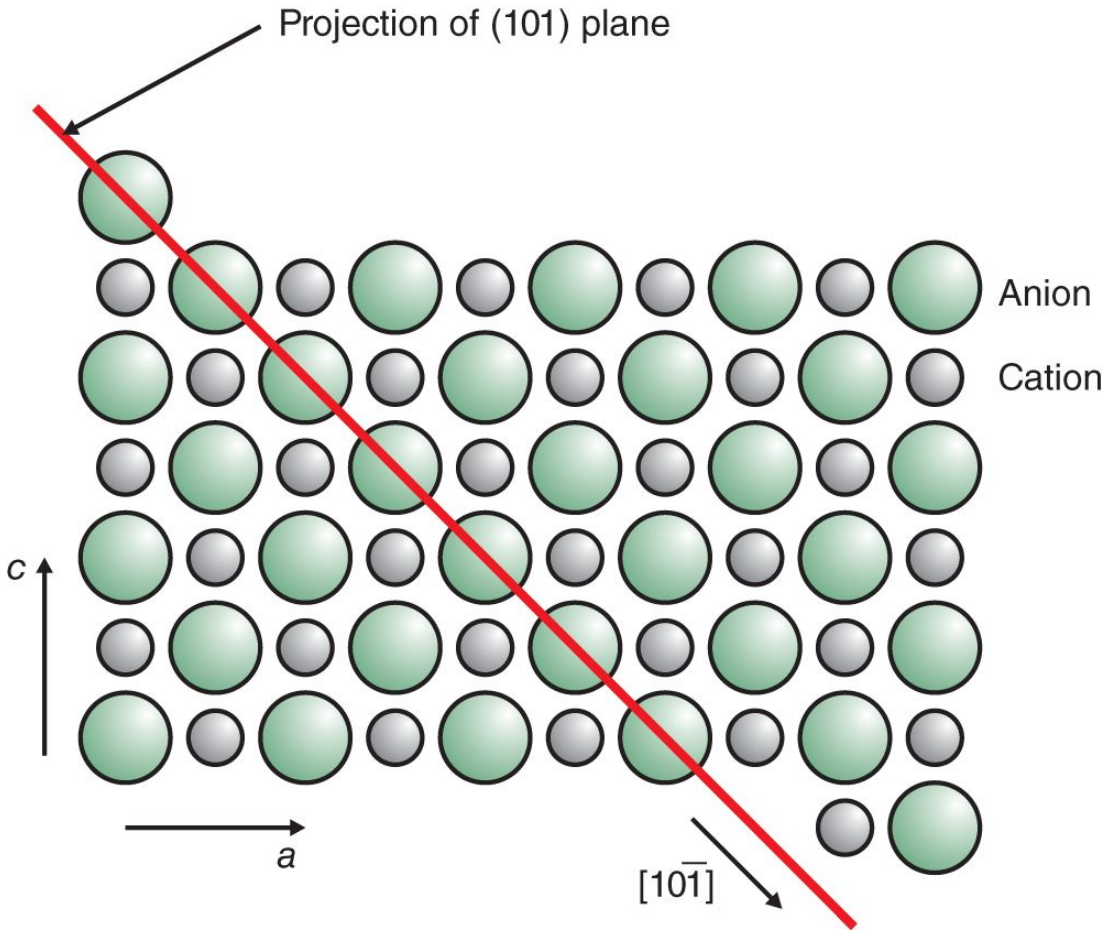


Figure 8.21 A (110) slip-plane in the halite (NaCl) structure.

For some metals, notably steels, the onset of ductility is obscured by an abrupt break in the stress–strain plot at the *upper yield point* (Figure 8.22). This is followed by deformation at a lower yield stress, at the *lower yield point*, before

the curve rises again. Between the upper and lower yield points, deformation occurs in localised regions that have the form of bands, rather than across the specimen in a uniform manner. The reason for this is that the dislocations, which would move during plastic deformation, are pinned (immobilised) in the steel, mainly by the interstitial carbon atoms present. At the upper yield stress, these become mobile, and once released they can move and multiply at a lower stress value. This is analogous to sticking and slipping when a body overcomes friction. The region in which dislocation movement starts forms a band. Ultimately the whole of the dislocation network is mobile, at which point the stress–strain curve begins to rise again, as in a ductile material.

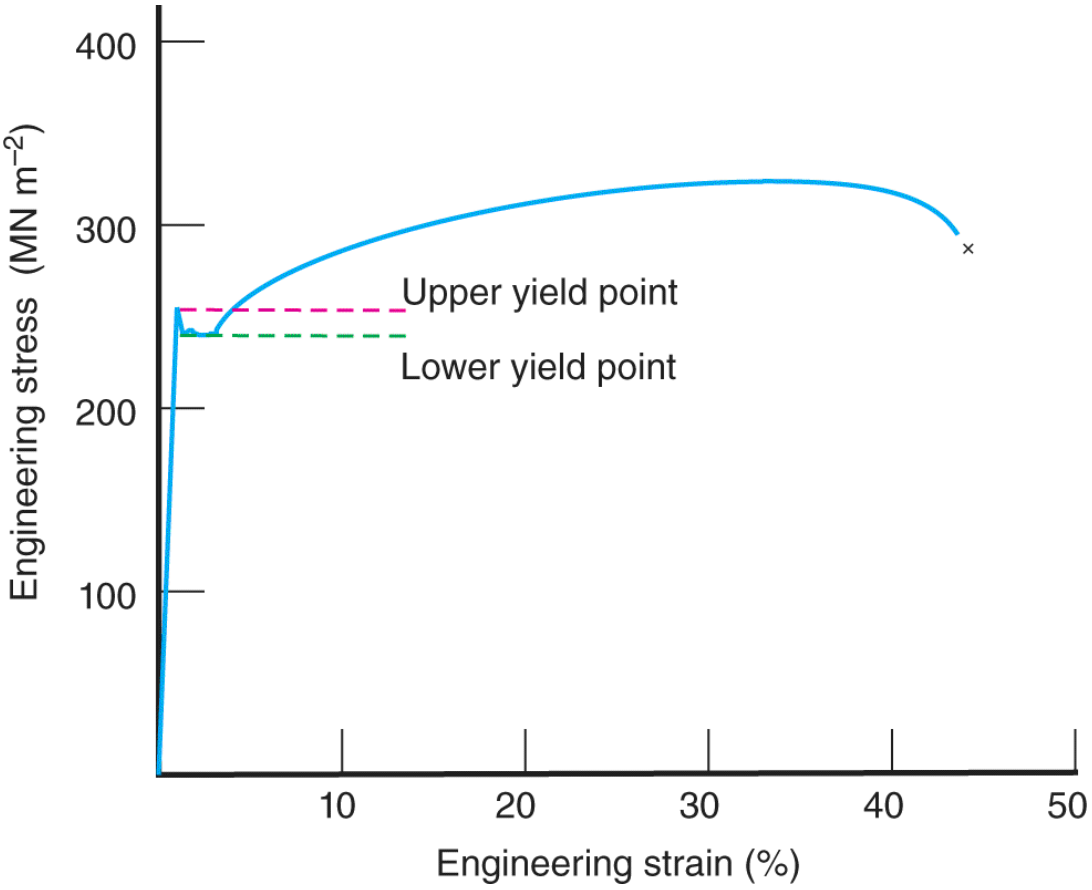


Figure 8.22 Schematic engineering stress-strain curve for a carbon steel.

The hydrogen embrittlement of steel is well-known. In the presence of hydrogen gas, many steels become weak, showing early fracture and a much-reduced toughness. This occurs in a variety of steels, including well annealed steels, which contain relatively few dislocations, and steels with a high density of dislocations. Electron microscopy has shown that hydrogen dissociates on the metal surface and H atoms preferentially migrate to dislocations, grain boundaries, and carbide precipitate-steel boundaries. The presence of the excess H atoms at these defects considerably modifies their response to

applied stress and is considered to be responsible for the resultant embrittlement of the metal.

8.3.4 Plastic Deformation of Polymers

Polymers that show considerable amounts of plastic deformation are usually partly crystalline thermoplastics or elastomers. The yield strength is taken as the initial maximum of the curve, and the tensile strength as the fracture point (Figure 8.4c). The tensile strength can be less than the yield point, especially at higher temperatures. In partly crystalline polymers, the stress is imposed upon both the amorphous and crystalline regions. The weakest links are those between the coiled chains in the amorphous regions, and initially these slip past each other as the sample elongates (Figure 8.23). Ultimately the chains become more or less aligned, and the stress now acts upon the chains themselves, the atoms of which are linked by strong covalent bonds. At this point, the crystalline regions can begin to slip and ultimately fracture will occur. Elastomers behave differently in that the coiled regions predominate, and can be uncoiled to give enormous extension before the molecules are all more or less parallel and before the strong covalent bonds are stressed.

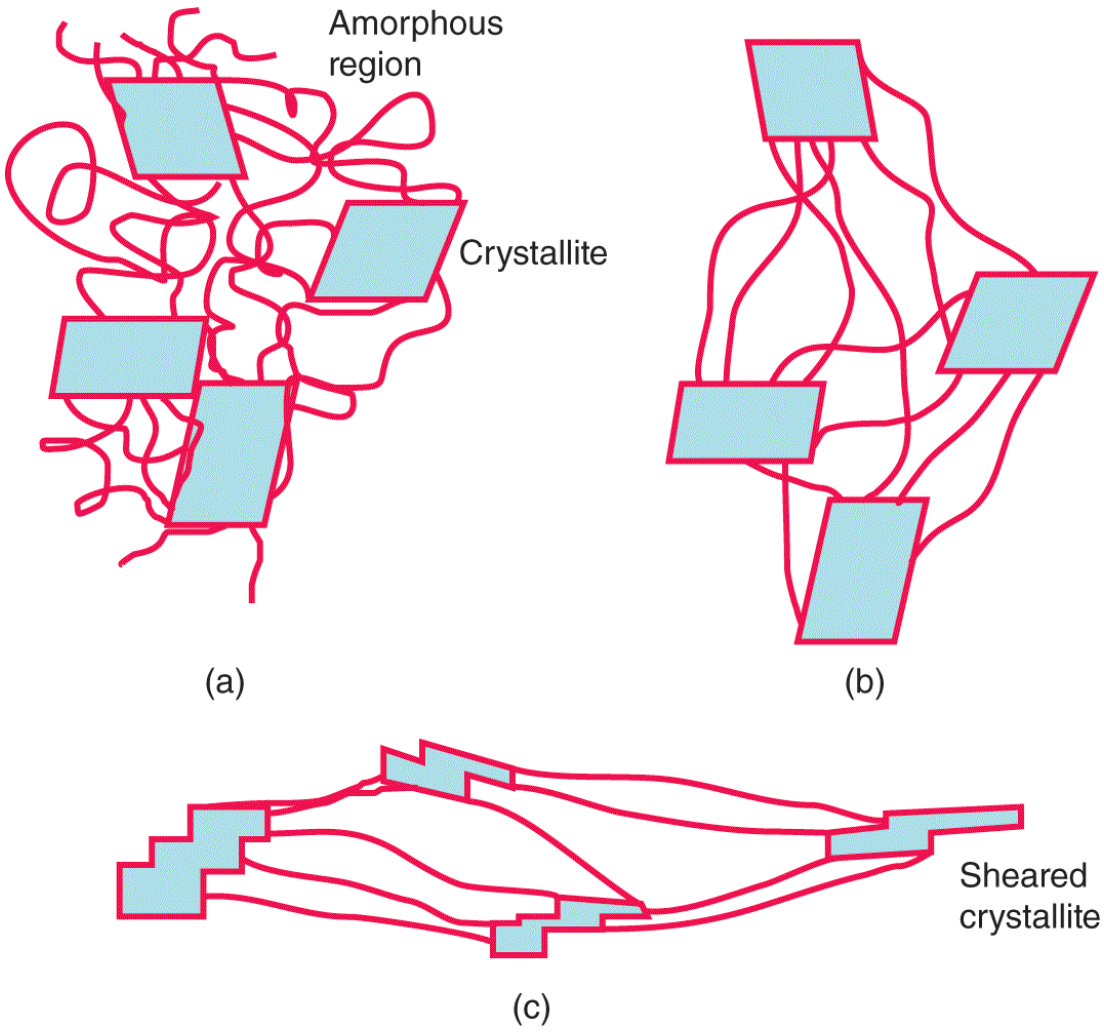


Figure 8.23 Deformation of a semi-crystalline polymer: (a) unstressed state; (b) initial loading causes chain elongation: (c) final loading causes crystal slip.

8.3.5 Fracture Following Plastic Deformation

Solids that show considerable plastic deformation before fracture generally fail in a different way than brittle solids. An important parameter that characterises this type of failure is the ductility of the material. The ductility of metals can be estimated by measuring the *percentage elongation* of a sample after fracture, where:

$$\% \text{ elongation} = [(l - l_0)/l_0] \times 100$$

where l is the final length and l_0 the initial length. During the early stages of deformation of a metal, the test sample retains its original shape, although when the stress is released the original dimensions are not recovered. When a metal fractures after considerable plastic deformation, the central part of the

sample is found to have formed a neck. This occurs after the highest point in the stress–strain curve (Figure 8.24). Because a large amount of the final deformation occurs in a rather small neck region, the percentage elongation measured will depend upon the value taken for l_0. It is for this reason that standard gauge lengths, usually 5 cm for metal samples, are used for specimens that are to undergo tensile testing.

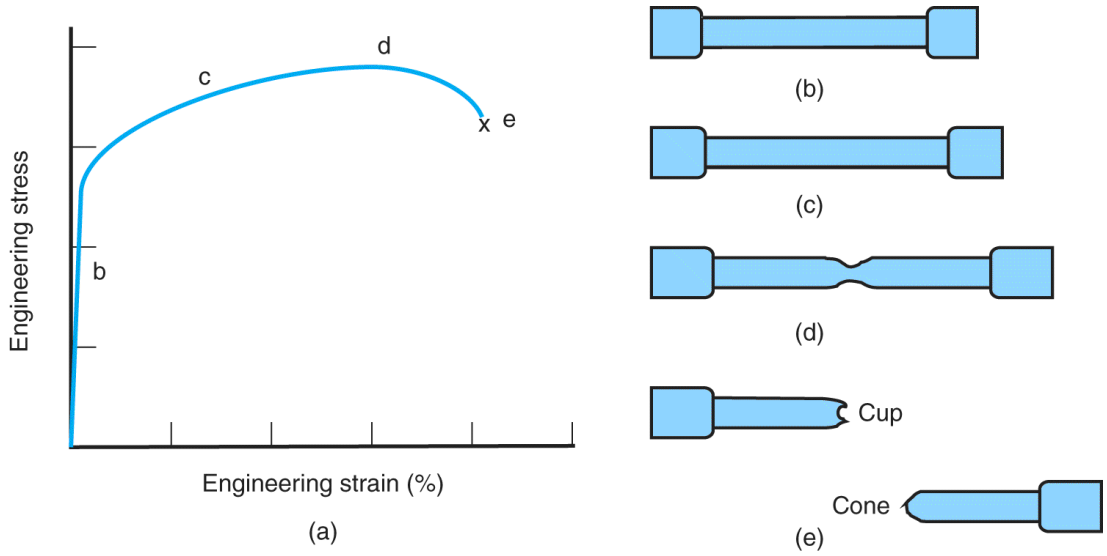
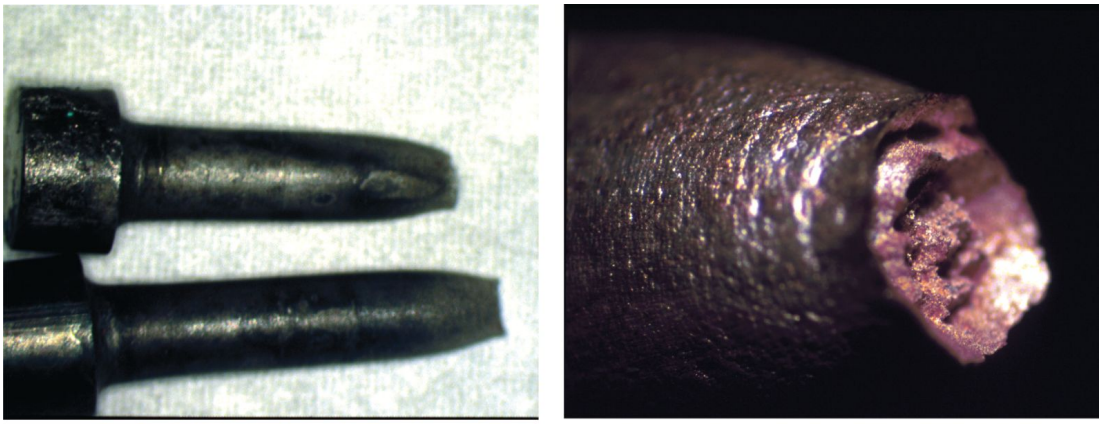


Figure 8.24 Metal ductility: (a) schematic engineering stress vs. engineering strain curve for a ductile metal; (b–e) the corresponding shape of a test piece during the test.

The ductility can also be estimated by a measurement of the reduction in the cross-sectional area of the sample at the break. The *percentage reduction in area* of a sample after fracture is:

$$\%\ \text{reduction in area} = [(A_0 - A_f)/A_0] \times 100$$

where A_0 is the initial area and A_f the final area. The fracture surface after ductile fracture has a characteristic shape, one side resembling a cup and the other a cone (Figures 8.24e and 8.25).



(a) (b)

Figure 8.25 (a) A fractured semi-ductile steel tensile test specimen showing necking and cup and cone fracture; (b) detail of fracture from a copper tensile test specimen. The neck diameter is approximately 2 mm.

The term ductility is generally reserved for metals, and in the case of polymers, ductility is replaced by *elongation*. The elongation is simply measured as the length of the polymer after stretching, l, divided by the original length, l_0, often given as a percentage:

$$\% \text{ elongation} = (l/l_0) \times 100$$

The deformation of a polymer tensile specimen is rather different from that of a metal. Polymers neck and the neck region itself will extend in a ribbon until the material ultimately tears (Figure 8.26). The term *percentage elongation at fracture* used for metals is generally replaced by *percentage elongation at break* for polymers.

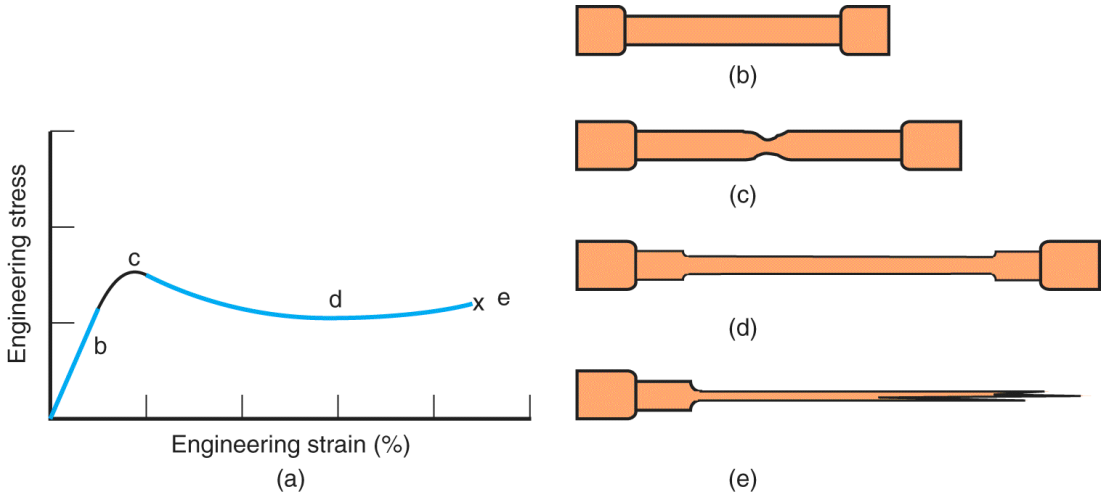


Figure 8.26 Polymer elongation: (a) schematic engineering stress vs. engineering strain curve for a polymer; (b–e) the corresponding shape of a test piece during the test.

In both metals and polymers, fracture is preceded by the formation of voids or holes. In a metal, these voids tend to segregate at the central region of the neck, and coalesce into a large crack. This crack ultimately spreads to the edge of the neck. The fracture surface has a pulled out fibrous texture exhibited in both the 'cup' and 'cone' regions of the break. In the case of a polymer sheet, the holes are formed by the continual alignment of the polymer chains. Ultimately the separate holes coalesce to from a long tear. As with metals, the edges of the tear have a pulled out, fibrous texture.

8.3.6 Strengthening

Pure metals tend to be soft and relatively weak. Ductility can be reduced and the metal strengthened by restricting dislocation movement. However, if this is continued too far the metal will become brittle. A compromise is often required. Historically, three methods have been used to achieve this objective, *alloying*, *reduction in grain size*, and *work hardening*.

Alloying, first widely used to transform soft copper into much stronger bronze, relies upon blocking dislocation movement due to strain by the introduction of impurity or dopant atoms. This strain field impedes dislocation movement because dislocations also generate a strain in the structure and the two components mutually repel each other to hinder slip. Additionally, if sufficient of the second component is added, precipitates of a second phase can form in the crystal matrix, leading to *precipitation strengthening*. These hinder dislocation movement simply by blocking slip planes.

Reduction of grain size has a similar strengthening effect because dislocations can only cross from one grain to another by expending energy, and often grain boundaries block almost all dislocation movement. Metal strengthening by

way of grain size reduction is widely described in terms of the *Hall–Petch equation*:

$$\sigma(d) = \sigma_0 + k_{\mathrm{HP}}/\sqrt{d}$$

where $\sigma(d)$ is the yield stress of a material consisting of grains of size d, σ_0 is the yield stress of a large single crystal, sometimes called the bulk stress, or simply a material constant, and k_{HP} is a constant appropriate to the material, sometimes called the strengthening coefficient. However, at a critical grain size, approximately 20 nm, the relationship seems to fail, as the mechanism of deformation changes from dislocation movement to shape change via grain-boundary movement.

Although the broad viewpoint that smaller grains gives stronger materials holds true, there is uncertainty over whether the Hall–Petch equation is valid for all materials. In particular, thin semiconductor films, foils, and thin wires do not accord well with the Hall–Petch formalism, and even in metal samples, there is considerable difficulty in measuring grain size throughout the sample accurately. Because of this, although the general premise that smaller grains gives stronger materials holds, there have been a number of alternative modified forms of the equation, such as:

$$\sigma(d) = \sigma_0 + k_{\mathrm{HP}}/\sqrt{d} + k_1/d$$

where k_1 is a constant and:

$$\varepsilon(d) = \varepsilon_0 + k \ \ln \ d/d$$

where $\varepsilon(d)$ is the elastic strain in a material of grain size d, k is a dimensionless constant, approximately equal to 1, and ε_0 is the elastic strain defined by Young's modulus equation, σ/E have also been described. (See Further Reading for more information on these alternatives.)

Work hardening, which is also called *cold working* or *strain hardening*, is the result of repeated deformation brought about by heating and mechanical deformation, which causes dislocation numbers to increase. Initially this can lower the strength but beyond a certain dislocation density, dislocations become tangled and movement becomes impeded. There are a number of mechanisms by which new dislocations can form and so become entangled, but most require that an existing dislocation becomes pinned in the crystal, so that glide can no longer occur. One of these involves a length of dislocation pinned at each end, a *Frank–Read source* (Figure 8.27a). When stress is applied the pinned dislocation cannot glide to relieve the stress, but it can bulge out from the pinning centres to achieve the same result (Figure 8.27b). Further stress

increases the degree of bulging, until both sides of the bulge can unite to form a dislocation loop plus a new dislocation between the pinning centres (Figure 8.27c–e). A Frank–Read source can continually emit dislocation loops during stress, thus significantly multiplying the dislocation density in a crystal.

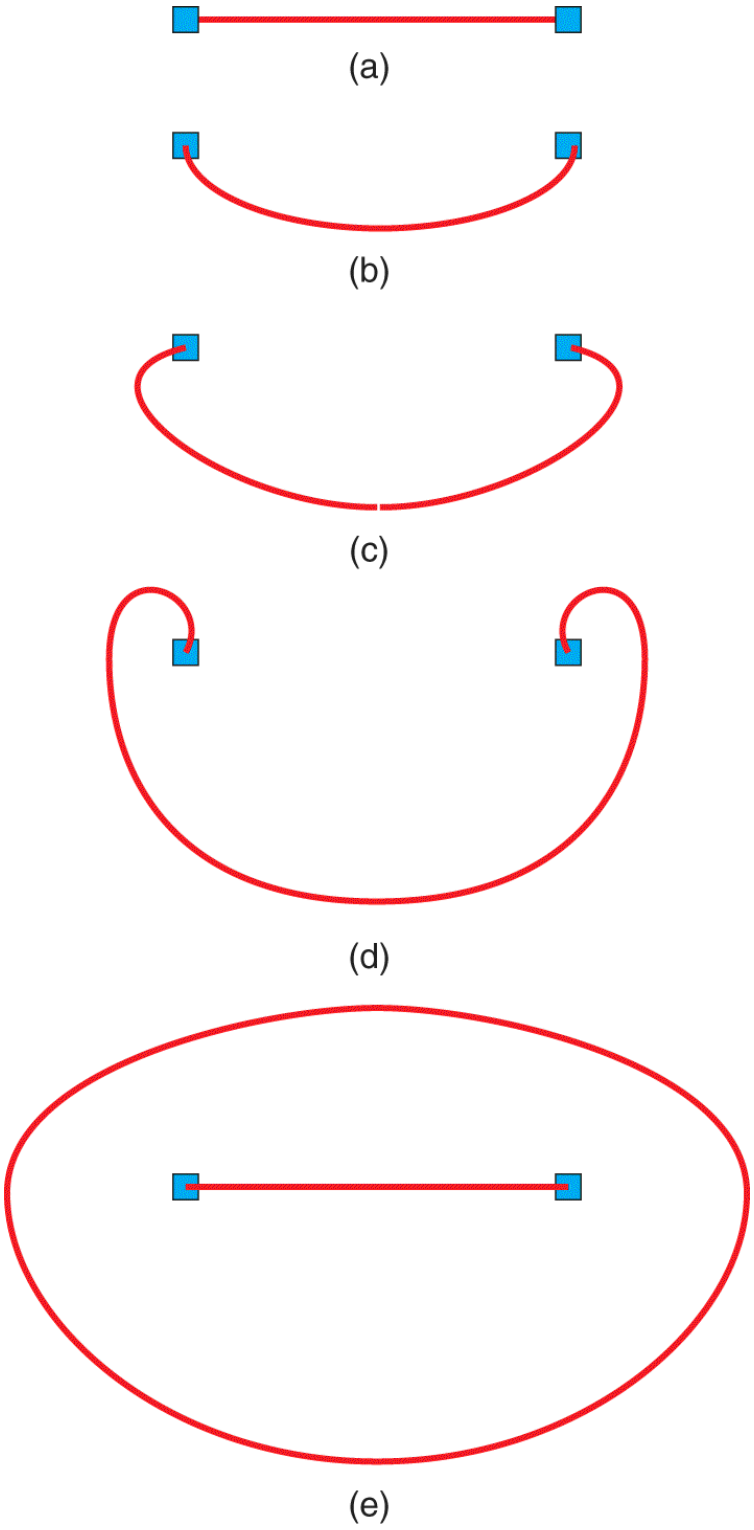


Figure 8.27 Dislocation multiplication at a Frank–Read source: a dislocation pinned at each end (a) can respond to stress by bowing (b), eventually generating a dislocation loop plus the original pinned dislocation (c–e).

Ceramics normally fail in a brittle mode and can be strengthened by the removal of the Griffith flaws. Although this is not always possible, careful protection of the surface from contamination, chemical reaction, or mechanical damage will help. Optical fibres, when freshly drawn, are extremely strong, but reaction of the surface with water vapour present in the air rapidly degrades the fibre strength. For this reason, freshly drawn fibres are immediately coated with a polymer to maintain strength. Similarly, single crystal ceramics can have surfaces carefully polished, to increase strength. Polycrystalline ceramics need to be fabricated in such a way as to minimise the amount of internal porosity.

The strength of a polymer is generally enhanced by increasing the amount of crystalline phase present with respect to the amount of amorphous phase present. Similarly, cross-linking will transform a soft thermoplastic into a hard and brittle solid. A strengthening route available to polymer science that is not available for metals and ceramics is the formation of block copolymers. Copolymers of an elastomer and a fibre will give a strong but elastic fibre, and copolymers of elastomers with brittle polymers will give a tough and flexible solid.

For many engineering (and biological) applications, solids are strengthened by combining materials, rather like block copolymer strategy. In the case of metals and ceramics, the two phases remain separate, and the material is called a *composite* (Section 8.6).

8.3.7 Computation of Deformation and Fracture

Deformation, fracture, and strengthening are of vital importance. Attempts to quantify brittle fracture in the early years of the twentieth century, described earlier, were the start of a long evolution of mathematical approaches to the subject. At present, computations of the response of solids to applied stress, the accompanying deformation and eventual fracture are all the object of considerable effort. The methods of computation are not new, but have becoming increasingly sophisticated as computing power has increased.

For details at an atomic level, the simplest method, in principle, is to apply the methods of *atomistic simulation* (Section 1.2.3) to describe deformation and fracture. In the computational procedure, electrons are ignored and the interactions between the atoms that make up the solid are represented by pair potentials – potential energy curves that vary with atomic separation. Using static lattice methods, the unit cell of the structure to be is subjected to forces or pressures causing deformation, which, in turn, allows the elastic constants to be estimated.

If the atoms are given velocities and their trajectories calculated using Newton's laws, *molecular dynamics simulations* will allow dislocation movement or the growth of a crack to be followed at an atomic level of

discrimination. Naturally, although simple in concept, the practice is by no means trivial in reality. Extremely sophisticated computation is required to ensure that sufficient atoms are included in the simulations to give meaningful results.

Equilibrium properties can also be assessed via density functional theory calculations (Section 1.6). This powerful computational technique, mainly associated with band structure calculations, can provide values for the lattice parameter of a crystal, elastic constants such as B, G, E, and ν, and the hardness of specific crystal faces.

These calculations, which are at a microscopic level, are quite unsuitable for the study of large structures such as aeroplane fuselages, bridges, or cars. The computational technique used here is that of *finite element analysis*. This mathematical method was developed in the latter half of the twentieth century, as computational power became more widely available. In this technique, the massive body to be simulated is conceptually decomposed into a mesh of smaller fragments (elements), very often small tetrahedra, chosen so as to faithfully model the body as a whole. The effect of forces or deformations acting upon the body are then applied to the elements and the effects then computed. In this way the behaviour of the whole structure and its most likely mode of failure may be assessed.

The technique is now commonplace across a wide variety of disciplines. Apart from traditional engineering, it is used to study the mechanical properties of, for example, biomaterials such as dinosaur bones, nacre, and sea shells, all of which are composites with very desirable mechanical properties.

8.4 Time-Dependent Properties

8.4.1 Fatigue

When placed under a *cyclical* or *varying* load for a period of time, a material may fail due to *fatigue*. This is invariably at a much lower stress than the part can withstand during a single application of the stress and failures under repeated cycling are called *fatigue failures*. Fatigue affects moving parts in machinery and is also of importance in components that are slightly flexed in a repetitive fashion, such as an aircraft fuselage under the varying pressure regimes in the atmosphere.

The stress pattern imposed upon a solid can vary greatly, from a completely irregular pattern to regular sinusoidal changes in which the stress can always remain positive, or can vary from positive to negative (Figure 8.28). In the laboratory, a sample is tested by imposing a cyclic strain and testing until the part fails. The results are plotted as the stress amplitude, σ, plotted against the logarithm of the number of cycles that the sample can tolerate, log N, before failure. Ferrous alloys (alloys of iron) have a behaviour represented by a curve of σ vs. log N that shows a distinct change of slope, called a *knee* (Figure 8.29). In curves of this type the part of the curve parallel to the log N-axis is called the *endurance limit* or *fatigue limit*. At stress levels below the endurance limit, fatigue failure never occurs. In most non-ferrous alloys and pure metals such as copper or aluminium, the curve is smooth up to fracture (Figure 8.29).

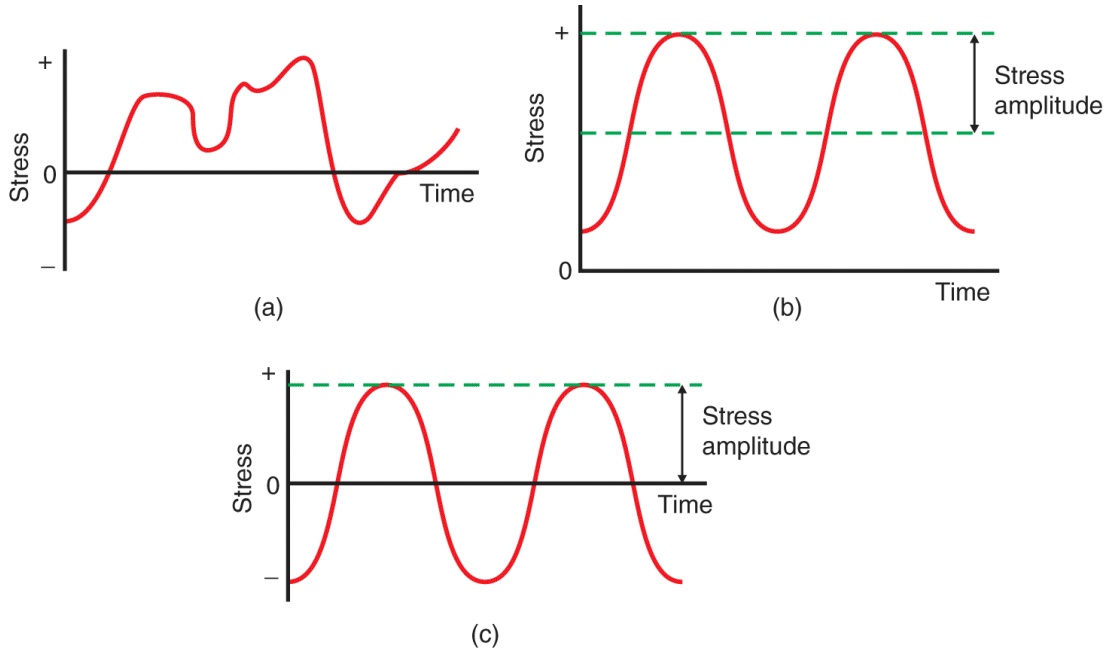


Figure 8.28 Stress cycles: (a) irregular; (b) positive sinusoidal cycle; (c) positive and negative sinusoidal cycle.

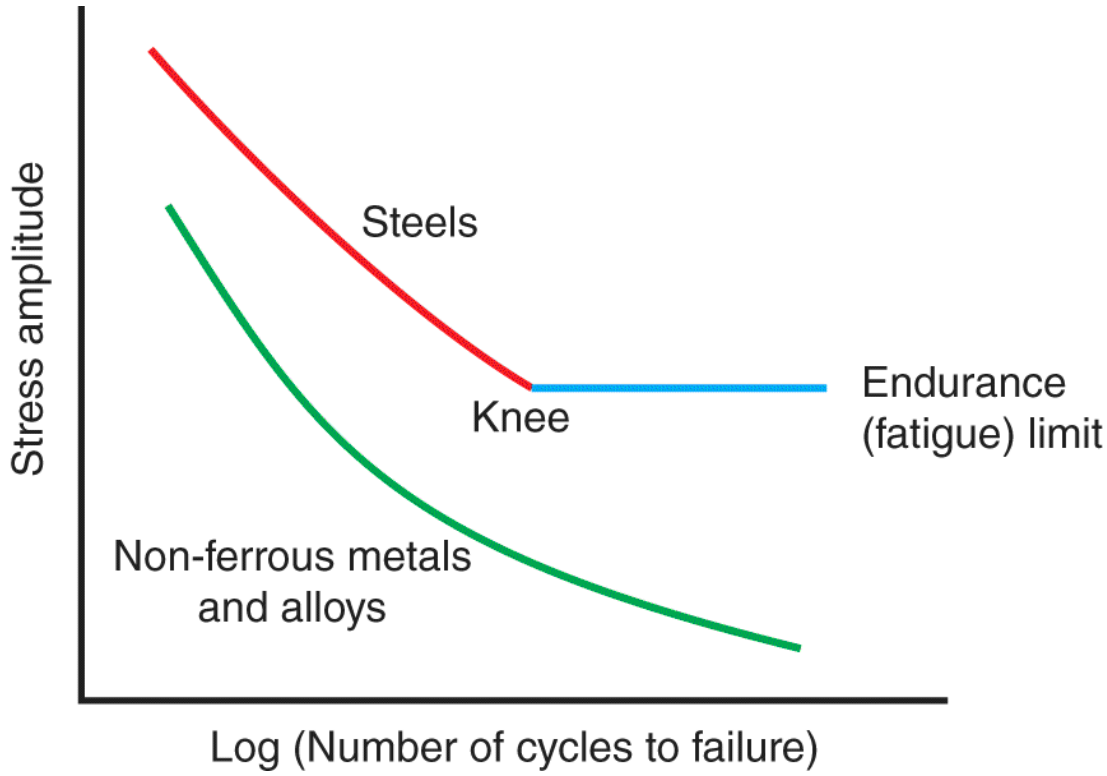


Figure 8.29 Stress amplitude plotted against log (number of cycles to failure) for typical steels and many non-ferrous metals and alloys.

The process of fatigue is empirically divided into a number of stages. Initial damage occurs in *Stage 1*. An originally perfect specimen will crack at an angle of approximately 45° to the direction of the cyclic stress along a slip plane (Figure 8.30). The surface also becomes distorted at the crack, giving rise to intrusions and extrusions, called *persistent slip bands*. The crack propagates across a small number of grains at this time. *Stage 2* of the failure is indicated by a change in crack direction to approximately normal to the cyclic stress direction along a cleavage plane. During this stage, the crack opens a little during each cycle, gradually spreading across much of the specimen. In practice, it is not easy to separate stages 1 and 2 of the process. *Stage 3* represents the final failure of the component. This occurs when the crack in stage 2 has grown to such an extent that rapid crack propagation occurs and the whole part fails catastrophically, by ductile or brittle failure, depending upon the nature of the material.

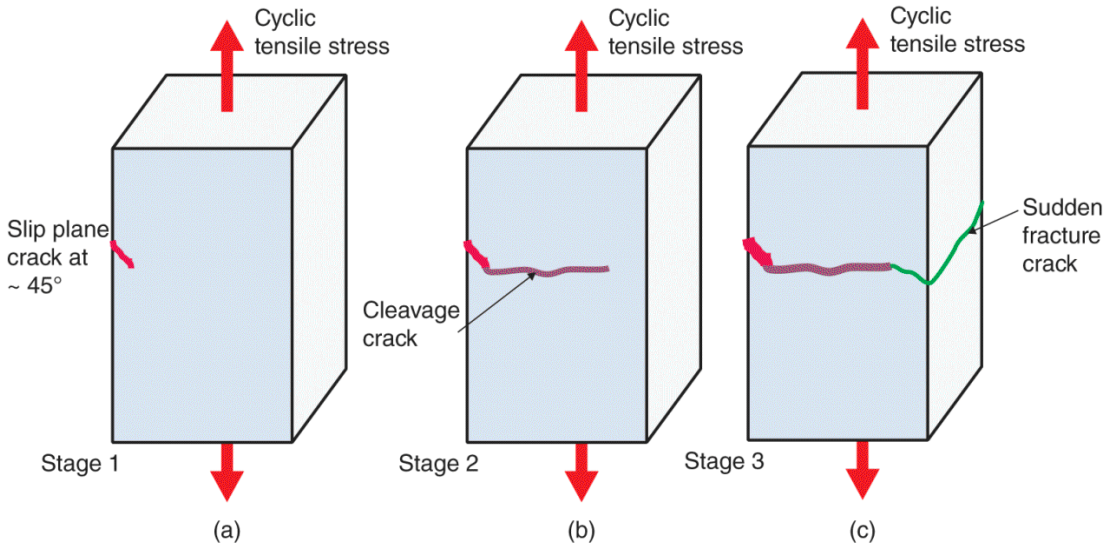


Figure 8.30 Fatigue failure: (a) stage 1, initial crack formation; (b) stage 2, crack growth; (c) stage 3, failure due to rapid crack propagation.

As might be anticipated, an important factor in fatigue is the state of the surface. Surface flaws, roughness, notches, sharp edges, holes, or abrupt changes in cross sectional area, called *stress raisers*, can greatly increase the stress locally, thus initiating stage 1 of the process. (Aircraft have rounded windows, rather than rectangular ones, for this reason!) The local environment is also relevant, and fatigue initiated by corrosion, called *corrosion fatigue*, is often important in chemical plants.

8.4.2 Creep

Creep is the *gradual elongation* of a material (strain) over time under a constant load (stress). Although creep is not very important at normal temperatures for most metals and ceramics, creep of polymers can be extensive. For metals it is especially important in high temperature applications such as turbine blades, where even the smallest change in dimensions can lead to catastrophic failure. Creep is mainly due to progressive plastic deformation of the solid under the constant load and is caused by the transport of material. This can be attributed to the movement of crystallites at grain boundaries, movement of dislocations, or diffusion of atoms. The mechanism of creep will often depend upon the temperature as well as the solid in question. For example, diffusion is more likely at high temperatures in metals and ceramics. Low temperature creep in plastics is due to movement between the coiled polymer chains in the material.

Creep is usually displayed as a graph of strain against time; a *creep curve* at a constant temperature (Figure 8.31). The section of the curve OA is the extension that occurs when the specimen is first stressed. This is elastic deformation for most ordinary loads, and corresponds to the straight-line part

of a stress–strain curve. Creep as such is indicated by the remaining parts of the curve. The whole of the creep curve is enhanced at higher temperatures or at greater stress loading (Figure 8.32). With reference to this figure, when the stress values, σ_1, σ_2, and σ_3 are equal, the temperature values lie in the sequence $T_1 > T_2 > T_3$. When the temperatures T_1, T_2, and T_3 are equal, the stress values lie in the sequence $\sigma_1 > \sigma_2 > \sigma_3$.

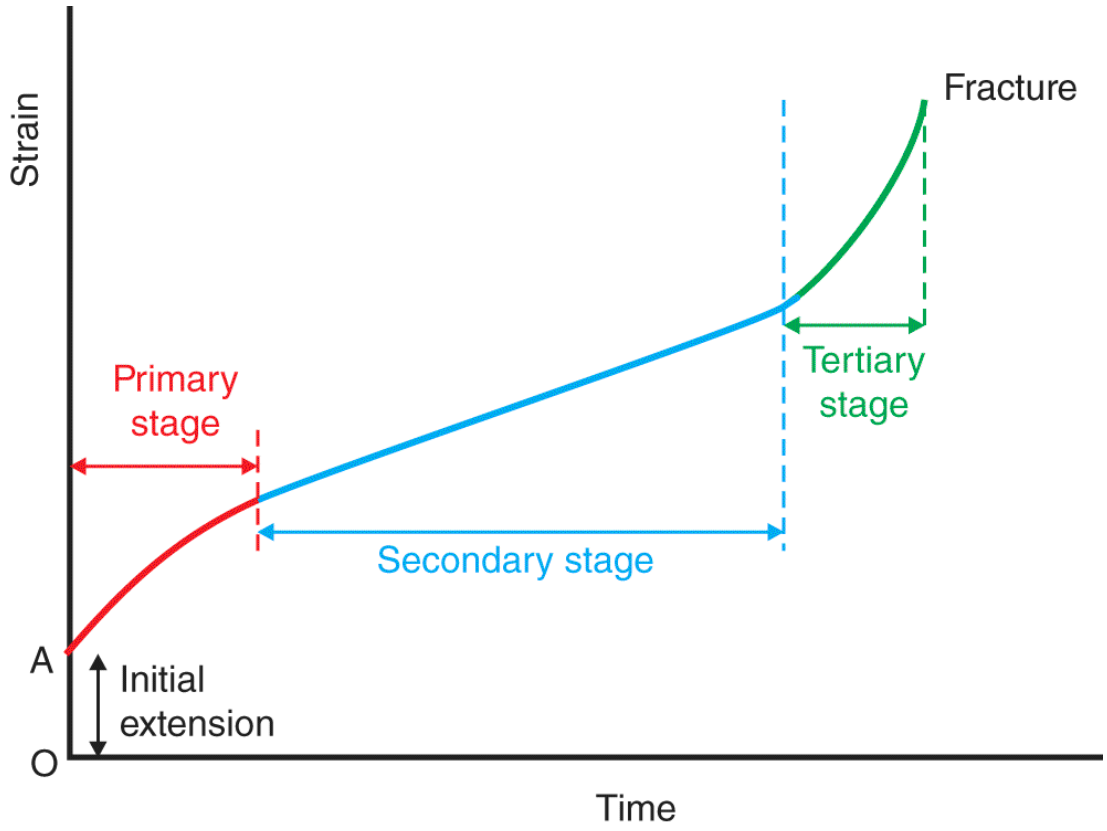


Figure 8.31 A normal creep curve for a solid subjected to constant load at a constant temperature.

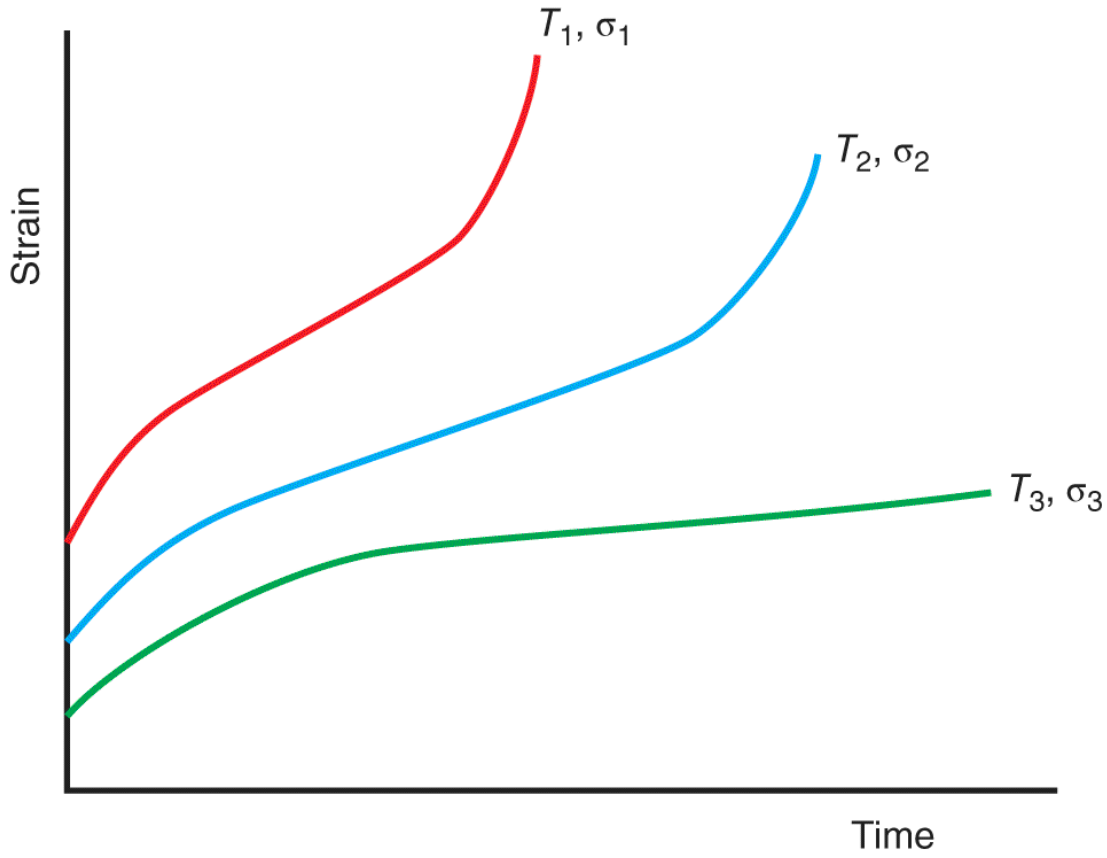


Figure 8.32 The effect of temperature on creep.

The initial deformation, which occurs in the *primary* stage, is referred to as *primary creep*. In this regime, the creep rate is continually *decreasing*. The dimensional change is largely believed to be the result of dislocation movement in the case of metals and ceramics, and by polymer chains sliding past each other in plastics. Primary creep is described by one of two equations:

$$\varepsilon = \alpha \log t$$

where ε is the strain, t is the time, and α is a constant. This equation is applicable at lower temperatures, fitting well with polymers and rubber. At higher temperatures the equation:

$$\varepsilon = \beta t^m$$

is applicable, where ε is the strain, β and m are constants, and t is the time. The value of m can vary between about 0.03 and 1.0, a range of about 10^2.

The decreasing rate of creep is attributed to cold working in the case of metals. Here the dislocations move until they begin to impede each other. In the case of polymers, uncoiling of tangled polymer chains proceeds at slower rates as

the energetically easiest rearrangements give way to reorganization involving cross-linking or crystal movement.

The middle part of a creep curve is linear and refers to *secondary* or *steady-state* creep. In this case, the internal relaxation in the solid is balanced by the strain induced by the load. This linear portion of the curve is approximately described by an equation:

$$\varepsilon = K\,t$$

where ε is the strain and K is a constant, equal to the slope of the linear portion of the curve.

A number of mechanisms have been proposed to explain secondary creep. At lower temperatures, the rearrangements of dislocations into lower energy configurations, by dislocation glide, are thought to be the most important process. As the temperature increases, dislocation climb – a process whereby point defects add or subtract from a dislocation line to cause it to shorten – is thought to become increasingly important. These processes lead to the reduction in the number of dislocations present, a process called *recovery*. This mechanism of creep, in which dislocation movement controls the creep rate, is called *dislocation creep* or *power law creep*, as the slope of the creep curve, K, is given by an equation of the general type:

$$K = \mathrm{d}\varepsilon/\mathrm{d}t = \pi^2\, D\, \sigma^n/(\sqrt{bN})\,(G^n)\,k_\mathrm{B}T$$

where D is the diffusion coefficient of the rate limiting species, G is the shear modulus of the solid, b is the Burgers vector of the rate-limiting dislocations, N the number of dislocations, k_B is Boltzmann's constant, T the temperature (K), and σ is the stress. The exponent n has a magnitude of about 5–6 for metals and ceramics and about 2 for plastics.

At higher temperatures, diffusion is widely believed to control the rate of creep in ceramics and metals, and this regime is labelled *diffusion creep*. Two mechanisms have been suggested. At relatively lower temperatures, creep rate is limited by diffusion along the grain boundaries (short-circuit diffusion), referred to as *Coble creep* (Figure 8.33a). In Coble creep atoms diffuse to grain boundaries parallel to the compressive stress. The slope of the secondary creep curve, K, for Coble creep is given by:

$$K = \mathrm{d}\varepsilon/\mathrm{d}t = A\,\sigma\,\Omega\,D_\mathrm{b}\,\delta/k_\mathrm{B}Td^3$$

where A is a numerical constant of the order of 47, σ is the stress, Ω is the vacancy volume, D_b is the grain boundary diffusion coefficient, δ is the grain

boundary width, d is the grain diameter, k_B is the Boltzmann constant, and T is the temperature (K).

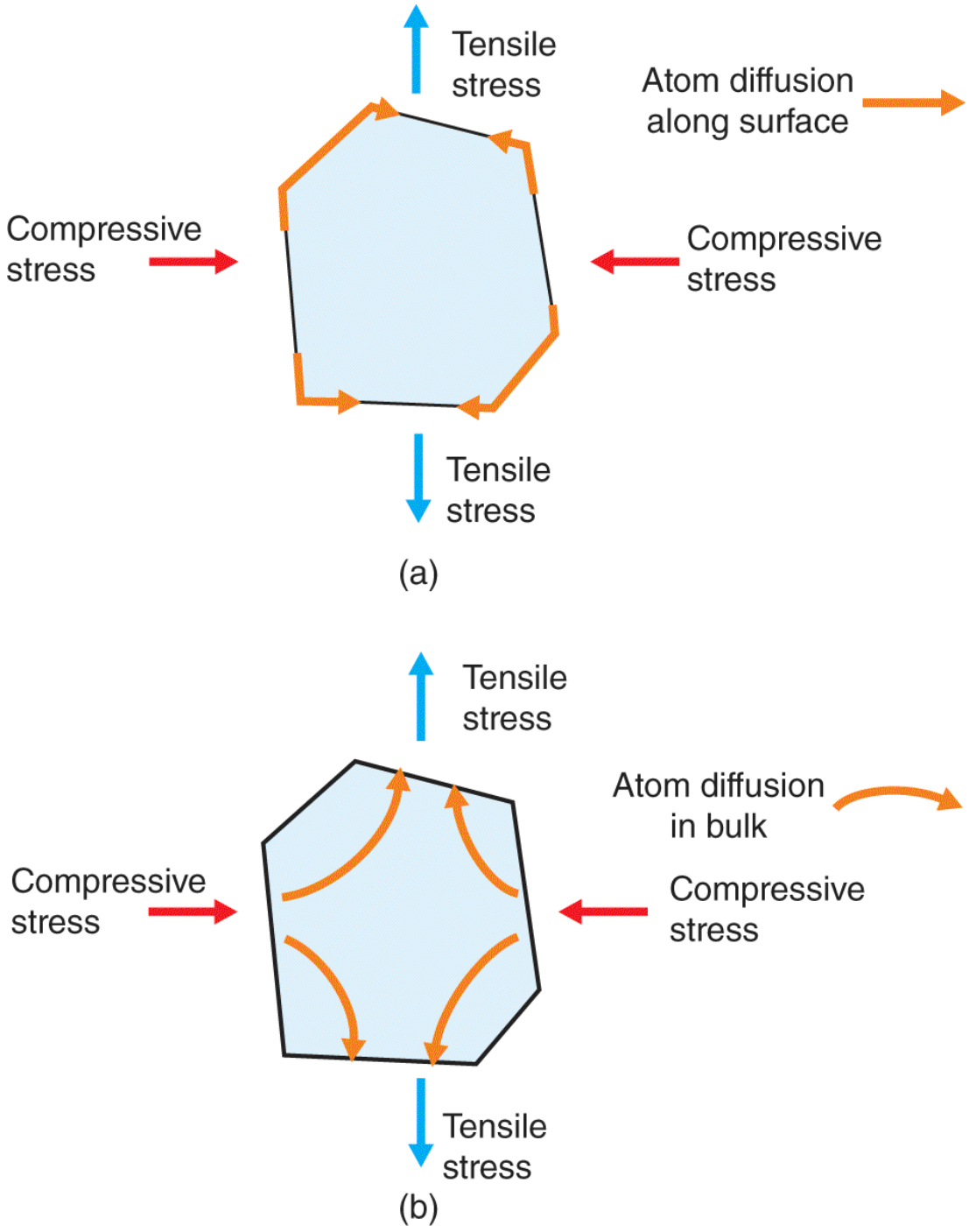


Figure 8.33 Creep mechanisms: (a) Coble creep: atoms diffuse along grain boundary surfaces; (b) Herring–Nabarro creep: atoms diffuse within grains.

At relatively higher temperatures, diffusion of atoms within the grains (bulk diffusion), becomes more important in both metals and ceramics. This is called

the *Herring–Nabarro creep* (Figure 8.33b). In this model, the slope of the secondary creep curve is given by:

$$K = \mathrm{d}\varepsilon/\mathrm{d}t = A\,\sigma\,\Omega\,D/k_{\mathrm{B}}\,T\,d^2$$

where A is a numerical constant of the order of 13, σ is the stress, Ω is the atomic volume, D is the vacancy diffusion coefficient for diffusion within the grains, d is the grain diameter, k_{B} is the Boltzmann constant, and T is the temperature (K). In Herring–Nabarro creep atoms diffuse within grains towards boundaries parallel to the compressive stress. The key difference between these two equations is that Coble creep is proportional to $1/d^3$ and Nabarro–Herring creep is proportional to $1/d^2$.

In both of these processes, atom diffusion is away from boundaries under compression towards boundaries under tension, leading to relief of the stress at the grain boundaries generated by the load on the sample. This causes the grains elongate along the direction of the stress. (Note that diagrams often show the flow of vacancies rather than the flow of atoms. Vacancy flow is opposite to atom flow, and is from boundaries more or less parallel to the compressive stress to boundaries more or less perpendicular to the compressive stress.)

Both of these equations reveal that the rate of diffusion creep increases rapidly as the grain size decreases. This has implications for the creep of thin films and nanomaterials. In such fine-grained solids, creep is often surprisingly large, compared with the bulk phase, due to grain boundary effects. Recall that strengthening of a metal can be brought about by decreasing grain size, so that creep and strengthening are, to some extent, in conflict, so that a compromise of grain size and texture is often desirable.

The various mechanisms proposed for secondary creep do not operate in totally separate temperature regimes. All tend to overlap, and the rate of creep can often best be generalised as a complex function of the many materials parameters used in the equations above. Despite this complexity, the temperature dependence of the creep rate can often be given by an Arrhenius-type equation:

$$K = \mathrm{d}\varepsilon/\mathrm{d}t = A\,\sigma^n\,\exp\left[-E/RT\right]$$

where A and n are constants and E is the activation energy for creep, which is found to have a similar magnitude to the activation energy for grain-boundary diffusion (Coble creep) or self-diffusion (Nabarro–Herring creep) in the material. The constant n varies from about 2 to about 6, dependent upon the material in question.

The *tertiary* stage, or *tertiary creep*, is characterised by a rapid increase in the strain, leading to failure or *creep rupture*. At this stage, voids form in the region of the fracture and there is considerable grain boundary movement. As would be anticipated, this part of the curve is difficult to analyse theoretically.

8.5 Nanoscale Properties

In this section, three illustrative examples of the impact of scale upon mechanical properties are outlined. The first, solid lubricants, underlines the connection between crystal structure and observed mechanical properties. In the second, auxetic materials, crystal structure and microstructure combine to produce materials with negative values of Poisson's ratio. Finally, methods of assessing the mechanical properties of thin films and nanowires are described.

8.5.1 Solid Lubricants

Solid lubricants embody the opposite properties to those of hardness; they are soft and feel greasy to the touch. As with all lubricants, solid lubricants reduce friction and wear, and prevent damage between surfaces in relative motion. Solid lubricants have advantages over the normal liquid lubricants in certain conditions. These include high temperatures, where liquids can decompose or boil, low temperatures where liquids can freeze, or in a vacuum, where liquids can evaporate or contaminate the vacuum. In this latter respect, space applications are important and new solid lubricants that can operate under the harsh conditions imposed by space exploration are being continually sought.

Solid lubricants need low shear strength in at least one dimension. They fall into three main classes, inorganic solids with lamellar (layer-like) crystal structures, solids that suffer plastic deformation easily, and polymers in which the constituent chains can slip past each other in an unrestricted way. The two categories of most importance are layer structures and soft inorganic compounds.

The most familiar solid lubricant is the layered material graphite. The carbon sheets are about 0.335 nm apart, and are linked by weak van der Waals bonds giving a large inter-sheet spacing called the *van der Waals gap* (Figure 8.34a). However, dry graphite is not a perfect lubricant, and much of the lubricating action seems to stem from adsorbed water vapour on and between the layers. This prevents pure graphite from being used for vacuum applications. The lubrication properties are greatly enhanced by forcing the layers apart by inserting atoms into the van der Waals gap. Of these materials, fluorinated graphite, CF_x, in which x can take values between 0.3 and 1.1, is one of the most successful. This material, originally developed for use in lithium batteries, has a structure in which fluorine atoms covalently bond to the carbon atoms. This destroys the sp^2-bonded skeleton of the planar carbon layers, which become puckered as in diamond (Figure 8.34b). However, the increase in spacing between the layers, from 0.335 nm in graphite to 0.68 nm in CF_x weakens the interlayer interaction so much that lubricating properties are greatly enhanced.

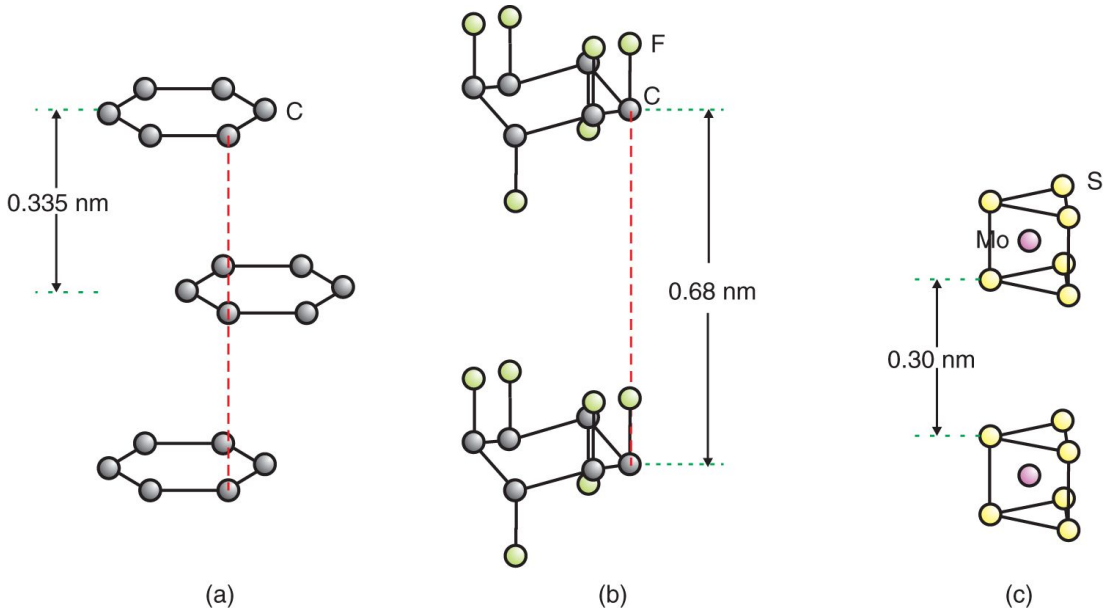


Figure 8.34 The structures of solid lubricants: (a) graphite; (b) graphite fluoride, CF_x; (c) molybdenum disulfide, MoS_2.

Molybdenum disulfide, MoS_2, is another layer structure widely used as a solid lubricant (Figure 8.34c). It is composed of MoS_2 layers weakly linked by van der Waals bonds. The Mo atoms are surrounded by six sulfur atoms in the form of a trigonal prism. Although the MoS_2 layers are closer together than the carbon layers in graphite, MoS_2 acts as a much better lubricant for most purposes because of the weakness of the bonds and the ease with which the strongly bound MoS_2 sheets can slide over each other.

Both graphite and molybdenum disulfide oxidise at high temperatures in air, and the main alternatives used are soft inorganic fluorides. Like many ceramics, although they show brittleness at room temperature, they display ductility at temperatures above the Tamman temperature (about half the melting point).

8.5.2 Auxetic Materials

Auxetic materials expand laterally when subjected to a tensile strain. That is, unlike elastic, which gets thinner when pulled, auxetic substances get fatter when pulled. They have a negative Poisson's ratio. This counter-intuitive mechanical property was first noticed in foam-like structures (Figure 8.35). In these, the links or bonds between elements of structure are of fixed length, but the joints where the bonds attach to elements are flexible. Under an applied force the network will open by pivoting at the flexible joints to cause the material to expand both parallel and perpendicular to the direction of the force. Since then the same effect has been found in non-foam-like materials.

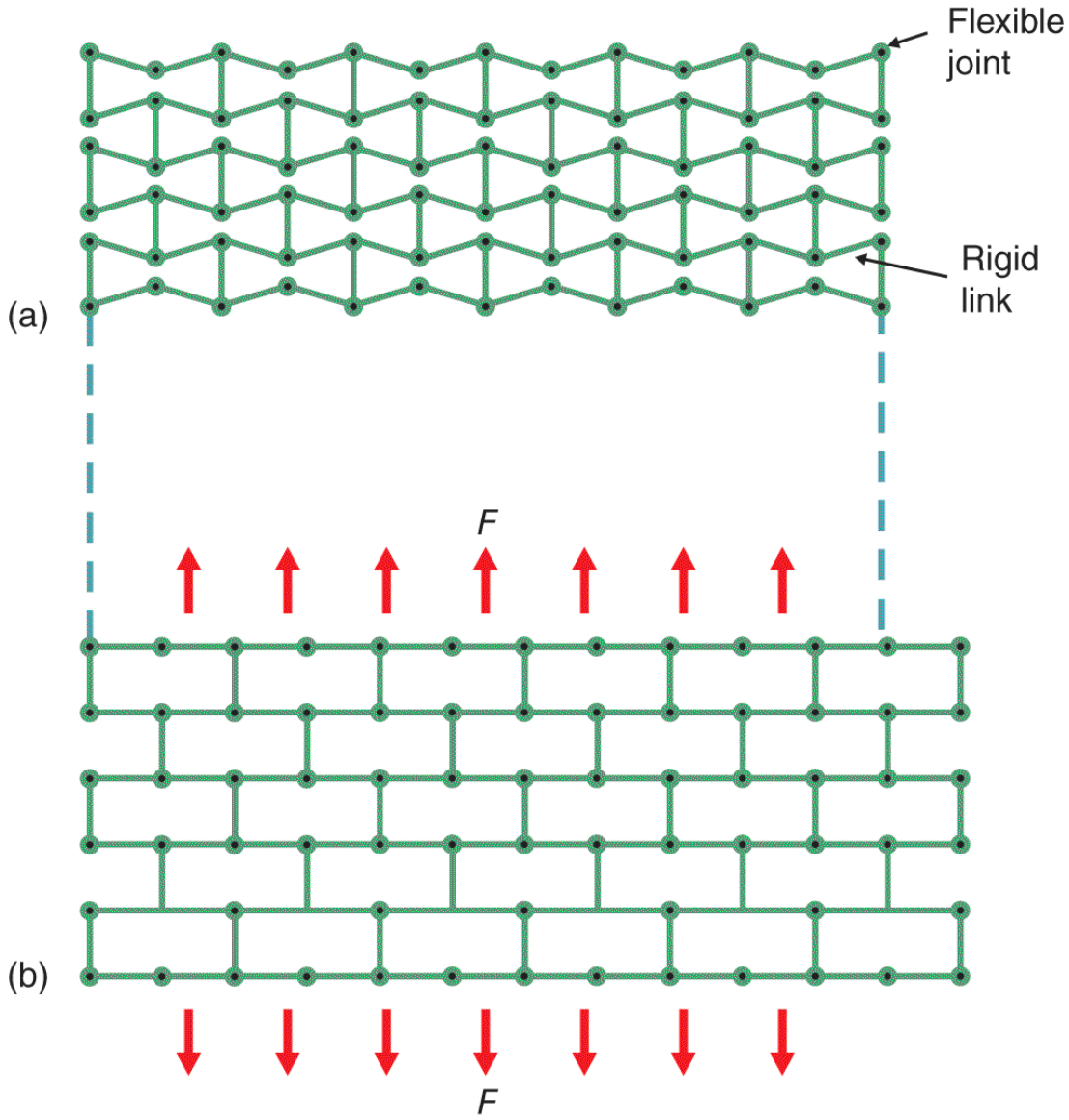


Figure 8.35 An auxetic network in which the links between elements of the structure are of fixed length but the joints are flexible: (a) initial state; (b) under an applied force *F*.

One method of producing auxetic structures is to utilise the microstructures of semicrystalline polymers. In these materials the crystallites form rigid blocks and disordered polymer chains form bonds between them. In the usual geometry of such solids (Figure 8.36a), the material becomes thinner when stretched (Figure 8.36b). However, if the bonds between the blocks have a re-entrant geometry (Figure 8.36c), a tensile force will cause the blocks to move apart and the material will become fatter (Figure 8.36d). This type of structure has been made in highly crystalline polyethylene (UHMWPE), and polypropylene fibres. The transformation from a normal polymer to an auxetic solid is achieved by making the crystalline portion the major component and

reducing the polymer chains linking the crystallites to such an extent that they become short non-coiled links that act as the rigid bonds depicted.

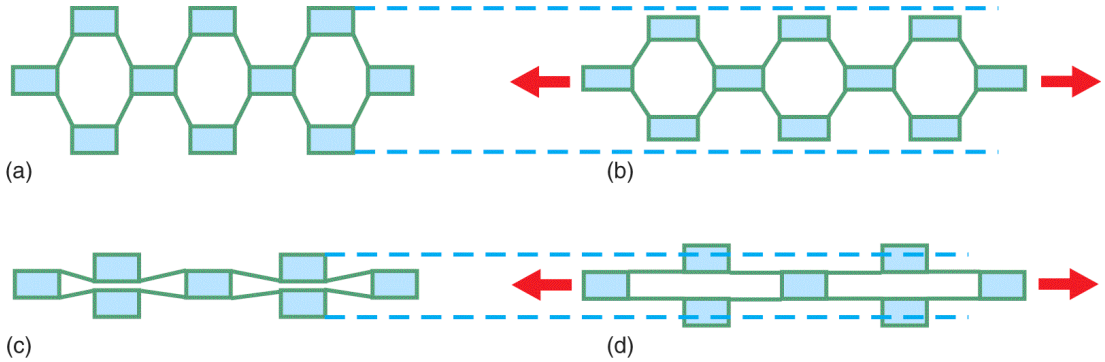


Figure 8.36 A solid composed of rigid blocks of crystal linked by bonds of constant length with flexible joints: (a, b) a normal solid becomes thinner under tension; (c, d) a material with re-entrant bonds is auxetic and becomes thicker under tension.

Surprisingly, many cubic metals behave in a similar fashion, although the effect has been masked by the fact that the mechanical properties are most often measured on polycrystalline solids. For example, the commonplace metallic alloy β-brass, CuZn, which has the CsCl structure, is noticeably auxetic. A tensile stress applied along [001] will result in a (normal) contraction along the lateral direction, perpendicular to (100) and (010) planes. Poisson's ratio has a value of +0.39. However, a tensile stress applied along [111] will result in an expansion in the lateral direction, with a Poisson's ratio of −0.39.

This behaviour is due to a specific combination of bonding and structure. The main bonds in β-brass are between the copper and zinc atoms and are directed along ⟨111⟩, towards the corners of the unit cell from the central atom. Weaker bonds lying along the cell edges link the copper atoms (Figure 8.37a). In mechanical terms, it is as if the central atom is linked to the atoms at the unit cell corners by strong springs and the atoms at the unit cell corners are linked to each other by weaker springs aligned along the unit cell edges. Tension along [001] will have a minimal effect on the strong bonds, and be taken up by the weaker bonds (Figure 8.37b). In this case, the ⟨111⟩ strong bonds are not stretched or compressed, but simply bend a little. As a result, the atoms at the cell corners move in slightly and the unit cell contracts perpendicular to the (100) and (010) faces, similar to a slight folding of a four-spoke umbrella (Figure 8.37c).

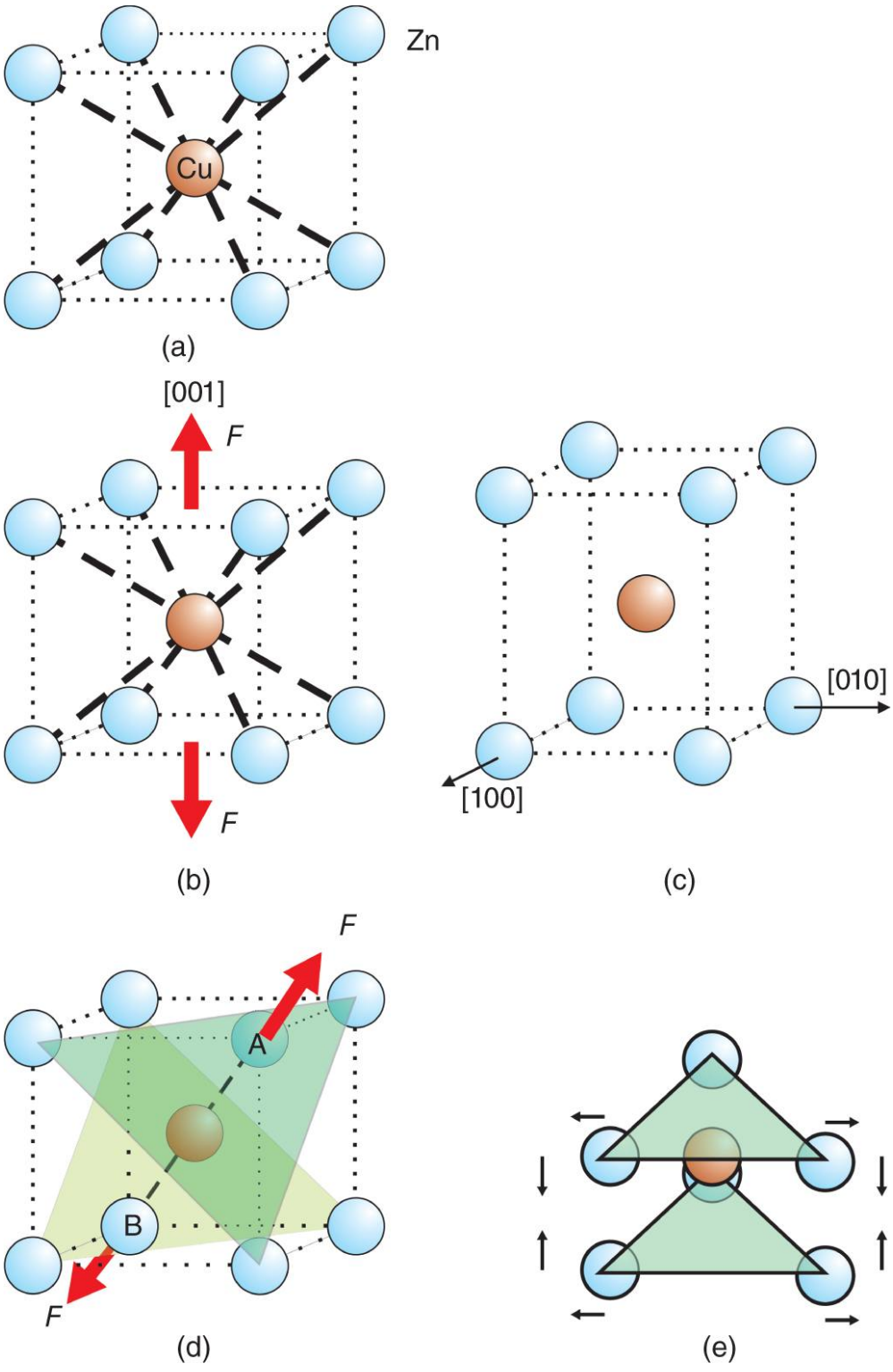

Zn
Cu
(a)
[001]
F
F
(b)
[010]
[100]
(c)
F
A
B
F
(d)
(e)

Figure 8.37 Auxetic β-brass: (a) the CsCl structure of β-brass; strong bonds, heavy dashed lines, weak bonds, light dotted lines; (b, c) application of a force along [001] results in contraction along the [100] and [010] directions; (d) the [111] planes in β-brass: a force along [111] acts directly along strong bonds; (e) the triangles of Zn atoms to come together slightly, making the structure expand perpendicular to [111].

The arrangement of the atoms perpendicular to (111) is rather different (Figure 8.37d). Between the two apical zinc atoms, A and B, are two triangles of zinc atoms, with the copper atom sandwiched in the centre. Tension along [111] is along AB, and hence directly along a strong bond and strongly resisted. As the atoms A and B move apart, the triangles of zinc atoms move slightly closer together, rather like a three-spoke umbrella opening slightly. This causes an expansion of the structure in the direction perpendicular to AB (Figure 8.37e).

It has been found that many body-centred cubic metals are auxetic. A number of silicates also show this unusual property. In each case, the relationship between the bonding and structure controls the response to the tensile stress.

8.5.3 Thin Films and Nanowires

Films with a thickness of the order of nanometres are at the heart of microelectronics and it has become apparent that the mechanical properties of these films must be known in order to ensure complete control over device manufacture and operation. There are a number of microscopic methods available for the determination of the mechanical properties of such films including indentation, beam deflection, and disc deflection. The technique most widely used is the small-scale equivalent to the hardness test, called *nanoindentation*. It has the advantage that the properties can be measured while the film is attached to a substrate and the surface can be tested in a large number of different places. It is also useful for measuring the properties of surfaces that have been modified by, for example, laser irradiation or optical coatings. In a nanoindentation test both the load applied to the indenter and the resultant displacement are measured as a function of time. In this sense, the test mirrors the conventional tensile or compression test. The parameters obtained are Young's modulus, Poisson's ratio, and the hardness of the film. However, there are considerable differences between nanoindentation and conventional tests. For example, there is no clearly distinguished elastic region, and the deformation produced during nanoindentation involves both elastic and plastic deformation. Additionally, the area over which the load is applied changes continually during indentation, and because of the small scale of the measurements, the interaction between the indenter, the film, and the substrate cannot be ignored.

The indenter geometry is different from that employed in large-scale testing, and in general a diamond indenter with a triangular pyramidal shape, called a *Berkovich indenter*, is used (Figure 8.38a). The loads used can be as small as 0.01 μN and displacements as small as 0.1 nm can be measured. The displacement of the indenter tip follows a different path on unloading as compared to loading, and considerable hysteresis is found. This means that interpretation of the results is less obvious than for large-scale experiments.

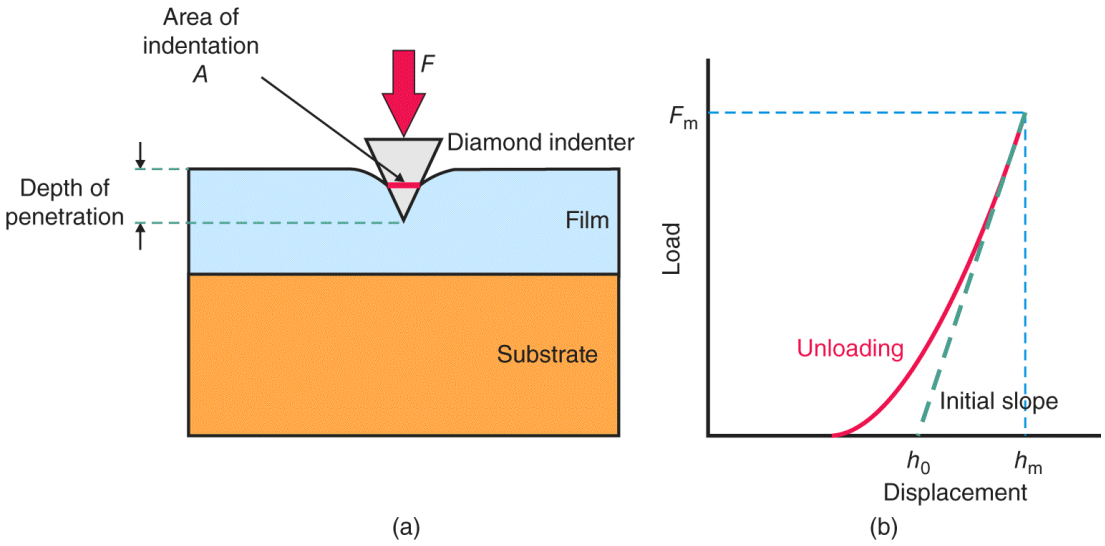


Figure 8.38 Nanoindentation: (a) indentation of a thin film; (b) schematic unloading curve for a nano-indentation test.

A representative load–displacement curve for unloading shows the maximum load applied, F_m, and the corresponding maximum displacement, h_m (Figure 8.38b). The initial slope of the unloading curve, dF/dh, is a measure of the initial stiffness, S_i, and is given by:

$$S_i = dF/dh = F_m/(h_m - h_0)$$

where h_0 is the extrapolation of the initial slope to $F = 0$. If the indentation does not exceed approximately 20% of the film thickness, the substrate does not need to be taken into account. In this case, the hardness of the film is given by:

$$H = F_m/A$$

where A is the area of the indentation. The value of the contact area, A, is a function of the depth of penetration, and must often be determined experimentally. For an ideally sharp indenter:

$$A = 24.5{h_0}^2$$

The slope is related to Young's modulus of the indenter, the film, and the substrate, and when the substrate is ignored the initial slope is given by:

$$S_\mathrm{i} = (2/\sqrt{\pi})\, E_\mathrm{r} \sqrt{A}$$

where A is the area of indentation at maximum load, F_m, and E_r is the *reduced Young's modulus*, given by:

$$1/E_\mathrm{r} = [(1 - n_\mathrm{fi}^2)/E_\mathrm{fi}] - [(1 - n_\mathrm{in}^2)/E_\mathrm{in}]$$

where the subscripts *in* and *fi* refer to indenter and film, respectively, E is the relevant Young's modulus, and ν is the relevant value of Poisson's ratio.

Nanowires and nanotubes are of potential importance in many applications. Gold nanowires, for example, have potential as use in interconnects in nanoscale electronics because of the inert nature of the metal coupled with excellent electrical conductivity. However, mechanical properties must be assessed for these objectives to be achieved. In recent years, the ability to clamp such microscopic units has allowed tensile testing to be carried out. For example, gold nanowires less than 20 nm in diameter have been tested in this way and show curious behaviour. The normal ductile fracture mode, in which the nanowire stretches, then necks and then fractures is frequently observed. In this case, dislocations nucleate at the surface and allow ductile deformation because pure gold is a well-known soft and ductile metal. However, under some circumstances, nanowires exhibit brittle fracture, and show no necking or thinning. In these instances (111) twins in the gold are responsible for brittle fracture.

8.6 Composite Materials

Composites are solids made of more than one material type, designed to have enhanced properties compared with the separate materials themselves. Such solids are commonplace in biological systems, where strength, lightness, and flexibility are the dominant criteria. Thus, wood, shells, bones, claws, feathers, and skin are all composites. These are usually composed of two phases with different and contrasting mechanical properties, a stiff or strong phase and a weak or soft phase, organised in alternating substructures. Thus, wood is composed of cellulose fibrils within a matrix of other cells, and many shells are composed of sheets or prisms of the mineral calcite ($CaCO_3$) alternating with lamellae of the softer material keratin. The mechanical properties of composite materials are often difficult to assess, not only because of the complex microstructures found, but also, and especially in biological structures, because the mechanical properties may be dependent upon the state of hydration/dehydration of the sample.

8.6.1 Elastic Modulus of Fibre Reinforced Composites

The objective of fibre reinforcement is to endow a lightweight matrix with high strength and stiffness. A critical fibre length, l_c, is necessary to achieve this:

$$l_c = \sigma_f \, d/2\tau_c$$

where d is the fibre diameter, σ_f is the (ultimate) tensile strength of the fibre, and τ_c is the shear yield strength of the matrix. The fibre length relative to the critical fibre length gives rise to two terminologies.

If $l > 15l_c$ the material is termed a *continuous fibre* composite.

If $l < 15l_c$ the material is termed a *discontinuous fibre* composite.

In practice, the fibre orientation, concentration, and distribution are all important in the final properties of the composite. Moreover, the mechanical properties of the composite depend upon whether the load is applied along the fibre direction or normal to it.

To obtain a value of Young's modulus of a composite in which the fibres are both continuous and aligned parallel to each other, with the applied force parallel to the fibre orientation, it is simplest to assume that the deformation of the fibres and the matrix is the same, that is, they are firmly bonded together (Figure 8.39a). This is termed the *iso-strain condition*, in which case:

$$\varepsilon_c = \varepsilon_m = \varepsilon_f$$

where ε_c is the strain experienced by the composite as a whole, ε_m is the strain experienced by the matrix, and ε_f is the strain experienced by the fibre. In this case it is reasonable to suppose that the tensile stress, σ, is divided into parts acting on the fibre and on the matrix in proportion to the volume fraction of each. With these assumptions:

$$E_{cl} = E_m V_m + E_f V_f$$

where E_{cl} is Young's modulus of the composite under a longitudinal load, E_m is Young's modulus of the matrix, E_f is Young's modulus of the fibre, and V_f and V_m are the volume fractions (V_f is the volume of fibre/total volume, V_m is the volume of matrix/total volume). For two components:

$$V_f = (1 - V_m)$$

$$E_{cl} = E_m(1 - V_f) + E_f V_f$$

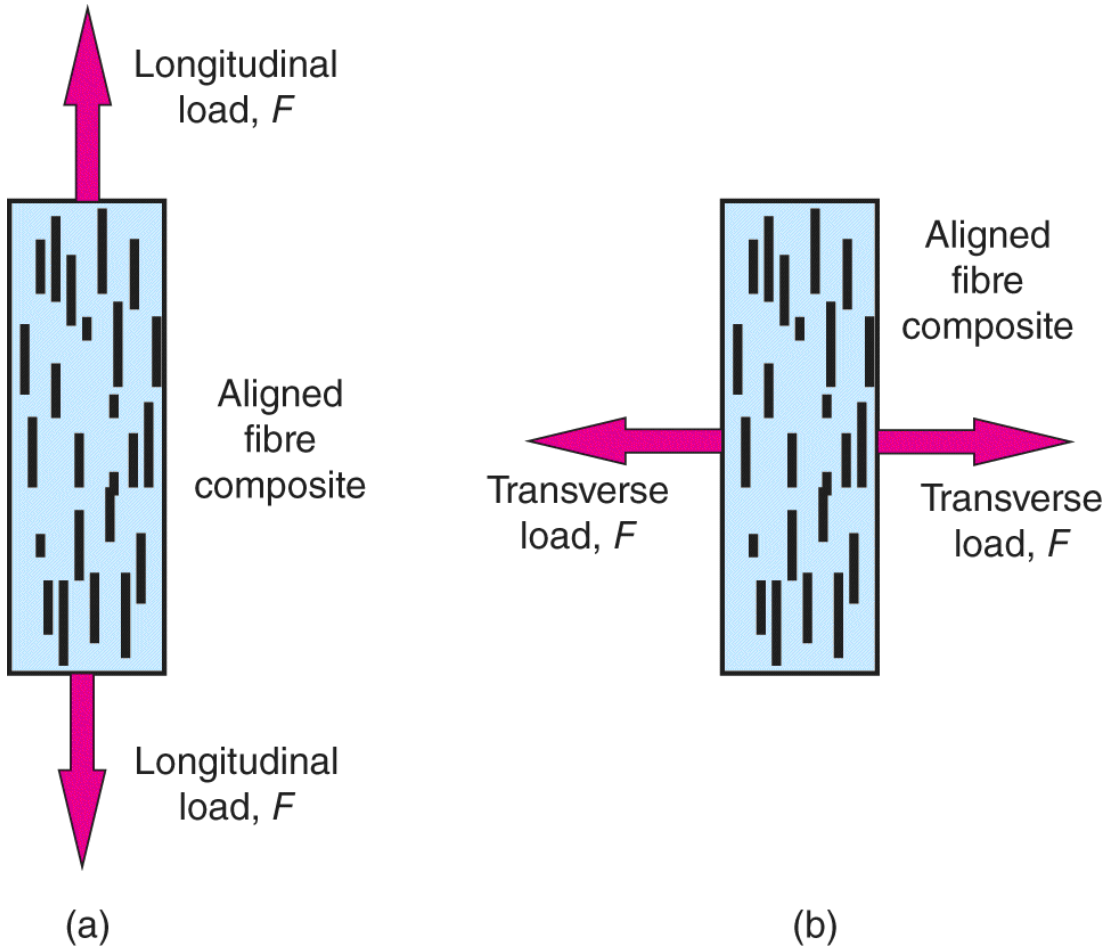


Figure 8.39 Aligned fibre composites: (a) longitudinal load; (b) transverse load.

This is the *upper bound* or *Voigt* bound or limit. The fraction of the load carried by each component is given by:

$$F_f/F_m = E_f V_f/E_m V_m$$

where F_f is the load carried by fibre and F_m is the load carried by matrix.

When a load is applied in a transverse direction to a composite that contains continuous and aligned fibres (Figure 8.39b), the iso-strain condition is unreasonable, and it is more appropriate to assume an *iso-stress condition* to apply:

$$\sigma_c = \sigma_m = \sigma_f = \sigma$$

where σ represents the stress and the subscripts have the same meaning as above. In this case:

$$E_{ct} = E_m E_f / (E_f V_m + E_m V_f)$$
$$= E_m E_f / (1 - V_f) E_f + E_m V_f$$

where E_{ct} is Young's modulus in a transverse direction for the composite.

This is the *lower bound* or *Reuss* bound or limit.

8.6.2 Elastic Modulus of a Two-Phase System

Many ordinary solids, such as ceramics and many biological solids, are made up of several phases. In reality, the equations that describe the elastic properties of a solid composed of several phases with a complex microstructure are frequently unwieldy and simpler approximations are useful for a number of well-defined geometries. To this end, the equations applicable to the continuous fibre-reinforced composite are taken over.

For a body composed of two phases, one of which is distributed as particles within the matrix of the other, Young's modulus is then taken to fall between the two limits described above, an upper (Voigt) and a lower (Reuss) limit.

$$E_c \text{ (upper limit)} = E_m V_m + E_p V_p$$

$$E_c \text{ (lower limit)} = E_m E_p / (E_m V_p + E_p V_m)$$

where E_c, E_m, and E_p are the Young's moduli of the composite, matrix, and particles, respectively, and V_m and V_p are the volume fractions (V_m is the volume of matrix/total volume and V_p is the volume of particles/total volume). For two components:

$$V_p = (1 - V_m)$$

$$E_c \text{ (upper limit)} = E_m (1 - V_p) + E_p V_p$$

The cemented carbides, typically tungsten carbide particles embedded in a matrix of cobalt metal, are large particle composites. These materials are used as cutting tools used to machine hard steels. The hard carbide cuts the steel, but is brittle. The toughness comes from the cobalt matrix. The large particle composite in greatest use is concrete.

A body composed of layers, A and B, aligned parallel to a uniaxial stress (Figure 8.40a) in which the strain is shared equally by the two phases, the iso-strain condition, has a Young's modulus:

$$E_c \text{ (upper limit)} = E_A V_A + E_B V_B$$

where E_A is Young's modulus of phase A and V_A the volume fraction of A and E_B is the Young's modulus of phase B and V_B the volume fraction of A. This gives a value for Young's modulus when the load is applied parallel to the sheets. Similarly, if the stress is applied perpendicular to the layers (Figure 8.40b), and an iso-stress condition applies, the Young's modulus is:

$$E_c \text{ (lower limit)} = E_A E_B / (E_A V_B + E_B V_A)$$

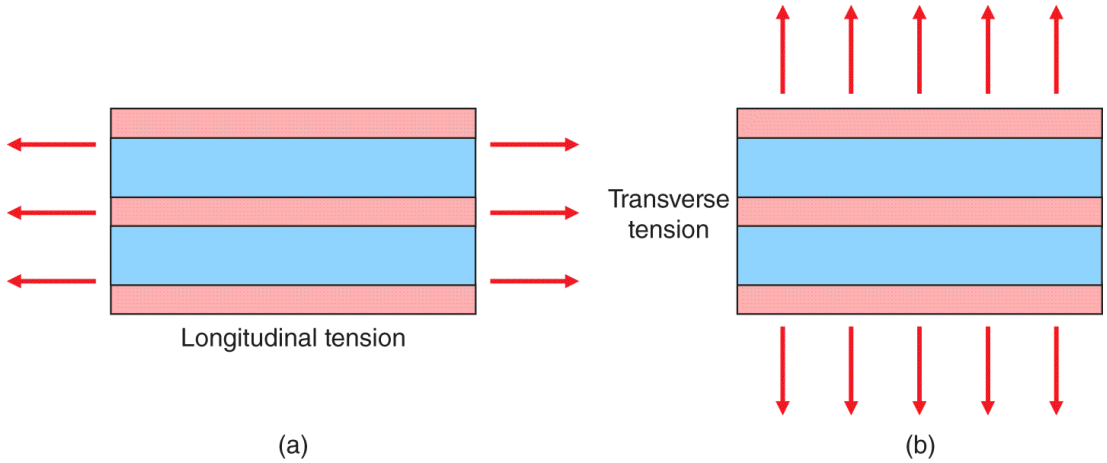


Figure 8.40 Layered materials: (a) longitudinal (parallel) tension; (b) transverse tension.

This gives a value of Young's modulus normal to the layers.

For situations when the stress is applied at an angle to the layers, Young's modulus of the material will lie between the upper, Voigt, value and the lower, Reuss value. In general, the relationship is not a simple one, and is given by a complex function of the angle between the direction of the applied load and the direction of the lamellae in the material (Figure 8.41).

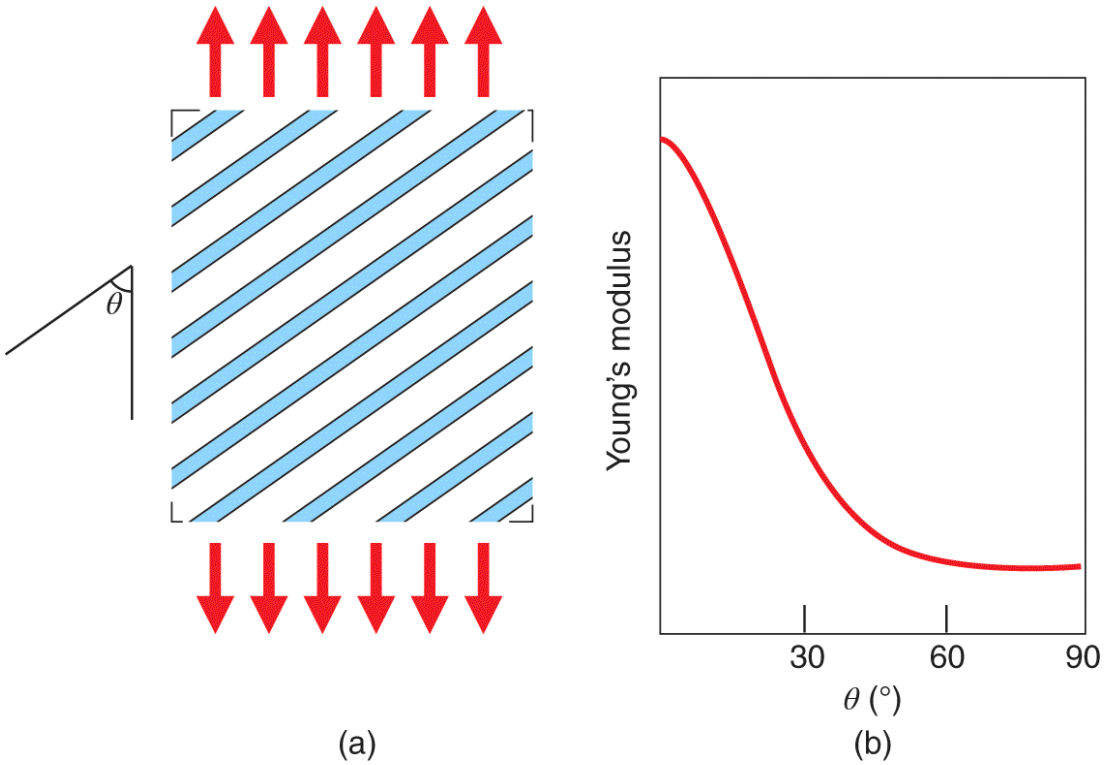


Figure 8.41 Young's modulus of a composite, schematic: (a) configuration of layers; (b) variation of Young's modulus with relative orientation of the load.

The presence of pores in ceramics usually leads to weakness. Young's modulus of a body with a Poisson's ratio of 0.3, containing isolated closed pores, is described by the equation:

$$E_c = E_0 [1 - 1.9V_p + 0.9V_p{}^2]$$

where E_c is the modulus of the porous ceramic, E_0 is the modulus of the non-porous ceramic, and V_p is the volume fraction of pores.

Further Reading

Ductility and Fracture:

Anderson, T.L. (2005). *Fracture Mechanics, Fundamentals and Applications*, 3e. Boca Raton, FL: Taylor & Francis.

Bryant, L.C. and Bewlay, B.P. (1995). The Coolidge process for making ductile tungsten. *Mater. Res. Soc. Bull.* **XX**: 67.

Chen, Y.-S., Lu, H., Liang, J. et al. (2020). Observation of hydrogen trapping at dislocations, grain boundaries and precipitates. *Science***367**: 171.

Eberhart, M.E. (1999). Why things break. *Sci. Am.***281**: 44.

Freiman, S. and Mecholsky, J.J. Jr. (2012). *Fracture of Brittle Materials: Testing and Analysis*. Hoboken, NJ: Wiley.

Griffith, A.A. (1920). The phenomena of rupture and flow in solids. *Philos. Trans. R. Soc. Lond.***A221**: 163–198. [This is the original paper describing the Griffith's theory of brittle fracture.].

Liu, Q., Wang, Z., Ell, J. et al. (2020). Making ultrastrong steel tough by grain-boundary delamination. *Science***368**: 1347.

Sun, C.T. and Lin, Z.-H. (2012). *Fracture Mechanics*. Oxford: Elsevier Academic Press.

Mechanical Properties of Biological Materials:

Liu, Z., Zhang, Z., and Ritchie, R.O. (2020). Structural orientation and anisoptopy in biological materials: functional design and mechanisms. *Adv. Funct. Mater.***30**: 1908121.

Hall–Petch Effect:

Li, Y., Bushby, A.J., and Dunstan, D.J. (2016). The Hall-Petch effect as a manifestation of the general size effect. *Proc. R. Soc., A***472**: 20150890.

Zhou, X., Feng, Z., Zhu, L. et al. (2020). High-pressure strengthening in ultrafine-grained metals. *Nature***579**: 67.

Computation of Properties:

Buehler, M.J. (2008). *Atomistic Modelling of Materials Failure: Deformation and Fracture of Brittle, Ductile and Nanoscale Materials*. New York: Springer Science+Business Media.

Buehler, M.J. and Gao, H. (2006). Dynamical fracture instabilities due to local hyperelasticity at crack tips. *Nature***439**: 307–310.

Holland, D. and Marder, M. (1998). Ideal brittle fracture of silicon studied with molecular dynamics. *Phys. Rev. Lett.***80**: 746–749.

Lesar, R. (2013). *Computational Materials Science*. Cambridge: Cambridge University Press.

Finite Element Methods:

Zienkiewicz, O.C., Taylor, R.L., and Zhou, J.Z. (2005). *The Finite Element Method: Its Basis and Fundamentals*, 6e. Amsterdam: Elsevier Butterworth Heinemann.

Nanoscale Methods:

Lu, Y., Song, J., Huang, J.Y., and Lou, J. (2011). Fracture of sub-20nm ultrathin gold nanowires. *Adv. Funct. Mater.***21**: 3982–3987.

Mitschke, H., Schwerdtfeger, J., Schury, F. et al. (2011). Finding auxetic frameworks in periodic tessellations. *Adv. Mater.***23**: 2669–2674.

Pharr, G.M. and Oliver, W.C. (1992). Measurement of thin film mechanical properties using nanoindentation. *Mater. Res. Soc. Bull.***XV11**: 28.

Xu, B., Arias, F., Brittain, S.T. et al. (1999). Making negative Poisson's ratio microstructures by soft lithography. *Adv. Mater.***11**: 1186–1189.

Composites:

Lu, D. and Barber, A.H. (2012). Optimal nanoscale composite behaviour in limpet teeth. *J. R. Soc. Interface***9**: 1318–1324.

Problems and Exercises

Quick Quiz

8.1 A material under a tensile force is:

a. Stretched.

b. Twisted.

c. Compressed.

8.2 Torsional forces occur in a material that is:

a. Sheared.

b. Twisted.

c. Compressed.

8.3 Both tensile and compressive forces occur in a rod that is:

a. Bent.

b. Twisted.

c. Stretched.

8.4 For a rod-shaped specimen of a metal, the stress is defined as:

a. The change in length per unit length.

b. The change in length per unit force.

c. Force per unit area.

8.5 For a rod-shaped specimen of a metal, the strain is defined as:

a. The change in length per unit length.

b. The change in length per unit force.

c. Force per unit area.

8.6 The engineering strain is given by:

a. Force divided by the original cross-sectional area.

b. The change in length divided by the original length.

c. Force divided by the original length.

8.7 The initial (linear) part of an engineering stress–engineering strain curve represents:

a. Plastic deformation

b. The tensile strength.

c. Elastic deformation.

8.8 In comparison with a strongly cross-linked polymer, a weakly cross-linked form of the polymer will have:

a. A higher Young's modulus.

b. A lower Young's modulus.

c. The same Young's modulus.

8.9 Permanent deformation of a solid when all stress is removed is a sign of:

a. Plastic deformation.

b. Elastic deformation.

c. Tensile deformation.

8.10 The defects mainly held responsible for plastic deformation are:

a. Point defects.

b. Precipitates.

c. Dislocations.

8.11 The tensile strength of a solid is:

a. The maximum load that can be carried before fracture.

b. The maximum load before plastic deformation occurs.

c. The load when fracture occurs.

8.12 The yield stress indicates the point at which:

a. The solid starts to thin down (neck).

b. The elastic behaviour changes to plastic behaviour.

c. The solid starts to fracture.

8.13 The specific strength of a material is:

a. The tensile strength/specific gravity.

b. The tensile strength/volume.

c. The tensile strength/weight.

8.14 An engineering stress–engineering strain curve does NOT give information on:

a. The modulus of elasticity of the material.

b. The yield strength of the material.

c. The hardness of the material.

8.15 The lateral dimensional change that accompanies a longitudinal strain is:

a. Poisson's modulus.

b. Poisson's ratio.

c. Poisson's strain.

8.16 Fracture of a polycrystalline solid that takes place between the crystallites is:

a. Intergranular fracture.

b. Transgranular fracture.

c. Cleavage.

8.17 Conchoidal fracture is displayed by brittle solids that are:

a. Polycrystalline.

b. Amorphous.

c. Natural (biomaterials).

8.18 The Griffith theory of brittle fracture postulates that the fracture is due to:

a. Dislocations.

b. Precipitates.

c. Small cracks.

8.19 Ceramics are more brittle than metals because:

a. They contain fewer dislocations than metals.

b. Dislocation movement is simpler than in metals.

c. Dislocation movement is more difficult than in metals.

8.20 The initial process of plastic deformation in semicrystalline polymers is due to deformation in:

a. The crystalline regions.

b. The amorphous regions.

c. Both the crystalline and amorphous regions together.

8.21 The ductility of a metal can be estimated from:

a. The elongation at fracture.

b. The stress applied at fracture.

c. The strain at the yield point.

8.22 Ceramics can be strengthened by:

a. Immobilising dislocations.

b. Removing Griffith flaws.

c. Increasing the degree of crystallinity.

8.23 One of the following does NOT describe hardness:

a. Knoop.

b. Brinell.

c. Poise.

8.24 Materials that fail after repeated cycles of stress are said to suffer:

a. Fatigue failure.

b. Creep failure.

c. Wear failure.

8.25 Fatigue failure does not occur at stress levels below:

a. The persistence limit.

b. The fatigue limit.

c. The endurance limit.

8.26 The gradual elongation of a material under a constant load is called:

a. Fatigue.

b. Creep.

c. Yield.

8.27 Steady state creep is defined by:

a. The initial part of a creep curve.

b. The middle part of the creep curve.

c. The final part of the creep curve.

8.28 Both Coble creep and Herring–Nabarro creep describe:

a. Creep due to twinning.

b. Creep due to dislocation movement.

c. Creep due to atomic diffusion.

8.29 The temperature dependence of creep is often described by:

a. An Arrhenius law.

b. A linear law.

c. A parabolic rate law.

8.30 Solid lubricants often have:

a. Layer structures.

b. Amorphous structures.

c. Liquid crystal structures.

8.31 Materials that expand when under tension are called:

a. Nanomaterials.

b. Eutectic materials.

c. Auxetic materials.

8.32 The hardness of nanothickness films is measured using:

a. A Vickers indenter.

b. A Knoop indenter.

c. A Berkovich indenter.

8.33 Cemented carbides that are used in cutting tools are:

a. Longitudinally reinforced fibre composites.

b. Large particle composites.

c. Ceramic composites.

Calculations and Questions

8.1 A weight of 500 kg is hung from a 2 cm diameter rod of brass. What is the engineering stress?

8.2 A weight of 3500 kg is hung from a 1.5 cm diameter rod of nickel. What is the engineering stress?

8.3 A steel wire 75 cm long and 1 mm diameter is subjected to a load of 22 kN. Young's modulus of the steel 201.9 GPa. Calculate the new length.

8.4 A rod of bronze 150 cm long and 3 mm diameter is subjected to a load of 30 kN. Young's modulus of the bronze is 105.3 GPa. Calculate the new length.

8.5 A rod of copper 60 cm long is subjected to a tensile stress of 300 MPa. Young's modulus of copper is 129.8 GPa. Calculate the new length.

8.6 A rod of aluminium 100 cm long is subjected to a tensile stress of 250 MPa. Young's modulus of aluminium is 70.3 GPa. Calculate the new

length.

8.7 A cast iron rod of length 200 mm and dimensions 10 mm × 20 mm is subjected to a load of 70 kN. An extension of 0.46 mm is observed. Calculate Young's modulus of the cast iron.

8.8 A zinc bar of length 125 mm and dimensions 5 mm × 7.5 mm is subjected to a load of 40 kN. An extension of 1.23 mm is observed. Calculate Young's modulus of zinc.

8.9 Calculate Poisson's ratio for a bar of metal originally 10 mm × 10 mm × 100 mm, which is extended to 101 mm, if there is no change in the overall volume of the sample.

8.10 A copper bar of 10 mm^2 section is subjected to a tensile load that increases its length from 100 to 102 mm. The value of Poisson's ratio for copper is 0.343. Calculate the new dimensions of the bar.

8.11 A brass rod of 12.5 mm diameter is subjected to a tensile load that increases its length from 150 to 151.5 mm. The value of Poisson's ratio for brass is 0.350. Calculate the new diameter.

8.12 A cylindrical titanium rod of diameter 15 mm is subjected to a tensile load applied along the long axis. Young's modulus of the metal is 115.7 GPa and Poisson's ratio is 0.321. Determine the magnitude of the load needed to produce a contraction in diameter of 5×10^{-3} mm if the deformation is elastic.

8.13 A steel rod of diameter 16.2 mm and length 25 cm is subjected to a force of 50 kN in tension along the long axis. Young's modulus is 210 GPa and Poisson's ratio is 0.293. Determine:

a. the amount that the specimen will elongate in the direction of the applied force;

b. the change in diameter of the rod.

8.14 A niobium bar of dimensions 15 mm^2 and of length 300 mm is subjected to a tensile force of 25 kN. Young's modulus of niobium is 104.9 GPa and Poisson's ratio is 0.397. Determine:

a. the engineering stress;

b. the elongation;

c. the engineering strain;

d. the change in the cross section of the bar.

8.15 A tungsten rod of 12.5 mm diameter and of length 350 mm is subjected to a tensile force of 90 kN. Young's modulus of tungsten is 411.0 GPa and Poisson's ratio is 0.280. Determine:

a. the engineering stress;

b. the elongation;

c. the engineering strain;

d. the change in the diameter of the rod.

8.16 A tensile test specimen of magnesium has a gauge length of 5 cm. The metal is subjected to a tensile loading until the gauge markings are 5.63 cm apart. Calculate:

a. the engineering stress;

b. the percentage elongation.

8.17 A tensile test specimen of brass has a gauge length of 5 cm. The metal is subjected to a tensile loading until the gauge markings are 6.05 cm apart. Calculate:

a. the engineering stress;

b. the percentage elongation.

8.18 An aluminium alloy specimen 3 mm diameter with 50 mm gauge length was tested to destruction in a tensile test. The results are given in the following table. The maximum load applied was 810 N and the final length between the gauge marks was 54 mm.

a. Plot an engineering stress vs. engineering strain curve. Determine:

b. Young's modulus of the alloy;

c. the tensile strength;

d. the 0.2% offset yield strength of the alloy;

e. the percentage elongation at fracture.

Load (N)	**Extension (mm)**
1000	0.10
2000	0.20
3000	0.290
4000	0.402
5000	0.504
6000	0.697
7000	0.900
7500	1.297
8000	2.204
7150, fracture	3.200

8.19 A steel specimen 12 mm diameter with 50 mm gauge length was tested to destruction in a tensile test. The results are given in the table. The maximum load applied was 152 kN.

a. Plot an engineering stress vs. engineering strain curve. Determine:

b. Young's modulus of the alloy;

c. the tensile strength;

d. the 0.1% offset yield stress of the alloy;

e. the percentage elongation at fracture.

Load (kN)	**Extension (mm)**
10	0.030
20	0.064
30	0.098
40	0.130
50	0.170
60	0.195
70	0.218
80	0.256
90	0.294
100	0.335
110	0.400
120	0.505
130	0.660
140	0.898
150	1.300
152 (max)	1.500
150	1.700
140	1.960
133, fracture	2.070

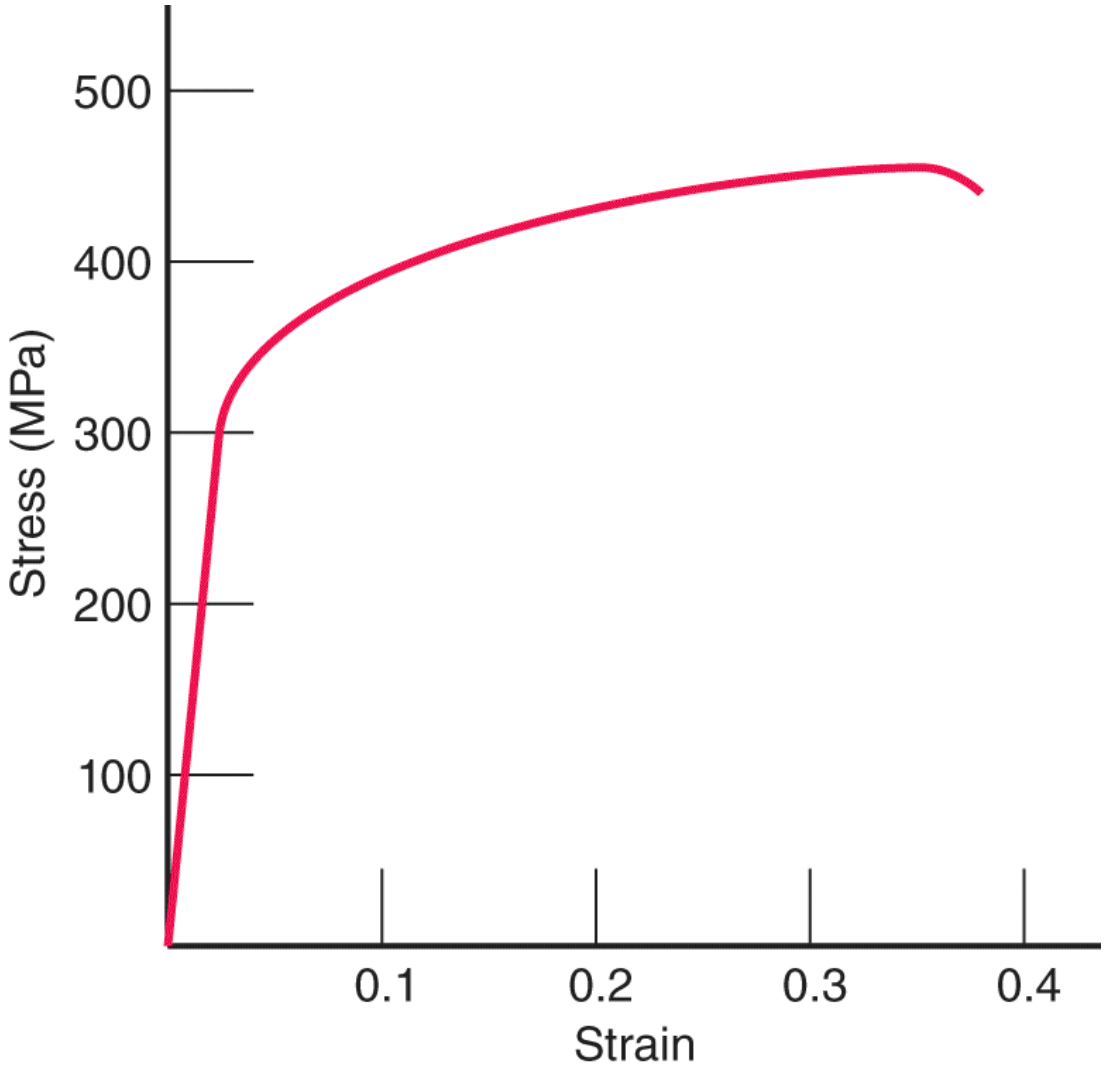
8.20 The figure below shows the engineering stress–engineering strain behaviour of a carbon steel. Determine:

a. Young's modulus.

b. The stress at 0.2% offset strain (proof stress).

c. The maximum load that can be sustained by a rod of diameter 12.5 mm.

d. The change in length of a rod originally 250 mm long subjected to an axial stress of 400 MPa.



8.21 A tensile test carried out on a sample of polypropylene of dimensions 12.5 mm width, 3.5 mm thick, gauge length 50 mm, gave the data in the table. The maximum load applied was 625 N and the length between the gauge marks at fracture was 53.8 mm. Estimate:

a. The initial modulus.

b. The secant modulus at 0.2% strain.

c. The tangent modulus at 0.2% strain.

d. The secant modulus at 0.4% strain.

e. The tangent modulus at 0.4% strain.

f. The percentage elongation at break.

Force (N)	Extension (mm)
25	0.018
50	0.042
75	0.071
100	0.115
125	0.145
150	0.187
175	0.230
200	0.285
225	0.345
250	0.387
275	0.460
300	0.543
286, break	0.720

8.22 A copper–nickel alloy has a 0.1% offset yield strength of 350 MPa and Young's modulus of 130 GPa.

a. Determine the maximum load that may be applied to a specimen of cross section 10 mm × 13 mm, without significant plastic deformation occurring.

b. If the original specimen length is 100 mm, what is the maximum length that it can be stretched to elastically?

8.23 Using the Griffith criterion, $[\sigma_c = (2\,\gamma\,E/\pi\,a)^{1/2}]$, estimate the stress at which a glass plate containing a surface crack of 1.2 μm deep will fracture due to a force applied perpendicular to the length of the crack. Young's modulus of the glass is 71.3 GPa and the surface energy of the glass is 0.360 J m^{-2}.

8.24 A glass plate has to withstand a stress of 10^8 N m^{-2}. Using the data in the Question 8.23, what will be the critical crack size for this to be achieved?

8.25 A plate of high-density polyethylene has a surface crack 7.5 μm in one face. The plate fractures in a brittle fashion when a force of 6×10^6 N m^{-2} is applied in a direction perpendicular to the crack. Young's modulus of the polyethylene is 0.95 GPa. Estimate the surface energy of the material.

8.26 Determine the upper bound and the lower bound Young's modulus of an ingot of magnesium metal containing 30 vol% magnesia (MgO) particles. Young's modulus of magnesium is 44.7 GPa and that of magnesia is 210.3 GPa.

8.27 An aluminium alloy is to be strengthened by the incorporation of beryllium oxide (BeO) particles. Calculate the upper and the lower bound elastic moduli of a composite consisting of 40 wt% alloy and 60 wt% BeO. Young's modulus of the alloy is 70.3 GPa and its density is 2698 kg m^{-3}. Young's modulus of BeO is 301.3 GPa and its density is 3010 kg m^{-3}.

8.28 Compute Young's modulus of a composite consisting of continuous and aligned glass fibres of 50 vol% fraction in an epoxy resin matrix under longitudinal and transverse loading. Young's modulus of the glass fibres is 76 GPa and that of the resin is 3 GPa.

8.29 Compute Young's modulus of a composite consisting of continuous and aligned carbon fibres of 60 wt% fraction in an epoxy resin matrix under longitudinal and transverse loading. Young's modulus of the carbon fibres is 290 GPa and the density is 1785 kg m^{-3}. Young's modulus of the resin is 3.2 GPa and its density is 1350 kg m^{-3}.

8.30 Determine the Voigt and the Reuss bounds to Young's modulus of a ceramic material consisting of layers of alumina and a high silica glass. Young's modulus of the alumina is 380 GPa and that of the glass is 72.4 GPa, and the glass comprises 30 vol% of the solid.

8.31 A mineral with an approximate formula $Mg_7Si_8O_{23}$ has a structure that is made up of alternating layers with compositions of 7MgO and $8SiO_2$. Estimate Young's modulus of the material when stressed parallel and perpendicular to the layers. Young's modulus and density of MgO are 210.3 GPa and 3580 kg m^{-3}, and for silica are 72.4 GPa and 2650 kg m^{-3}.

8.32 Young's modulus of sintered calcia stabilised zirconia with a porosity of 5% is 151.7 GPa. Estimate Young's modulus of completely pore free material.

8.33 Young's modulus of sintered silicon carbide with 5% porosity is 468.9 GPa. What is the porosity of a specimen with a Young's modulus of 350 GPa?

9
Insulating Solids

- What is a dielectric?
- How can quartz be used for telling the time?
- What is a ferroelectric crystal?

Solids have been traditionally divided into three classes when electrical properties are described. Those that conduct electricity well were called *conductors*; a group typified by metals. Solids that conducted poorly, like the element silicon or many minerals, were described as *semiconductors*. Solids that appeared not to conduct electricity were known as *dielectrics* or *insulators*. Many oxides and most polymers fall into this category (Figure 9.1). It has long been known that these divisions refer to the ability of the solid to transport electrons across the phase, so that in metals, electron transport is easy, and in insulators, no passage of electrons is possible.

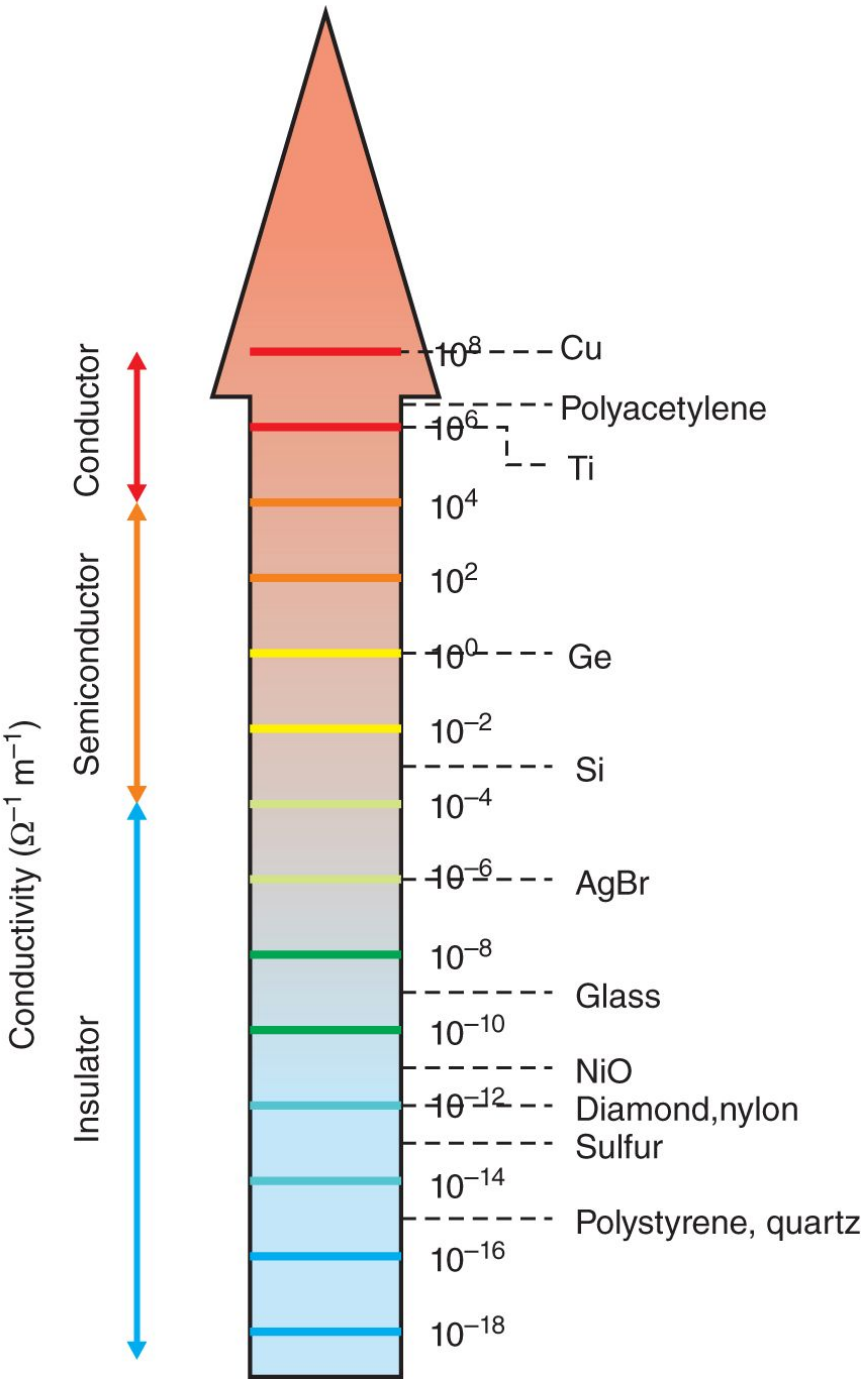


Figure 9.1 The range of electronic conductivity in solids.

This division is far too coarse to encompass the wide range of electrical properties that are now known. It is quite feasible to turn an insulating oxide into a very good conductor and metallic polymers are well known. However, the historic division is useful in broad-brush terms and is retained here. In this chapter the insulators are described. In Chapter 11 those materials that are conductors of electricity are discussed.

9.1 Dielectrics

9.1.1 Relative Permittivity and Polarisation

Insulators are explained in terms of chemical bonding as those solids in which the outer electrons are unable to move through the structure. They are localised in strong bonds if the material is considered to be a covalent compound, or else are restricted to the region close to an atomic nucleus if the compound is supposed to be ionic. In either case, these electrons are trapped and cannot move from one region to another. Insulating materials are often referred to as *dielectrics*. One of the most important parameters used to describe an insulator is its *dielectric constant*, properly called the *relative permittivity*, ε_r.

Dielectrics form the working material in *capacitors*. In its simplest, traditional form, a capacitor consists of two parallel metal plates separated by a dielectric. If we arrange for the two plates to be connected to a battery, a certain amount of charge will accumulate on the plates (Figure 9.2). In a vacuum (or air, in practice), we find

$$q = \varepsilon_0 A V/d$$

where q is the charge on the capacitor, ε_0 is a constant, the *permittivity of free space*, A is the area of the plates, and d their separation. The capacitance, c, of the arrangement is equal to the ratio of the charge on either of the metallic plates, q, to the potential difference, V, between them:

$$c = q/V = \varepsilon_0 A/d$$

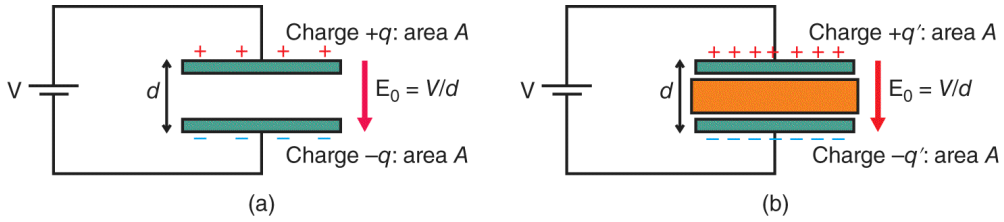


Figure 9.2 Capacitors: (a) charges will accumulate on metal plates due to an applied voltage; (b) a slab of dielectric inserted between the plates will increase the charges held on the plates.

If the region between the plates is now filled with a dielectric, the charge and the capacitance increase by an amount ε_r:

$$q' = \varepsilon_0 \varepsilon_r A V/d$$

$$c' = \varepsilon_0 \varepsilon_r A/d$$

where ε_r is the relative permittivity of the material.

The relative permittivity describes the response of a solid to an *electric field*. The electric field is a vector quantity, **E**, which has a direction pointing from positive to negative. When an insulating material is exposed to an external electric field, $\mathbf{E}_0$,

arising from, for example, two charged metallic plates in a capacitor; the constituents of the insulator, ions, atoms, or molecules, become polarised (Section 9.1.2). The result is the formation of induced *internal electric dipoles* on these components. These dipoles are described by vectors that run from the negative charge to the positive. The electric dipole moment of a pair of charges $\pm q$ is given by a vector **p**, where:

$$\mathbf{p} = q\mathbf{r}$$

and **r** is the vector describing the location of the charges. When direction is unimportant, the dipole moment p is simply given by:

$$p = qr$$

where r is the distance between the charges.

The induced dipoles on the internal constituents add together, with a result that opposite surfaces of the solid become positively and negatively charged and the solid becomes *polarised* (Figure 9.3). The polarisation of the dielectric, **P**, is a vector quantity, defined as the electric dipole moment per unit volume. The polarisation vector points from the negative surface to the positive surface, and the vectors **P** and **p** are parallel.

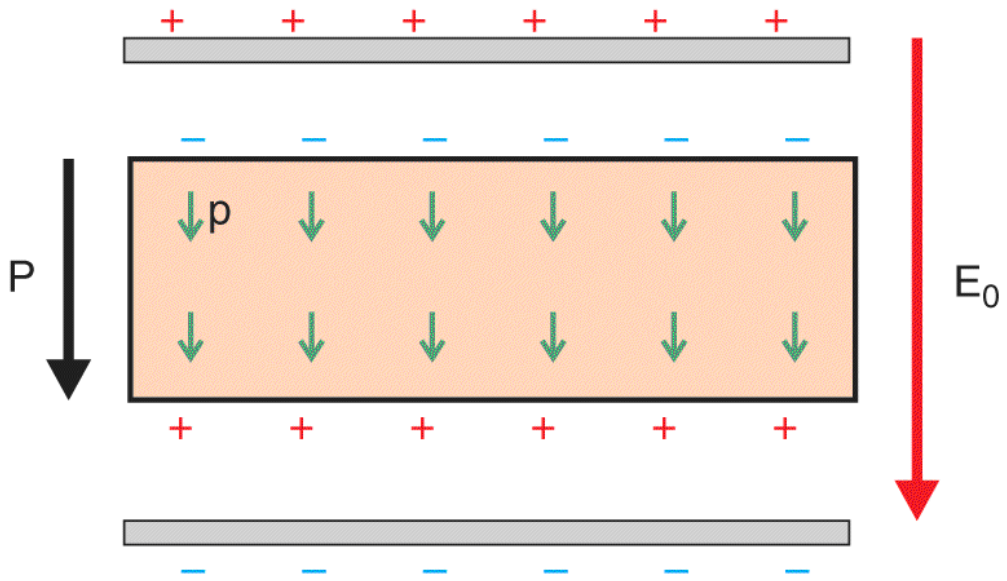


Figure 9.3 In an applied electric field, $\mathbf{E}_0$, an insulator gains a surface charge due to the formation of internal dipoles, **p**, that induce an observable polarisation, **P**.

In the case of a solid that is isotropic (uniform in all directions), for example, glass or a cubic crystal, it is found that at ordinary field strengths **P** is proportional and parallel to the applied electric field, $\mathbf{E}_0$, and:

$$\mathbf{P} = \varepsilon_0 \chi \mathbf{E}_0 \qquad (9.1)$$

where **P** has units C m^{-2}, **E** has units V m^{-1}, and χ is called the *dielectric susceptibility* of the material. (At higher electric field strengths such as those found in laser beams, it

is necessary to replace the right-hand side of the equation by a series, with $\varepsilon_0\chi\mathbf{E}$ as the first term.) In mathematical terms, dielectric susceptibility is the derivative of polarisation with respect to applied electric field. At normal field strengths, the dielectric susceptibility is related to the relative permittivity by the equation:

$$\chi = (\varepsilon_r - 1) \qquad (9.2)$$

$$\mathbf{P} = (\varepsilon_r - 1)\varepsilon_0\mathbf{E_0}$$

For ordinary electric field strengths, the electric dipole moment **p**, induced in a constituent in the solid, is proportional to the polarisability of the constituent, α,[1] and the *local* electric field acting on the constituent, thus:

$$\mathbf{p} = \alpha\,\mathbf{E}_{\text{loc}}$$

Note that the local field, $\mathbf{E}_{\text{loc}}$, is not the same as the applied field, $\mathbf{E}_0$, as it will also include contributions from dipoles present in the structure, some of which may be permanent dipoles, but others will be temporary dipoles, induced by the applied electric field itself.

If there are N identical dipoles per unit volume:

$$\mathbf{P} = N\,\alpha\,\mathbf{E}_{\text{loc}}$$

In principle, all constituents of a solid, including defects and internal surfaces, contribute to the polarisability. Thus, a solid with two polarisable components A and B containing 10 units of A and 15 units of B, of polarisability α_A and α_B, would have a polarisability:

$$\mathbf{P} = 10\mathbf{p}_A + 15\mathbf{p}_B = 10\alpha_A\mathbf{E}_{\text{loc}} + 15\alpha_B\mathbf{E}_{\text{loc}}$$

If there are N_j types of constituents, of polarisability α_j, in a solid, the observed polarisation, **P**, is:

$$\mathbf{P} = \sum_j N_j\mathbf{p}_j = \sum_j N_j\alpha_j\mathbf{E}_{\text{loc}}$$

where the local electric field acting on the constituent, $\mathbf{E}_{\text{loc}}$, may vary from site to site in the crystal.

A universal response to the application of an electric field to a dielectric is a change in the sample dimensions, called *electrostriction*. Some materials get thinner while others get thicker in the direction of the electric field. This effect is not reversible – that is a deformation does not produce any polarisation. The effect is generally very small in crystalline ceramic dielectrics except for *relaxor ferroelectrics* (Section 9.3.8) which may show high deformation. (Note that electrostriction is a property of all dielectrics and is quite different from piezoelectricity, Section 9.2).

9.1.2 Polarisability

The polarisation of an insulating solid is derived from the atomic constituents that make up the material and the defects that may be present. In the absence of an electric field, the electronic charge cloud surrounding an atom (at a little distance from the atom) is symmetrically disposed around the nucleus. In an electric field this charge cloud becomes deformed and the centre of the electronic negative charge is no longer coincident with the positive nuclear charge (Figure 9.4a), and a dipole will arise. This effect is termed *electronic polarisability* α_e. Charged ions in a solid will suffer a displacement in an electric field resulting in *ionic polarisability* α_i (Figure 9.4b). A number of common molecules, including water, carry a permanent dipole. If such molecules are exposed to an electric field, they will try to orient the dipole along the field (Figure 9.4c) leading to *orientational polarisability* α_d. As the movement of molecules in solids is restricted, orientational polarisability is more often noticed in gases and liquids. If a material has mobile charges present, electrons, holes, or ions, they will move under the influence of the electric field, with positive charges moving towards one electrode and negative charges towards the other. These will tend to build up at grain boundaries within the solid and at the electrode surfaces until the resulting charge inhibits further movement and equilibrium is reached, producing *space charge polarisability* α_s. Good ionic conductors often show pronounced space charge effects. Other defects such as vacancies can also make a significant contribution to the observed polarisation. The observed total polarisability α_t will arise from the sum of all of the separate terms, written as:

$$\alpha_t = \alpha_e + \alpha_i + \alpha_d + \alpha_s + \cdots$$

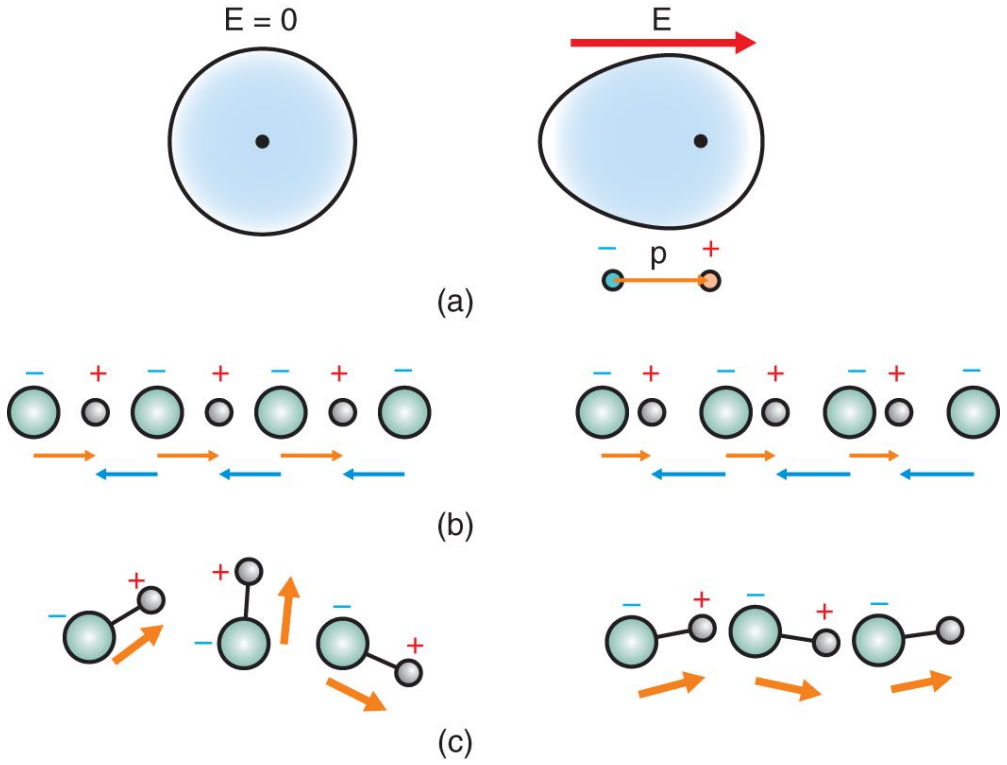


Figure 9.4 Polarisation: (a) electronic; (b) ionic; (c) orientational. Dipoles are shown as arrows.

To relate polarisability to the relative permittivity, it is necessary to remember that each constituent of the solid is polarised by a *local* electric field. This local field, $\mathbf{E}_{loc}$, is not the same as the applied field, $\mathbf{E}_0$, but will also include contributions from internal fields $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, etc. arising from the induced and permanent dipoles in the structure.

$$\mathbf{E}_{loc} = \mathbf{E}_0 + \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 + \cdots$$

Lorentz, using classical electrostatic theory, showed that the local field in an isotropic insulator such as a gas, a glass, or a crystal with cubic symmetry is uniform everywhere and given by:

$$\mathbf{E}_{loc} = \mathbf{E}_0 + \mathbf{P}/3\,\varepsilon_0 \tag{9.3}$$

Using the three Eqs. , it is possible to derive the most widely used relationship between relative permittivity and polarisability, the *Clausius–Mossotti relation*, Eq. (9.4), usually written:

$$(\varepsilon_r - 1)/(\varepsilon_r + 2) = N\alpha/3\,\varepsilon_0 \tag{9.4}$$

where α is the polarisability of the material and N is the number of atoms or formula units of structure per unit volume. If there are j types of polarisable species present, the sum of $N_j\alpha_j$ is needed. Remember that this equation is only applicable to homogeneous isotropic materials that do not contain permanent dipoles or dipolar molecules. However, it is often taken to be approximately true for crystals of lower symmetry, provided that they do not contain permanent dipolar molecules.

Several alternative forms of the Clausius–Mossotti equation are encountered. Frequently the term N is replaced by its reciprocal, the volume of one atom or one formula unit of structure, $V_m = 1/N$, and is set out in terms of α, thus:

$$\alpha = 3\,\varepsilon_0\,V_m\,[(\varepsilon_r - 1)/(\varepsilon_r + 2)]$$

Because the polarisability volume, α', is often used rather than the SI polarisability, α, a commonly encountered form of the Clausius–Mossotti equation is:

$$\alpha' = (3V_m/4\pi)\,[(\varepsilon_r - 1)/(\varepsilon_r + 2)]$$

where V_m is the volume of one formula unit of structure. The units of α' will be the same as those of V_m. Quite often the equation is expressed in terms of the *molar polarisability*, P_m. This form is obtained by multiplying both sides of Eq. (9.4) by M/ρ, where M is the molar mass of the material and ρ is its density, to obtain:

$$(\varepsilon_r - 1)\,M/(\varepsilon_r + 2)\,\rho = N_A\alpha/3\,\varepsilon_0 = P_m$$

where N_A is the Avogadro constant, given by $M\,N/\rho$.

As the total polarisability of a material is made up several contributions, the relative permittivity can also be thought of as made up from the same contributions. In a static electric field, all the various contributions will be important and will arise from

electrons, ions, dipoles, defects, and surfaces. However, if a variable, especially alternating, electric field acts on the solid, the situation changes.

At low enough frequencies the value of the relative permittivity measured will be identical to the static value and all polarisability terms will contribute to ε_r. However, space charge polarisation is usually unable to follow changes in electric field that occur much faster than that of radio frequencies, about 10^6 Hz, and this contribution will no longer be registered at frequencies much higher than this value. Similarly, any dipoles present are usually unable to rotate to and fro in time with the alternations of the electric field when frequencies reach the microwave region, about 10^9 Hz, and at higher frequencies this contribution will be lost. Ionic polarisability, involving the movement of atomic nuclei, is no longer registered when the frequency of the field approaches that of the infrared range, 10^{12} Hz. Electrons, being the lightest components of matter, still respond to an alternating electric field at frequencies corresponding to the visible region, 10^{14} Hz, but even the contribution of electronic polarisability drops out at ultraviolet frequencies (Figure 9.5).

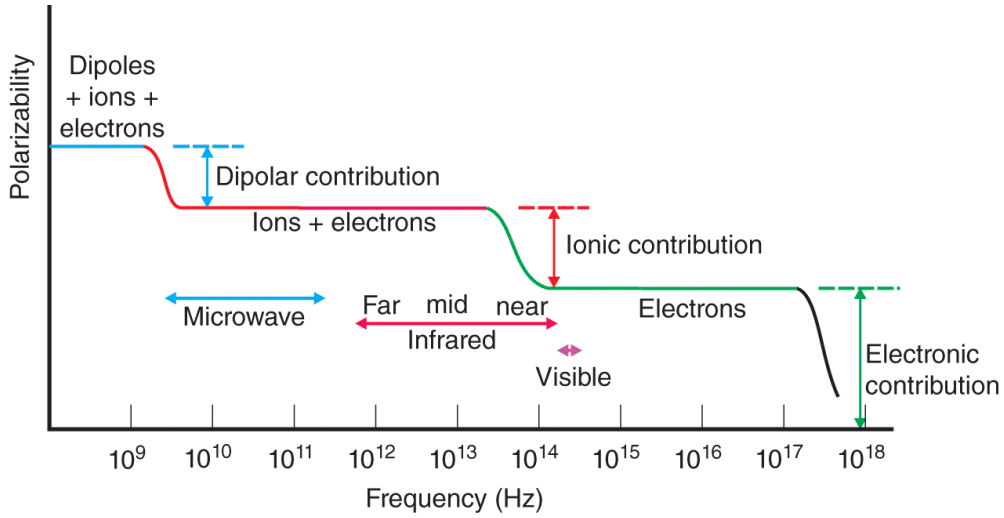


Figure 9.5 The contribution of dipoles, ions and electrons to the polarisability of a solid, schematic.

The to and fro interaction of the components of a solid with an alternating electric field dissipates energy and results in heating of the dielectric. It also causes a lag between the phase of the input field and the phase of the output field. The action of an alternating electric field in best described by using the complex dielectric constant

$$\varepsilon_r = \varepsilon' - i\varepsilon''$$

The loss tangent, tan δ, then specifies the phase lag, where:

$$\tan \delta = \varepsilon''/\varepsilon'$$

The loss tangent is a measure of the energy loss in a capacitor. For good dielectrics, tan δ is about 10^{-4} and is relatively insensitive to the frequency of the applied field.

The interaction of an alternating electric field with a solid in the frequency range between the infrared and ultraviolet, the optical range, is more commonly expressed as the refractive index, because a light wave consists of oscillating electric and magnetic fields. The relative permittivity in the optical region is related to the refractive index, n, by:

$$n^2 = \varepsilon_r \tag{9.5}$$

This relationship as such is not well obeyed if the normal static or low frequency relative permittivity is used when this contains significant contributions from non-electronic polarisability (Table 9.1). With this in mind, substitution of the relationship given in Eq. (9.5) into the Clausius–Mossotti equation yields the *Lorentz–Lorenz* equation:

$$(n^2 - 1)/(n^2 + 2) = N\alpha_e/3\,\varepsilon_o$$

Table 9.1 The relative permittivity and refractive index of some crystals.

Compound	Symmetry	Relative permittivity[a]	Frequency (Hz)	Refractive index (n)	n^2
Diamond	Cubic	5.66	10^3	2.418	5.85
Periclase MgO	Cubic	9.65	10^2–10^8	1.735	3.010
Spinel $MgAl_2O_4$	Cubic	8.6	—	1.719	2.955
Fluorite CaF_2	Cubic	6.81	10^2–10^{11}	1.434	2.056
$CaCu_3Ti_4O_{12}$	Cubic	~5×10^5	—	—	—
Corundum Al_2O_3	Hexagonal	⊥ c, 9.34 || c, 11.54	10^2–10^9	⊥ c, 1.761 || c, 1.769	3.101 3.129
Beryl $Be_3Al_2Si_6O_{18}$	Hexagonal	⊥ c, 6.86 || c, 5.95	10^3	⊥ c, 1.589 || c, 1.582	2.525 2.503
Rutile TiO_2	Tetragonal	|| a, b, 86 || c, 170	10^4–10^6	|| a, b, 2.609 || c, 2.900	6.807 8.410

[a] ⊥ perpendicular to; || parallel to.

9.1.3 The Relative Permittivity of Crystals

In cubic crystals the relative permittivity does not depend upon the direction of the applied electric field. However, in other crystalline solids the relative permittivity varies with direction, subject to the symmetry of the crystal structure. To describe this variation, it is usual to define the electric field vector with respect to a set of orthogonal axes in the crystal. When a cubic crystal is placed in an electric field, **E**, at ordinary field strengths the induced polarisation, **P**, is proportional and *parallel* to the applied electric field vector, and is represented by Eq. (9.1). The directions of **E** and **P** are

usually referred to a set of orthogonal axes, OX, OY, OZ (1, 2, 3) in the crystal, so that the components of the vectors **P** and **E** along OX, OY, and OZ are:

$$\mathbf{P}_1 = \varepsilon_0 \chi \mathbf{E}_1$$

$$\mathbf{P}_2 = \varepsilon_0 \chi \mathbf{E}_2$$

$$\mathbf{P}_3 = \varepsilon_0 \chi \mathbf{E}_3$$

In crystals of lower symmetry than cubic, this set of equations must be replaced by a more general formulation:

$$\begin{aligned} \mathbf{P}_1 &= \varepsilon_0 \chi_{11} \mathbf{E}_1 + \varepsilon_0 \chi_{12} \mathbf{E}_2 + \varepsilon_0 \chi_{13} \mathbf{E}_3 \\ \mathbf{P}_2 &= \varepsilon_0 \chi_{21} \mathbf{E}_1 + \varepsilon_0 \chi_{22} \mathbf{E}_2 + \varepsilon_0 \chi_{23} \mathbf{E}_3 \\ \mathbf{P}_3 &= \varepsilon_0 \chi_{31} \mathbf{E}_1 + \varepsilon_0 \chi_{32} \mathbf{E}_2 + \varepsilon_0 \chi_{33} \mathbf{E}_3 \end{aligned} \tag{9.6}$$

This implies that an electric field lying along OX may induce a polarisation along OX, OY, and OZ. The array of nine coefficients χ_{ij} relating response polarisation to electric field are reduced considerably in higher symmetry crystals, as symmetry constraints make many equal to zero. Thus, cubic crystals are characterised by a single value of χ and relative permittivity, hexagonal and tetragonal crystals, have two values, corresponding to fields along the **a**- and **c**-axes, and orthorhombic crystals have just three values, corresponding to field along the **a**-, **b**-, and **c**-axes (Table 9.1).

The relative permittivity of polycrystalline samples does not reflect this because the crystallographic symmetry is averaged. In these materials though, the measured relative permittivity is often controlled more by defects than by crystal properties. Point defects, mobile charge carriers, and grain boundaries will all have an influence on the measured relative permittivity of the solid. This can be illustrated by the ceramic oxide $CaCu_3Ti_4O_{12}$. The relative permittivity, which can be as high as 2×10^5, depends upon the preparation route, the sintering conditions, and the grain size in the solid. The source of the high value (as yet uncertain) appears to arise in thin fully oxidised insulating grain boundaries enclosing large semiconducting grains containing a high population of mobile charges.

It may sometimes be necessary to estimate the polarisability of a solid in the absence of experimental data. Polarisability is not particularly easy to measure, but the relative permittivity is. The Clausius–Mossotti equation is usually used to obtain polarisability from relative permittivity. The equation gives reasonable values for isotropic solids showing only ionic and electronic polarisation. If the refractive index is known or can be estimated, the Lorentz–Lorenz equation will yield the electronic polarisability of the material. Hence, by difference, the ionic polarisability can be estimated.

In the absence of relative permittivity data for the solid under consideration, it is possible to make use of the *additivity rule*. In its simplest form:

$$\alpha(\text{compound}) = \Sigma\alpha(\text{components})$$

For example, for an oxide mineral:

$$\alpha(\text{mineral}) = \Sigma\alpha(\text{component oxides})$$

$$\alpha(Mg_2SiO_4) = 2\alpha(MgO) + \alpha(SiO_2)$$

A more extended form of the additivity rule is obtained if the 'components' are actual ions or atoms. In the example earlier we would then write:

$$\alpha(Mg_2SiO_4) = 2\alpha(Mg^{2+}) + \alpha(Si^{4+}) + 4\alpha(O^{2-})$$

9.2 Piezoelectrics, Pyroelectrics, and Ferroelectrics

9.2.1 The Piezoelectric and Pyroelectric Effects

In a normal dielectric, the observed polarisation of the material is zero in the absence of an electric field. In a *piezoelectric solid* a surface electric charge develops when the solid is subjected to a mechanical stress such as pressure, even in the absence of an external electric field. This is called the *direct piezoelectric effect*. The effect is reversible and the *inverse* (or *converse*) piezoelectric effect, in which a voltage applied to a crystal causes a change in shape, also occurs in piezoelectric crystals. In mathematical terms piezoelectricity is the derivative of polarisation with respect to strain. The piezoelectric effect generally varies from one direction to another in a crystal and in some directions a crystal may show no piezoelectric effect at all, while in other directions it is pronounced. Piezoelectric solids are a subset of dielectrics. All piezoelectrics are dielectrics, but only some dielectrics are piezoelectrics.

In the case of a *pyroelectric solid* a change of temperature induces a polarisation change. The change obtained on heating is reversed on cooling. In mathematical terms pyroelectricity is the derivative of polarisation with respect to temperature. Pyroelectric crystals are a subset of piezoelectrics. All pyroelectric crystals are piezoelectrics, but not all piezoelectrics demonstrate pyroelectricity. A material that is a pyroelectric is found to possess a *spontaneous polarisation*, $\mathbf{P}_S$. This means that a pyroelectric crystal shows a permanent polarisation that is present both in the absence of an electric field and mechanical stress. Despite this, it is a matter of common observation that a pyroelectric crystal does not usually show an external charge. This is because the surface charges are neutralised by charged particles picked up from the air. Nevertheless, when a pyroelectric crystal is heated or cooled the spontaneous polarisation will change but the collection of neutralising particles will take time to arrive, and a surface charge will be detected. Pyroelectric crystals kept clean and in a vacuum maintain the surface charges for many days.

Ferroelectrics also possess a spontaneous polarisation, $\mathbf{P}_S$, in the absence of an electric field and a mechanical distortion. They are, therefore, a subset of pyroelectrics, and as such, all ferroelectrics are also pyroelectrics and piezoelectrics. The feature that distinguishes ferroelectrics from pyroelectrics is that the direction of the spontaneous polarisation, $\mathbf{P}_S$, can be *switched* (changed) in an applied electric field. Additionally, ferroelectric crystals often possess very high values of relative permittivity, especially over narrow temperature ranges close to crystallographic symmetry transitions.

The hierarchy of insulating properties can be summarised thus (Figure 9.6):

> If polarisation, **P**, changes with applied electric field, **E**, we have a *dielectric*.
>
> In some dielectrics the polarisation, **P**, can arise from mechanical stress, σ, to give a *piezoelectric*.
>
> In some piezoelectrics, there is a spontaneous polarisation, $\mathbf{P}_S$, when the applied electric field, **E**, and the stress, σ, are zero, that changes with temperature, T, to give *pyroelectrics*.

In some pyroelectrics the direction of the spontaneous polarisation, $\mathbf{P}_S$, is easily switched in an electric field, to give a *ferroelectric*.

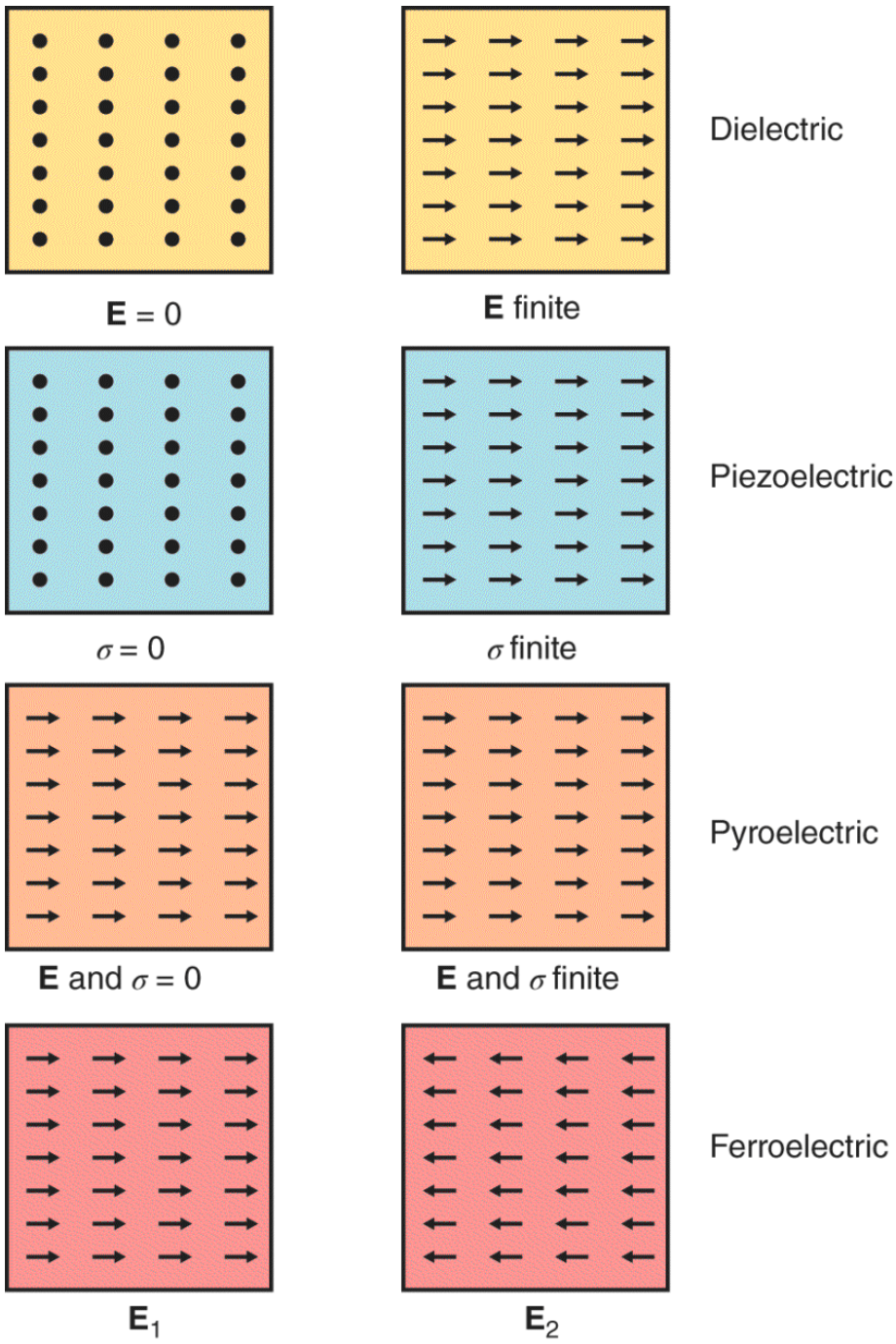


Figure 9.6 Schematic relationship between insulating solids: **E** is an applied electric field and σ is an applied stress.

9.2.2 Crystal Symmetry and the Piezoelectric and Pyroelectric Effects

In the piezoelectric effect the polarisation results from the generation of internal dipoles in the crystal as a result of external stress. For example, the application of a stress can distort the anion coordination polyhedron, frequently a tetrahedron or octahedron, surrounding a cation. In some circumstances this can lead to the cation (+) no longer occupying the centre of gravity of the surrounding anions (−), producing an internal dipole in the polyhedron. Suppose that a dipole arises in a distorted tetrahedron of anions surrounding a central cation. The observed effect will be the sum of the all dipoles in the unit cell. The easiest way to obtain this sum is to use the symmetry of the unit cell, defined by the 32 different crystal classes. If the unit cell in the crystal has a centre of symmetry, the overall polarisation will always add to zero because any dipole **p** pointing in a direction [*uvw*] will be matched by one pointing in the opposite direction. Piezoelectricity can thus only arise in one of the 21 crystal classes that lack a centre of symmetry. In fact, one of these classes is an anomalous cubic group that is unable to support the piezoelectric effect so that piezoelectric materials must belong to one of the 20 (non-cubic) classes that lack of a centre of symmetry (Figure 9.7).

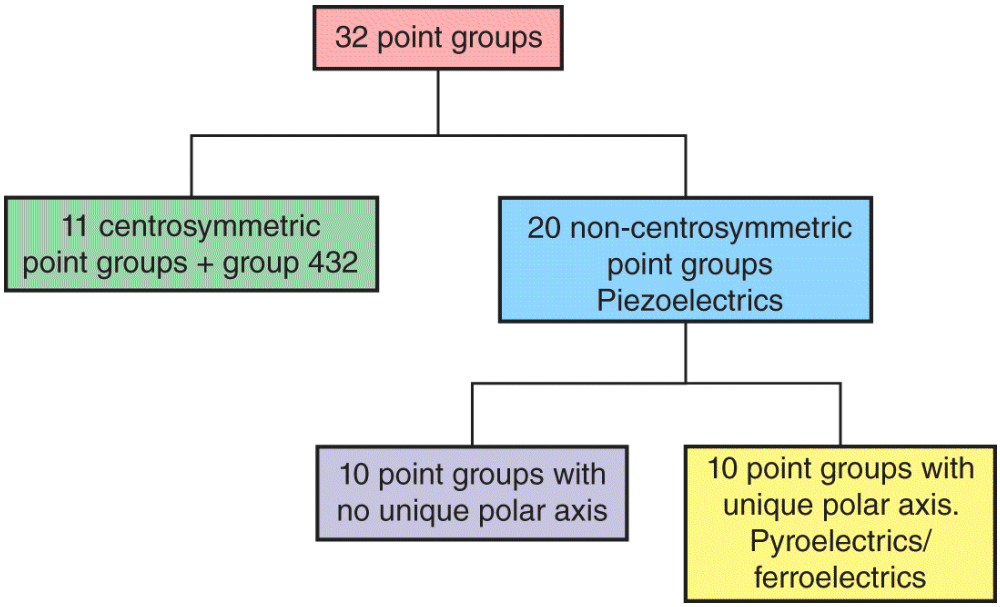


Figure 9.7 The relationship between point-group symmetry and piezoelectric and pyroelectric properties.

The relationship between the measured polarisation **P** and the stress applied σ can, to a *first approximation*, by written as:

$$\mathbf{P} = d\,\sigma$$

where d is the *piezoelectric modulus* or *coefficient*. The description of the piezoelectric effect is, in fact, far more complex than this because stress is a directional property and the degree of polarisation produced will depend upon both the magnitude and direction of the stress. To take all of the possibilities into account requires 27 piezoelectric moduli, d_{ijk}. However, depending upon the actual symmetry operators present in the unit cell, some of these d_{ijk} will be equal to each other and some will be

zero so that the number of different moduli will be less than 27 for most symmetry classes.

The requirement that the piezoelectric effect is restricted to non-centrosymmetric crystals implies that piezoelectricity should not be observed in a polycrystalline solid. This is because the individual grains will polarise in random directions that will cancel overall. It is possible to get around this problem in some piezoelectric materials, as detailed in Section 9.3.7.

A *pyroelectric* crystal possesses a spontaneous polarisation, $\mathbf{P}_s$ at all times, even in the absence of an electric field. This is due to elementary dipoles in the crystal that are aligned to give an observable external bulk polarisation. As in the case of piezoelectrics, any elementary dipoles will cancel out if the crystallographic unit cell has a centre of symmetry and the resultant $\mathbf{P}_s$ will be zero. However, another condition is also needed to produce a spontaneous polarisation, the presence of a *unique polar axis*, which is a direction in the crystal unrelated by symmetry to any other direction, not even the antiparallel direction. The dipoles must lie parallel to the polar axis of the crystal except for one monoclinic crystal class where the polarisation direction is allowed to lie in a plane. Of the 20 piezoelectric crystal classes, only 10 fulfil this criterion and give rise to the pyroelectric effect (Figure 9.7).

The relationship between the change in spontaneous polarisation, $\Delta\mathbf{P}_s$ and the change in temperature, ΔT, can be written as:

$$(\Delta \mathbf{P}_s)_i = \pi_i \, \Delta T$$

where π_i is the *pyroelectric coefficient* and i takes values of 1, 2, or 3 and refers to the unique x-, y-, or z-axis. Typical values of π are of the order of 10^{-5} C m^{-2} K^{-1}.

The pyroelectric effect that is normally observed in a crystal is, in fact, composed of two separate effects called the *primary* (or *true*) pyroelectric effect and the *secondary* pyroelectric effect. If a crystal is fixed so that its size is constant as the temperature changes, the primary effect is measured. Normally though, a crystal is unconstrained. An additional pyroelectric effect will now be measured, the secondary pyroelectric effect, due to strains in the crystal produced by the thermal change. In general, the secondary effect is much greater than the primary effect, but both are utilised in devices.

Among the structurally simplest pyroelectrics are hexagonal ZnO (zincite), and the isostructural hexagonal ZnS (wurtzite). In these crystals, the structure is built of layers of metal and non-metal atoms, with the metals surrounded by a tetrahedron of non-metals. These tetrahedra all point in the same direction, along the polar axis, which is the **c**-axis (Figure 9.8). The tetrahedra are *slightly flattened*, which gives rise to an internal electric dipole, which lies parallel to the **c**-axis. A change in temperature will modify the dimensions of the tetrahedron and so change the existing dipole moment in each of these units, giving rise to pyroelectricity. Materials showing the pyroelectric effect are used as infrared radiation detectors.

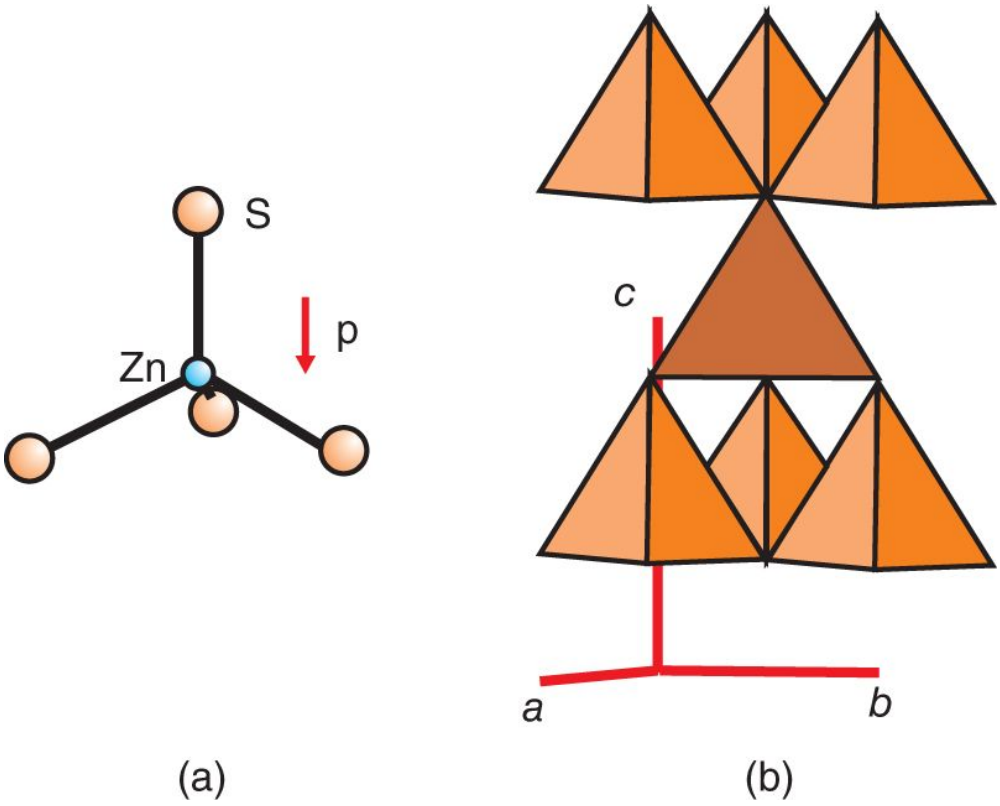


Figure 9.8 The structure of hexagonal ZnS (wurtzite): (a) a ZnS_4 tetrahedron shows an electric dipole, **p**, parallel to the **c**-axis of the unit cell; (b) the parallel orientation of the tetrahedra in a unit cell.

9.2.3 Piezoelectric Mechanisms

A piezoelectric crystal may develop surface charges on being stressed as a result of induced bulk polarisation due to the formation of internal dipoles or alternatively to the rearrangement of existing dipoles. Two examples follow.

In the first, suppose that a crystal is built up of metal-oxygen MO_4 tetrahedra. (Note that piezoelectricity is not confined solely to crystals containing tetrahedral groups.) In an *ideal* MO_4 tetrahedron, the centre of gravity of the negative charges, arising from the combined effects of the oxygen atoms and the chemical bonds, will coincide with the centre of gravity of the positive charges arising in the metal atom, M (Figure 9.9a). A force applied to the top of a tetrahedron will cause a deformation. The oxygen–metal bond in line with the force will resist deformation most, as the positive metal and negative oxygen atoms are being forced together. The basal triangle of oxygen atoms will be flattened, and as there are no metal atoms to directly oppose this change, it will occur to a greater degree than the other deformation. The centre of gravity of the negative charges will no longer coincide with the centre of gravity of the positive charge, and a dipole will result. However, the direction of the force is important, and in some directions, polarisation will not occur. A force directed perpendicular to a tetrahedron edge (Figure 9.9b) will deform the all bonds equally, and not give rise to any dipoles.

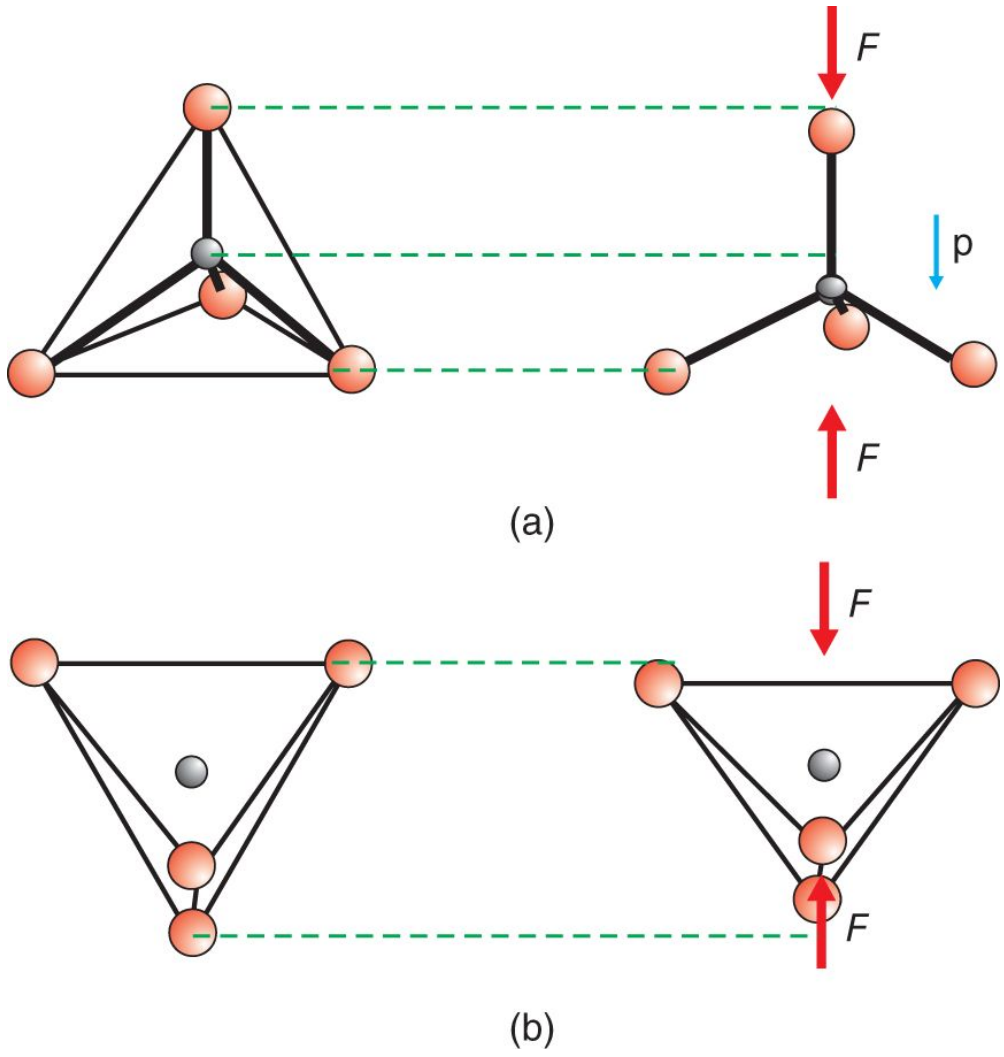


Figure 9.9 Piezoelectricity: (a) a force applied to a tetrahedron along a bond gives rise to a dipole due to distortion; (b) a force applied perpendicular to an edge does not.

The piezoelectric effect can also be generated in a crystal already containing dipoles. In some materials these dipoles add to zero in the absence of stress. When the crystal is deformed, the dipole directions change slightly, so that an overall polarisation is observed. This happens in quartz, SiO_2, which is the most widely used piezoelectric material. The room temperature structure of quartz (α-quartz) contains helices of distorted corner-connected SiO_4 tetrahedra running down the **c**-axis (Figure 9.10a). Each of these shows an internal electric dipole, but these add to zero over a unit cell. Application of a stress to the crystal distorts the structure, so that the dipoles no longer cancel, with the consequence that an overall polarisation is produced (Figure 9.10b).

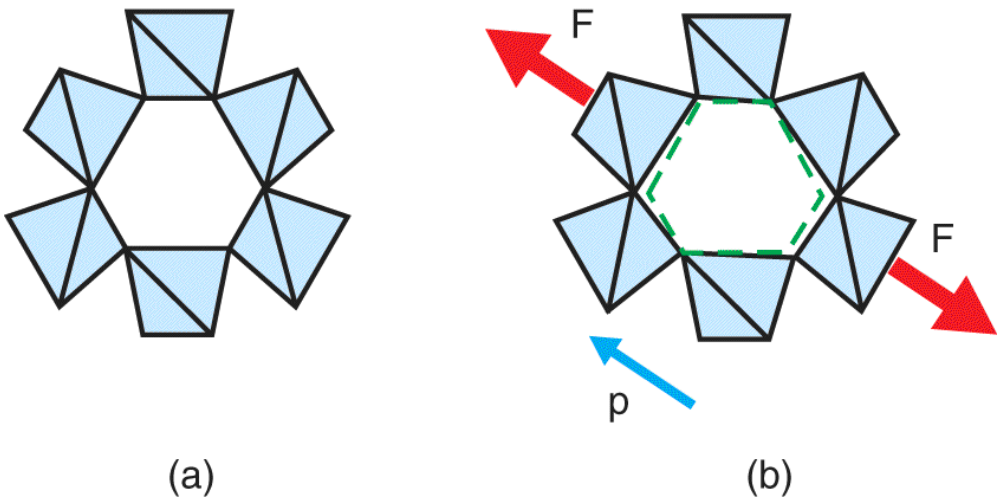


Figure 9.10 Piezoelectricity in quartz: (a) Part of the structure of room-temperature α-quartz, projected down the **c**-axis; (b) application of a force to the structure causes a distortion (exaggerated, dotted lines) so that the internal dipoles no longer cancel. (Note that the tetrahedra are arranged in a helix, not rings.)

9.2.4 Quartz Oscillators

One of the earliest technological uses of quartz crystals was in oscillators. Single crystals of quartz have a hexagonal section normal to the crystallographic **c**-axis (Figure 9.11a). The degree of polarisation produced depends upon the direction of the stress with respect to the crystallographic structure. For example, a stress along *Y* will generate a polarisation along *X*. Similarly, an electric field applied along *X* will cause the crystal to elongate or contract along *Y*. Similar relationships apply to other directions in the crystal.

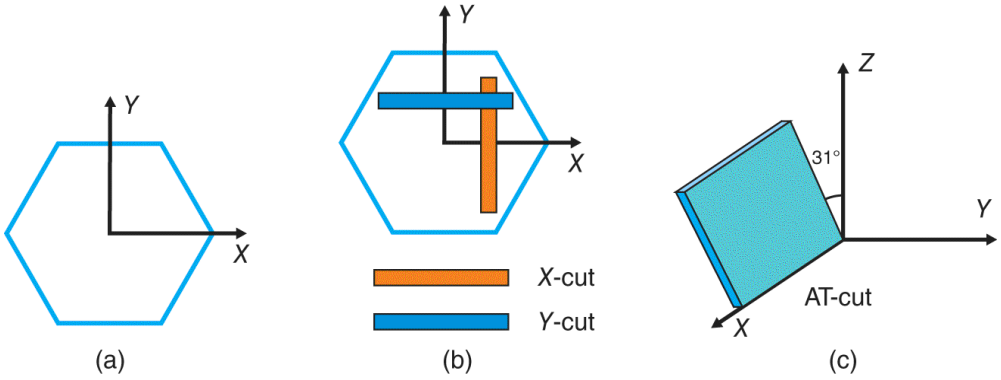


Figure 9.11 Quartz crystals: (a) section through a crystal normal to the **c**-axis; (b) *X*-cut and *Y*-cut plates; (c) AT-cut plate. (Note that *X* and *Y* are not crystallographic axes, but *Z* is parallel to the crystallographic **c**-axis.)

In their simplest form, oscillators consist of a slice of quartz with metal electrodes plated onto opposite faces. Utilising the inverse piezoelectric effect, the application of a varying voltage to the electrodes will cause the quartz to expand and contract, it will vibrate or oscillate. However, these vibrations will only mount to significant levels if the voltage variation has, more or less, the same frequency as the mechanical vibration

modes of the crystal. (An analogy is with a swing. Push a swing at any frequency you like, but you will only get the swing to 'work' if your pushes are more or less in time with the allowed undulations of the seat, which depend precisely upon the length of the ropes connecting the seat to the surrounding framework.) In the same way, the mechanical vibration frequency of a quartz plate is determined exactly by its dimensions and is very 'sharp'. To obtain an oscillator with the desired frequency, crystals are initially mechanically polished and then fine-tuned with lasers to obtain high dimensional precision. The crystal vibrations are then used to generate an output voltage with an extremely precise frequency by detecting the voltage change induced by the direct piezoelectric effect through a second set of electrodes. (The use of two separate sets of electrodes was rapidly outmoded and now all oscillators use just one pair for both input and output.)

Quartz crystal oscillators were used in 1926 for the generation of precise radio broadcasting frequencies. The accuracy of the vibrations also meant that they afford a means of measuring time superior to the best mechanical clocks and quartz crystal clocks were introduced in 1927. The first commercial quartz wristwatches were available in 1969, from Seiko. These use quartz crystals that are shaped like a tuning fork, about 3 mm long and 0.3 mm thick, chosen so as to generate a frequency of 32 768 (2^{15}) Hz because this can be conveniently halved multiple times to give an effective one second interval that can be displayed.

For precision work, the temperature variation of the oscillations of a crystal is of more importance than the magnitude of the polarisation produced and much effort has been spent on obtaining slices with a zero temperature-coefficient of oscillation. For example, an *X-cut*, a slice with faces parallel to Y and Z and normal to X, has a temperature coefficient of frequency variation of -22×10^{-6} per °C. A *Y-cut* has faces parallel to X and Z, perpendicular to Y with a temperature coefficient of frequency variation of $+90 \times 10^{-6}$ per °C (Figure 9.11b). In making a temperature stable oscillator, the aim is to make a cut, which balances negative and positive temperature coefficients so that at a chosen temperature, usually 25 °C, the temperature coefficient is zero. The first of these cuts, the AT-cut, contains the X direction and is tilted to the Z direction by approximately 31° (Figure 9.11c). Other thermally stable cuts are now known and are used for certain applications.

9.2.5 Piezoelectric Polymers and Biomolecular Materials

At first sight it might seem surprising that polymers can exhibit piezoelectricity, but it is so. Indeed, the requirements to produce piezoelectricity are the same as those just given. That is, the material should contain pressure induced or pressure sensitive elementary dipoles and these should be incorporated into a crystalline matrix that lacks a centre of symmetry. Piezoelectric polymers generally rely upon permanent dipoles on the polymer chains. There are two main sources of these dipoles, firstly strongly polar bonds, such as carbon–fluorine, carbon–chlorine, and carbon–nitrogen, and secondly, hydrogen bonds. Polar carbon–fluorine bonds are found in the polymers such as poly(vinyl fluoride) $[CH_2–CHF]_n$, known as PVF (Figure 9.12a) and poly(vinylidene fluoride) $[CH_2–CF_2]_n$, known as PVF_2 (Figure 9.12b). The negative end of the dipole is located on the fluorine atom (C ← F), and a smaller dipole is found on the carbon–hydrogen bond (C → H). In PVF, the overall dipole moment is greatest in the isotactic

form of the polymer, in which all of the fluorine atoms are on the same side of the carbon–carbon backbone (Figure 9.12c). As would be expected, atactic polymers, in which the fluorine atoms have a random distribution, do not show a significant piezoelectric effect. The polymer PVF_2 also has an overall dipole composed of a C ← F dipole opposed by a C → H dipole. Again, the isotactic form of the polymer has the highest net dipole moment (Figure 9.12d). Defects in the chain, especially caused by irregular linking of the monomer units during polymerisation, reduce the overall dipole moment of the chains.

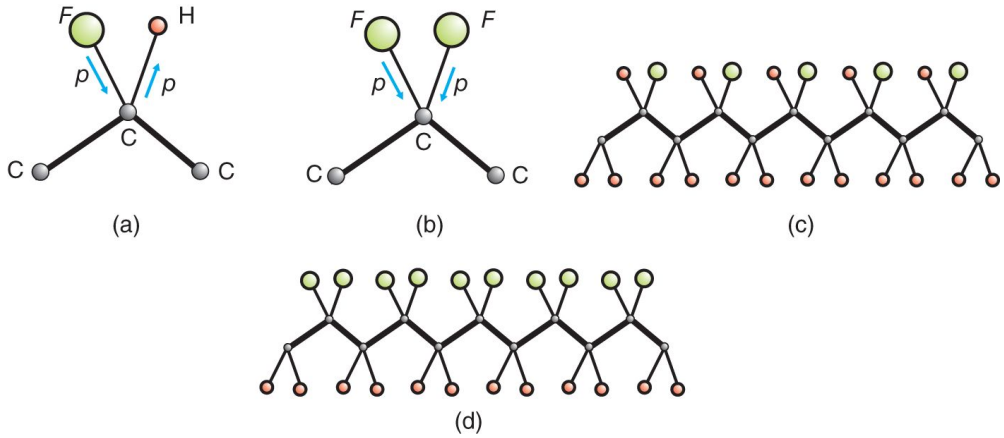


Figure 9.12 (a) dipoles present in a tetrahedral unit of PVF, poly(vinyl fluoride); (b) dipoles present in a tetrahedral unit of PVF_2, poly(vinylidene fluoride); (c) isotactic structure of a PVF chain; (d) isotactic structure of a PVF chain.

Hydrogen bonding produces the polarisation in polyamides, better known as nylons. The relative configurations of the hydrogen bond dipoles depend upon the spacing between the amide groups along the polymer chain. In the case of even polymers, such as nylon 6, dipoles are opposed along the chain (Figure 9.13a). In this case, the dipoles cancel internally and no overall dipole moment is observed. In odd polymers, such as nylon 5, the dipoles are aligned to give an observable polarisation (Figure 9.13b).

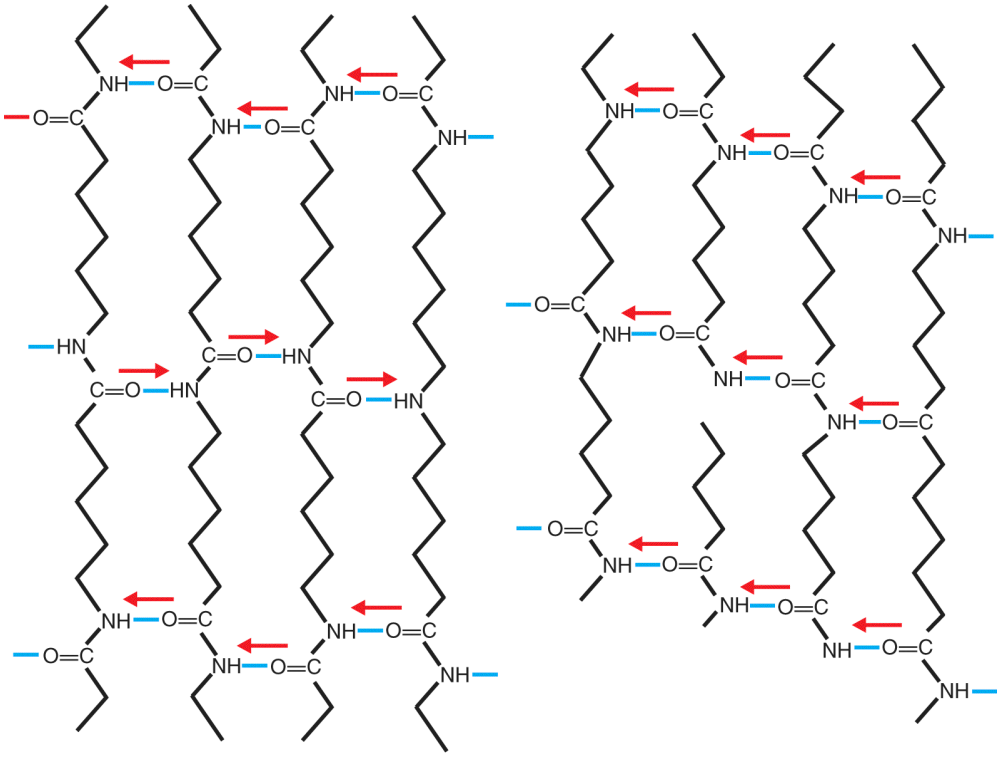


Figure 9.13 The electric dipoles present in (a) chains of nylon 6; (b) chains of nylon 5.

In addition to the presence of elementary dipoles, it is important for the polymer to form non-centrosymmetric crystals. Polymer chains can usually pack together in several different ways. For example, poly(vinylidene fluoride), PVF_2, can crystallise in four forms to produce either non-polar or polar crystals (Figure 9.14). In the centrosymmetric form the electric dipoles in a unit cell cancel and the material does not display piezoelectricity. In non-centrosymmetric structures the electric dipoles add, and the material displays piezoelectric effects. Naturally, the degree of crystallinity of the polymer strongly influences the magnitude of the observed piezoelectric effect. Careful processing is important in the production of good piezoelectric films.

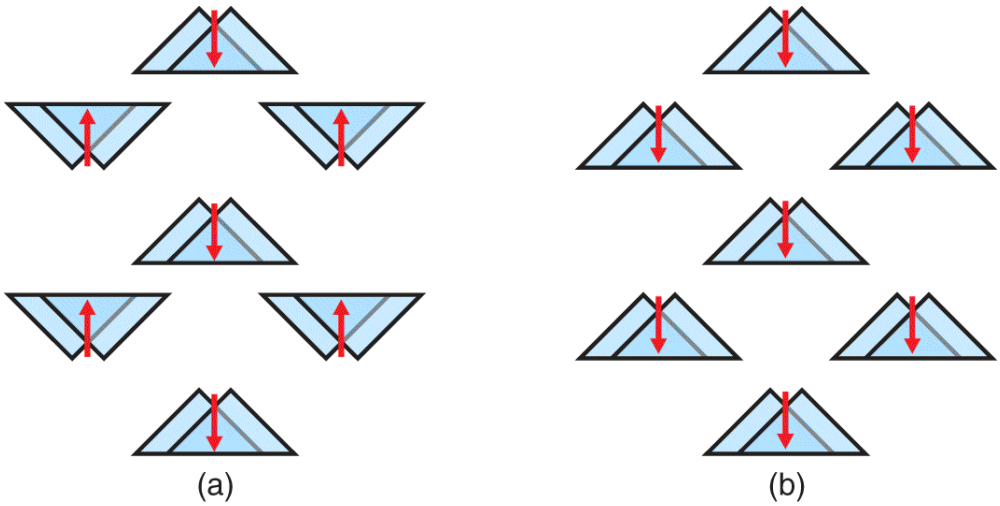


Figure 9.14 The crystal structures, schematic, of two forms of PVF_2 viewed down the polymer chains; (a) non-piezoelectric centrosymmetric structure; (b) piezoelectric non-centrosymmetric structure.

Piezoelectric plastic sheets can also be fabricated. These materials are known as *electrets*. Electrets are thin polymer films of high resistance that are polarised in a high field or by having an electric charge 'sprayed' onto the surface from a discharge. The resistance is so high that they retain the polarisation permanently. In these materials, the dipole moment is very large, due to the large (in atomic terms) separation of the charges on the opposed faces of the plastic.

Although polymer piezoelectrics do not generally show such high piezoelectric coefficients as ceramic materials, they have some important advantages. Among other things, polymer films are of low density and flexible, which makes them suitable for sensors and transducers in microphones, keyboards, and flat panel speakers.

Piezoelectric polymers are also of importance in living organisms, which have to react to mechanical stimuli such as pressure, touch or sound waves in many different tissue types, from hair cells in the inner ear to smooth muscle cells in blood vessels. These mechanical impulses are invariably transmitted as electrical signals generated by the piezoelectric effect. For example, the strengthening of bone under repeated stress is believed to be mediated by a piezoelectric response, as are reactions involving ion channels, which open in response to tension in the cell membrane to allow ions to pass.

The origin of piezoelectricity in biomolecules is similar to that in artificial polymers and stems from the orientation of individual polar molecules in the material. Proteins are polymers of amino acids. Molecules of amino acids contain an acidic –COOH group and a basic $–NH_2$, which imparts a dipole, running from COOH (–) to NH_2 (+). The naturally occurring amino acids, typified by glycine (NH_2CH_2COOH), are often called α-amino acids, this meaning that the $–NH_2$ group is bonded to the carbon atom (the α carbon), next to the –COOH group. Naturally occurring proteins are built from 20 natural α-amino acids via the reaction of the acidic –COOH group on one molecule with the basic $–NH_2$ group on another. The resulting link is called a *peptide bond*. If there are fewer than about 50 amino acids in the chain, the molecules are referred to as *polypeptides*, while if there are more than this number, they are usually called

proteins. As H_2O is lost during the formation of a peptide bond, the remaining section of amino acids are called amino acid *residues*. The sequence of amino acid residues is then called the primary structure of the protein and has an amino group ($–NH_2$) at one end of the molecular chain and a carboxyl (–COOH) group at the other.

Although the primary structure of a protein is of importance, the way in which the polypeptide chain folds controls the chemical activity of the molecule and the piezoelectric properties. These local forms make up the *secondary structure* of the protein. A common arrangement is the *α-helix*, a right-handed coil. The backbone of the helix is composed of the repeating sequence: (i) carbon atom double bonded to oxygen, C□O; (ii) carbon atom bonded to a general organic group, C□R; (iii) nitrogen, N,(Figure 9.15a). The helical conformation is a result of the geometry of the bonds around the carbon atoms and hydrogen bonding between hydrogen atoms on N and oxygen atoms on C□O groups. The repeat distance between any two corresponding atoms, the *pitch* of the helix, is 0.54 nm, and there are 3.6 amino acid residues, in each turn (Figure 9.15b). Because of this regular structure, the $–NH_2$ end of the helix has a positive charge relative to the negative –COOH end.

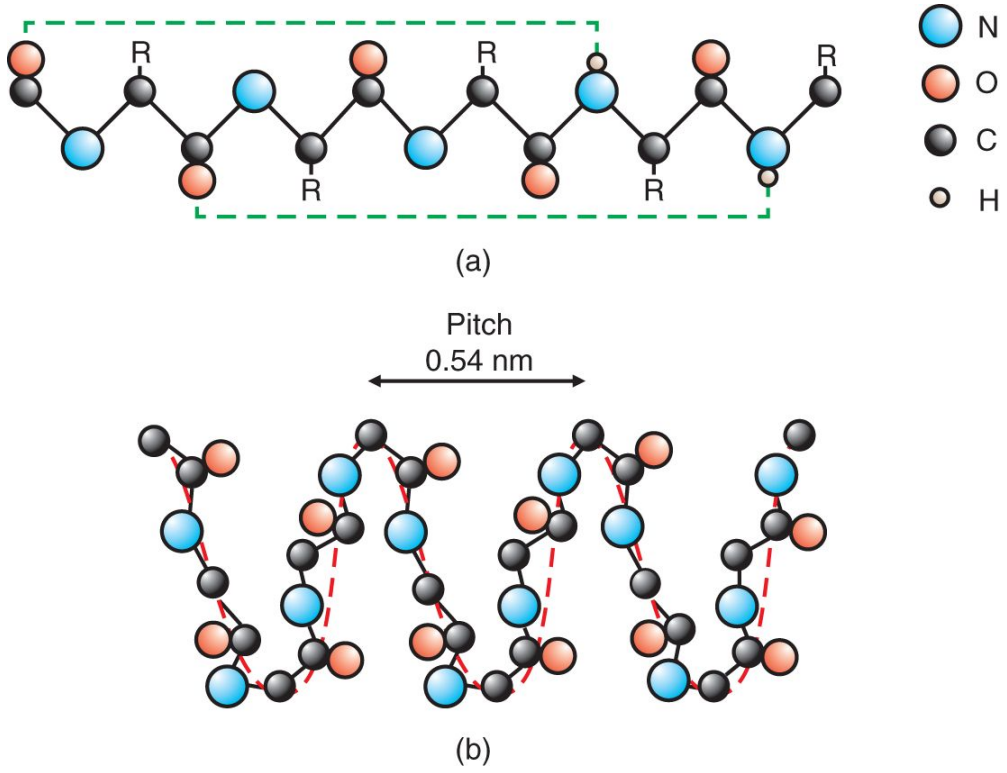


Figure 9.15 The α-helix, schematic: (a) the succession of carbon and nitrogen atoms along the backbone of an α-helix, hydrogen bonds form between O atoms on carbon and H atoms on nitrogen linked by dashed lines; (b) coiled α-helix.

Keratin and collagen are fibrous proteins, which preserve these helices and their piezoelectric properties intact. Keratin is essentially composed of single α-helix chains arranged in a more or less parallel fashion. The polarity of the chains then allows for piezoelectricity in the protein. Collagen has a more complex structure, in which three α-

helices are braided together in a chain; these chains then coalesce parallel to each other to form the protein. Again, as the chains are parallel to each other they exhibit piezoelectric effect.

9.3 Ferroelectrics

Ferroic is a generic term for crystals that have an internal structure that can be switched from one stable orientation to another that is equally stable, by the application of a suitable driving force along an appropriate direction. The best-known ferroic materials are the ferroelectrics (described in the following text) and ferromagnetics (described in Section 10.3), with switching driven by, respectively, electric or magnetic fields. The switching between stable orientations is called a ferroic transition. Ferroic crystals frequently show hysteresis loops under the application of the switching stimulus. The physical changes taking place in a ferroic transition are:

(i) *Ferroelectric*: spontaneous electrical polarisation, switched by electric fields;

(ii) *Ferromagnetic*: spontaneous magnetisation, switched by magnetic fields;

Ferroic materials generally show a domain structure that delineates regions with parallel electric or magnetic spin directions, which differ in direction to those in neighbouring domains. Multiferroics are single phases that exhibit more than one ferroic property simultaneously.

9.3.1 Ferroelectric and Antiferroelectric Crystals

Ferroelectrics are distinguished from pyroelectrics by virtue of the fact that the direction of the spontaneous polarisation, $\mathbf{P}_S$, can be switched. A variety of crystallographic features can result in ferroelectric behaviour and many different chemical compounds are classified as ferroelectrics (Table 9.2). In all of these materials, the root cause of the effect is a displacement of atoms from a symmetrical to a non-symmetrical environment, thus creating internal electric dipoles.

Table 9.2 Ferroelectrics and antiferroelectrics.

Compound	Formula	Curie temperature (K)	Spontaneous polarisation ($C\,m^{-2}$)	Relative permittivity[a]
Hydrogen bonded compounds				
Rochelle salt	$NaK(COO{\cdot}CHOH)_2{\cdot}4H_2O$	298	0.01	5×10^3
Triglycine sulfate	$(NH_2CH_2COOH)_3{\cdot}H_2SO_4$	322	0.03	2×10^3
Potassium dihydrogen sulfate	KH_2PO_4	123	0.05	6×10^5
Polar groups				
Sodium nitrite	$NaNO_2$	43	0.08	1.1×10^3
Perovskites				
Barium titanate	$BaTiO_3$	403	0.26	1×10^4
Lead titanate	$PbTiO_3$	763	0.80	9×10^3
Potassium niobate	$KNbO_3$	691	0.30	4.5×10^3
Tungsten bronzes				
Sodium barium niobate	$Ba_2NaNb_5O_{15}$	833	0.40	6×10^4
Antiferroelectrics				
Tungsten trioxide	WO_3	1010	0	300a
Ammonium dihydrogen phosphate	$NH_4H_2PO_4$	148	0	57, 10a
Lead hafnate	$PbHfO_3$	476	0	200a
Lead zirconate	$PbZrO_3$	503	0	150a
Sodium niobate	$NaNbO_3$	627	0	700, 70a

[a] These materials have anisotropic relative permittivity values that vary considerably with temperature. Upper and lower values are shown when these differ substantially. The values cited are to show orders of magnitude only.

The process can be illustrated by a simple model. Suppose we have a rectangular array of anions with centred cations (Figure 9.16a). This arrangement is non-ferroelectric. However, when relatively small cations are involved, the structure may gain stability when the cations are displaced slightly from the centre of the surrounding anion coordination polyhedron. The centre of gravity of the anion array will not now coincide with the positive cation and each unit now contains a dipole (Figure 9.16b). Repetition of this motif builds a structure with an aligned dipole array characteristic of a ferroelectric (Figure 9.16c,d). For an example of this type of behaviour, see Section 9.3.6, $BaTiO_3$.

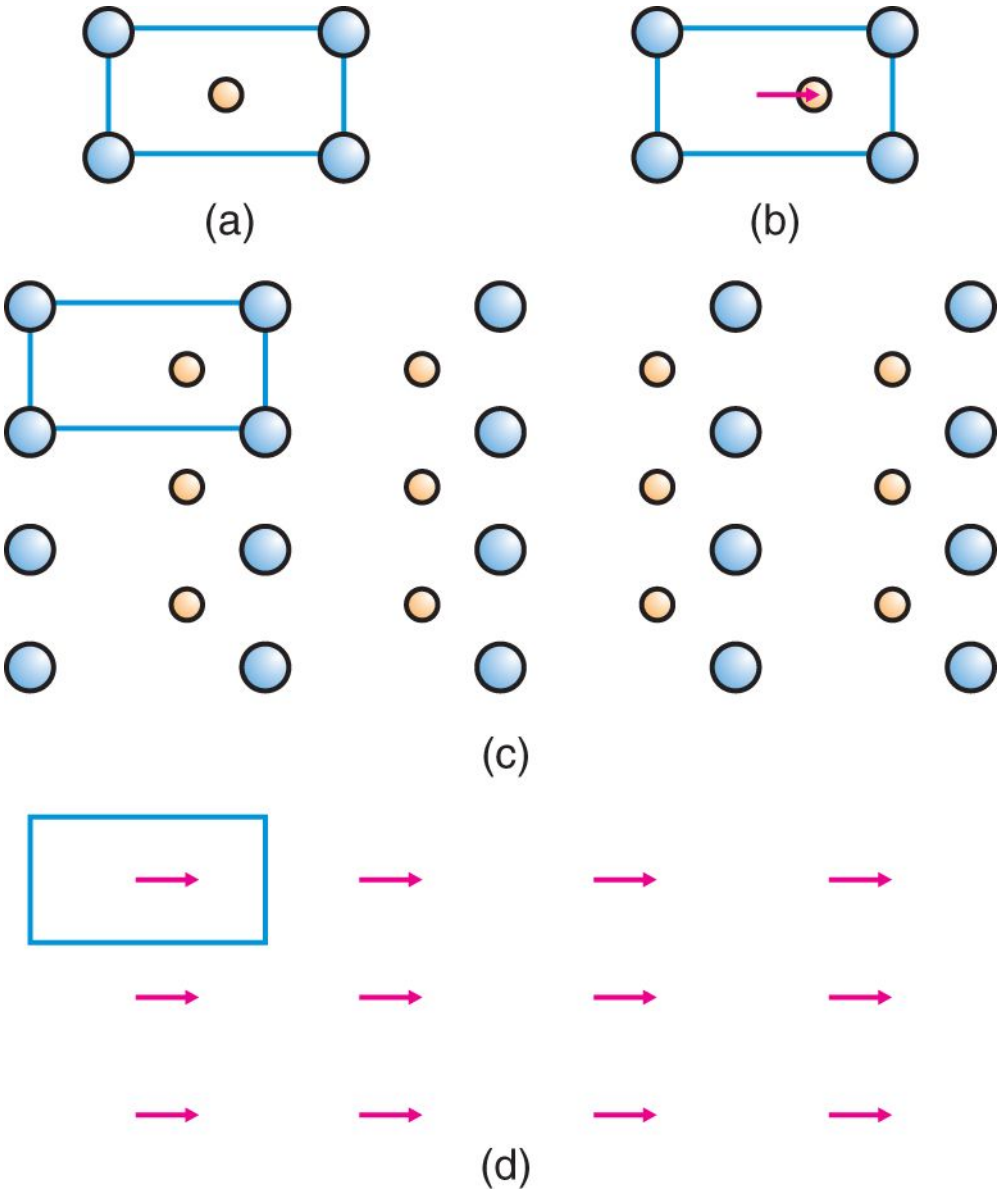


Figure 9.16 Ferroelectric crystal formation, schematic: (a) a cation-centred unit cell; (b) cation displacement and dipole formation; (c) the structure as an array of unit cells; (d) the structure as an array of electric dipoles.

In many instances, cation displacement can take place in one of 2 equiv directions. A graph of the potential energy of the cation against position will have two minima, separated by a potential energy barrier, ΔU, corresponding to the two alternative displacements (Figure 9.17). Within the crystal a cation in any particular region might occupy either of these positions at random. Thereafter, local interactions tend to make cations in adjoining units adopt the same displacement so as to form a parallel set of dipoles. Volumes in different parts of the crystal will then show either of the two possible orientations, and the crystal is said to contain *domains* of differing polarisation, separated by domain *boundaries* or domain *walls* (Figure 9.18). Domain walls are not atomically smooth, as figuratively represented, but have a thickness of between 0.5 and 1 nm. A single domain can have surface charges of the order of 1.5×10^{14} electrons cm^{-2}, which can generate an internal electric field of 300 MV or more.

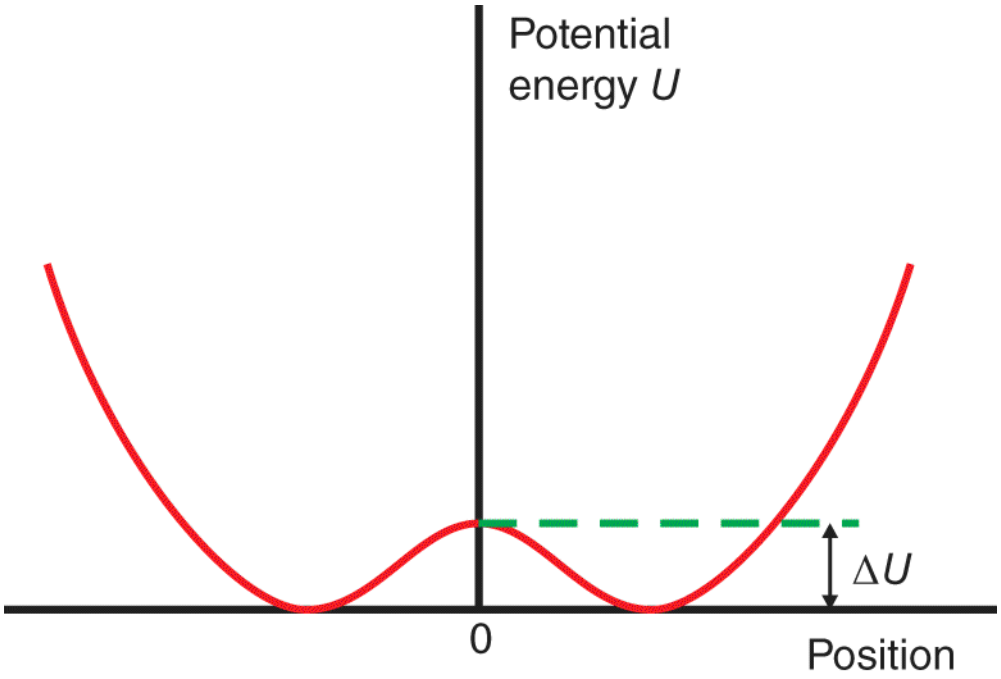


Figure 9.17 Variation of potential energy vs. position for cation displacement from the centre of a surrounding anion polyhedron.

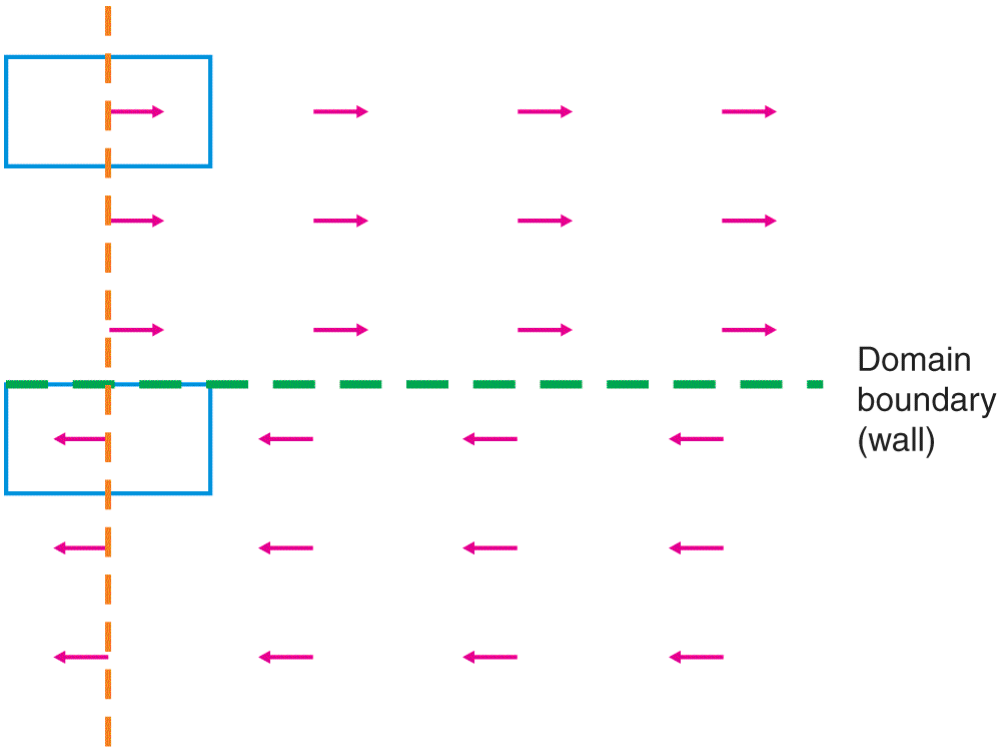


Figure 9.18 Two domains formed by differing alignment of electric dipoles in adjacent regions of a crystal.

Ferroelectricity is governed by two types of factors (i) chemical bonds, which are short-range forces, and (ii) dipolar interactions, which are long range. Calculations of the energy of ferroelectric crystals indicate that a minimum energy results when all the elementary dipoles are parallel or all the dipoles are in an antiparallel arrangement. The antiparallel arrangement is found in *antiferroelectrics* (Table 9.2, Figure 9.19).

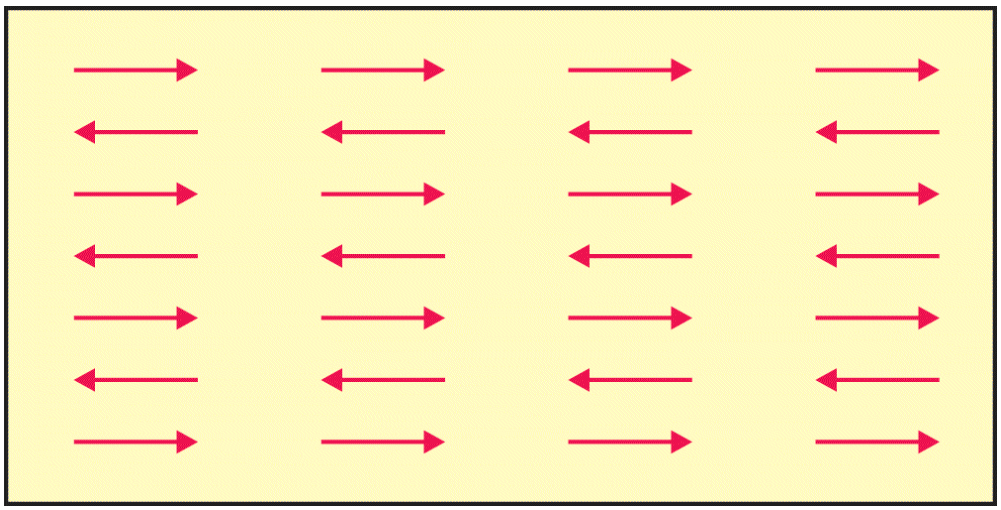


Figure 9.19 Antiparallel arrangement of electric dipoles in an antiferromagnetic, schematic.

The balance between ferroelectric and antiferroelectric states is delicately poised, and some antiferroelectrics readily transform to ferroelectric states. For example, orthorhombic lead zirconate, $PbZrO_3$, with a slightly distorted form of the cubic perovskite structure (Figure 3.11) is antiferroelectric at room temperature. To a first approximation the electric dipoles arise because the Zr^{4+} cations, located in ZrO_6 octahedra, are slightly off-centre at room temperature. The Zr^{4+} cations are readily replaced by Ti^{4+} cations and a solid solution is possible between $PbZrO_3$ and the structurally similar, but ferroelectric $PbTiO_3$. A substitution of 3% or 4% changes the phase from an orthorhombic antiferroelectric to a rhombohedral ferroelectric phase.

9.3.2 Hysteresis and Domain Growth in Ferroelectric Crystals

In general, a ferroelectric crystal will be composed of an equal number of domains oriented in all the equivalent directions allowed by the crystal symmetry. The overall polarisation of the crystal will be zero. If we now apply a small electric field, **E**, in a nominally positive direction, the crystal will behave like a normal dielectric, as the value of **E** is not great enough to overcome the energy barrier between alternative configurations. This corresponds to the segment O–A on a graph of the polarisation vs. field (Figure 9.20). As **E** increases, cations will start to gain sufficient energy to overcome this energy barrier and will be able jump from one potential well to the other. The elementary dipole direction will switch. Gradually all of the domains will change orientation and the observed polarisation will now increase rapidly, corresponding to section AB (Figure 9.20). Ultimately all of the dipoles will be aligned parallel and the crystal will (in principle) consist of a *single domain*. This is the state of *saturation*, B–C (Figure 9.20).

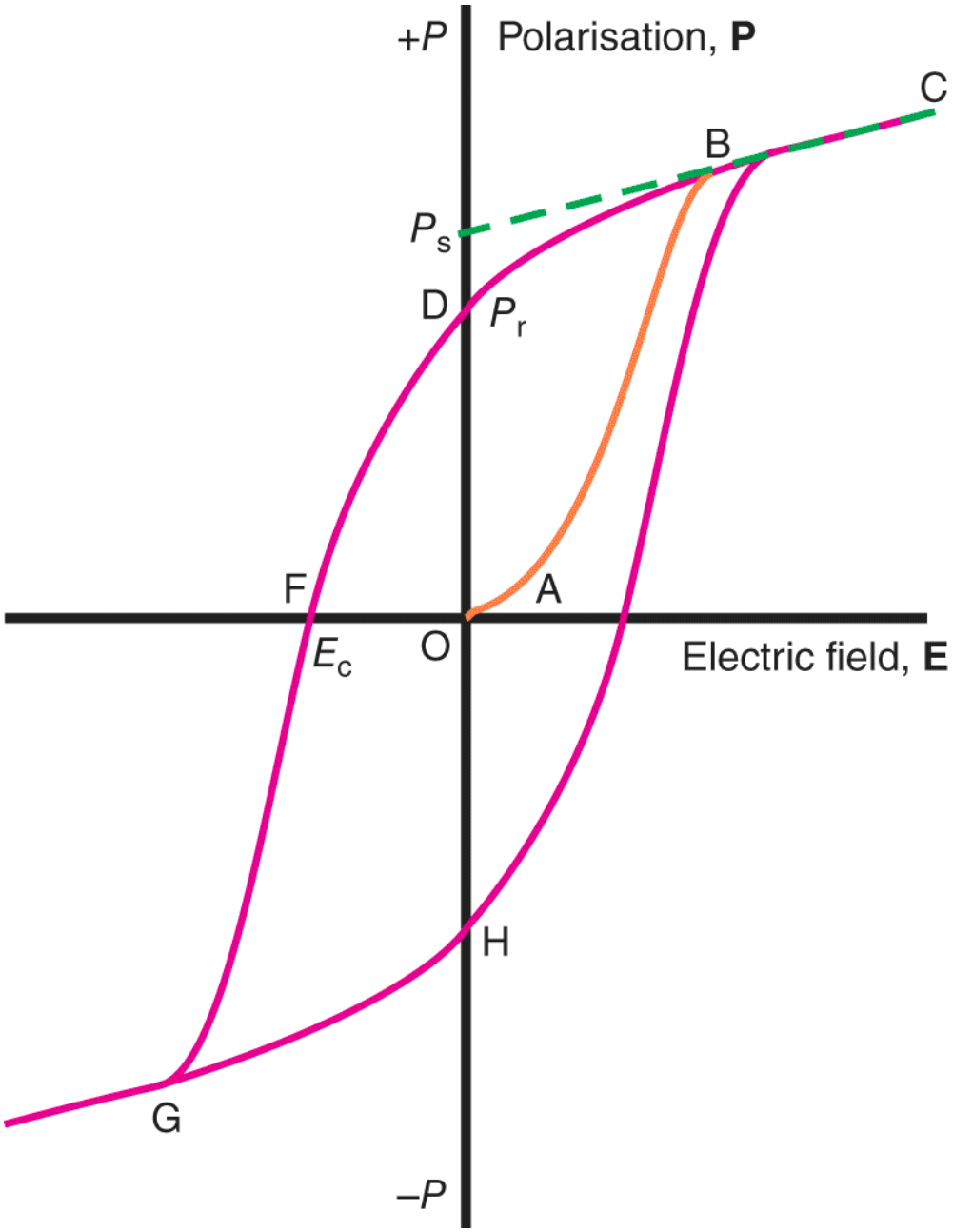


Figure 9.20 Hysteresis behaviour of the polarisation, **P**, vs. applied electric field, **E**, for a ferroelectric crystal.

On reducing and then reversing the applied electric field, the converse takes place but the **P**–**E** path will trace a new route that lags behind the old one because energy has to be supplied to create new domains. Gradually dipoles switch direction, following path C–D–F–G, to reach saturation with dipoles pointing in the opposite direction, at G (Figure 9.20). Reversal of the electric field again causes a reversal of dipole direction, and the curve will follow the path G–H–C. This closed circuit is called a *hysteresis loop* and the phenomenon is *hysteresis* (the meaning of *hysteresis* is 'to lag behind').

Hysteresis is perhaps the most characteristic feature of strongly ferroelectric materials. The value OD is called the *remanent* or *residual polarisation*, $\mathbf{P}_r$, and OF is called the *coercive field* $\mathbf{E}_c$, which tends to fall in the range of 10–100 V. Extrapolation of the linear portion of the curve B–C to $\mathbf{E} = 0$ gives the value of the *spontaneous polarisation* $\mathbf{P}_s$. All of these terms are adopted from the behaviour of ferromagnetic materials (Section 10.3).

The nucleation and growth of domains has long been observed using microscopy and these studies have shown that defects, especially surfaces, are of considerable importance in domain wall movement. As an example, the nucleation and growth of ferroelectric domains in ferroelectric perovskite structure $BiFeO_3$ has been observed when subjected to an electric field applied by a metallic probe. The samples consisted of single crystal (001) sheets of $BiFeO_3$ with the polarisation vector along the [111] direction (Figure 9.21a–c). Application of an electric field caused domains to be nucleated at the lower film surface, not at the metallic probe, as arrow shaped incursions into the film. These amalgamate into a single domain as the field strength increases, with a polarisation along $[11\bar{1}]$ (Figure 9.21d–e).

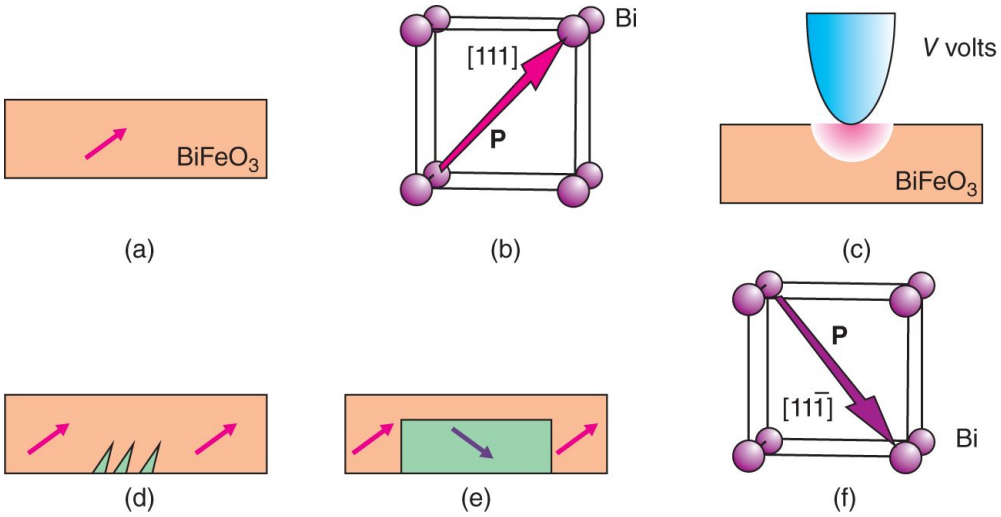


Figure 9.21 Domain growth in ferroelectric $BiFeO_3$: (a) (001) single domain sheet; (b) polarisation vector **P** along [111]; (c) applied voltage *V*; (d) initiation of new domains; (e) domain aggregation; (f) polarisation vector **P** in new domain along $[11\bar{1}]$.

9.3.3 The Temperature Dependence of Ferroelectricity and Antiferroelectricity

The fact that an applied electric field can cause the polarisation to alter its direction implies that the atoms involved make only small movements and that the energy barrier between the different states is low. With increasing temperature, the thermal motion of the atoms will increase and eventually this alone can overcome the energy barrier separating the various orientations. At high temperatures the distribution of atoms becomes statistical and the crystal behaves as a normal dielectric. It is no longer a polar material but is in the *paraelectric* state. The temperature at which this occurs is known as the *transition temperature*, *Curie temperature* or *Curie point*, T_C. The way

in which the spontaneous polarisation of the crystal varies as the Curie temperature is approached depends upon whether the transformation to the paraelectric state can be classified as first or second order. In a first-order transition the polarisation falls abruptly to zero from a finite value, while in a second-order transition the change is gradual and continuous (Figure 9.22a). In either case the relative permittivity rises to a sharp peak in the neighbourhood of T_C (Figure 9.22b).

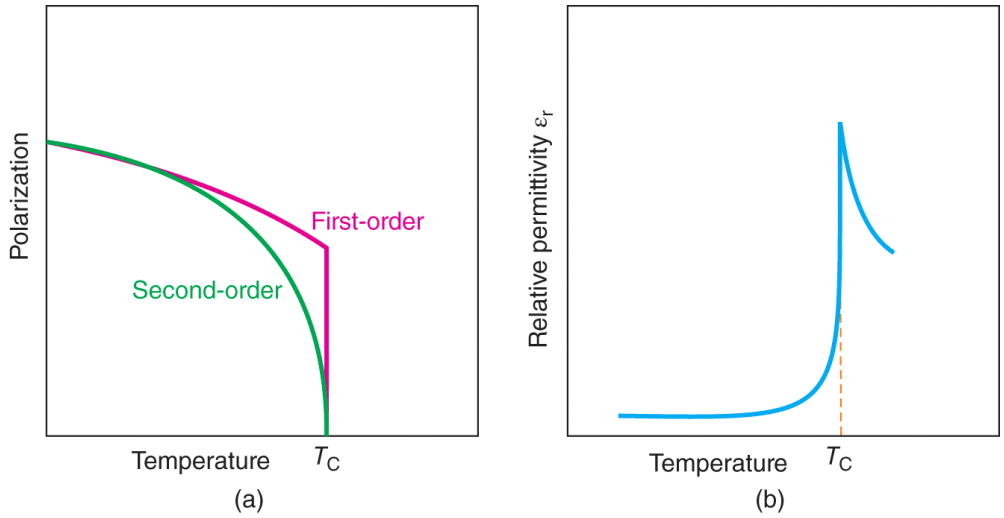


Figure 9.22 Ferroelectric behaviour near to the Curie temperature: (a) first-order and second-order polarisation changes; (b) relative permittivity change.

The temperature dependence of the relative permittivity ε_r of many ferroelectric crystals in the *paraelectric state* can be described fairly accurately by the *Curie–Weiss law*.

$$\varepsilon_r = C/(T - T_C)$$

where C is a constant, T_C is the *Curie temperature*, and T the temperature. The value of C is determined by a plot of $1/\varepsilon_r$ vs. T.

$$1/\varepsilon_r = T/C - T_C/C$$

Ideally the graph is linear, with a slope of $1/C$ and an intercept on the T-axis of T_C (Figure 9.23). Frequently the point of intercept, T_0, is slightly different from the measured value of T_C, and the Curie–Weiss equation is often written in the form:

$$\varepsilon_r = C/(T - T_0)$$

where T_0 is the *extrapolated Curie temperature* (Figure 9.23). Note that the Curie temperature (Curie point) is the temperature at which the structural transition to or from the paraelectric state takes place. The value T_0 (also confusingly referred to as the Curie temperature) is derived from extrapolation and does not coincide with the phase transition.

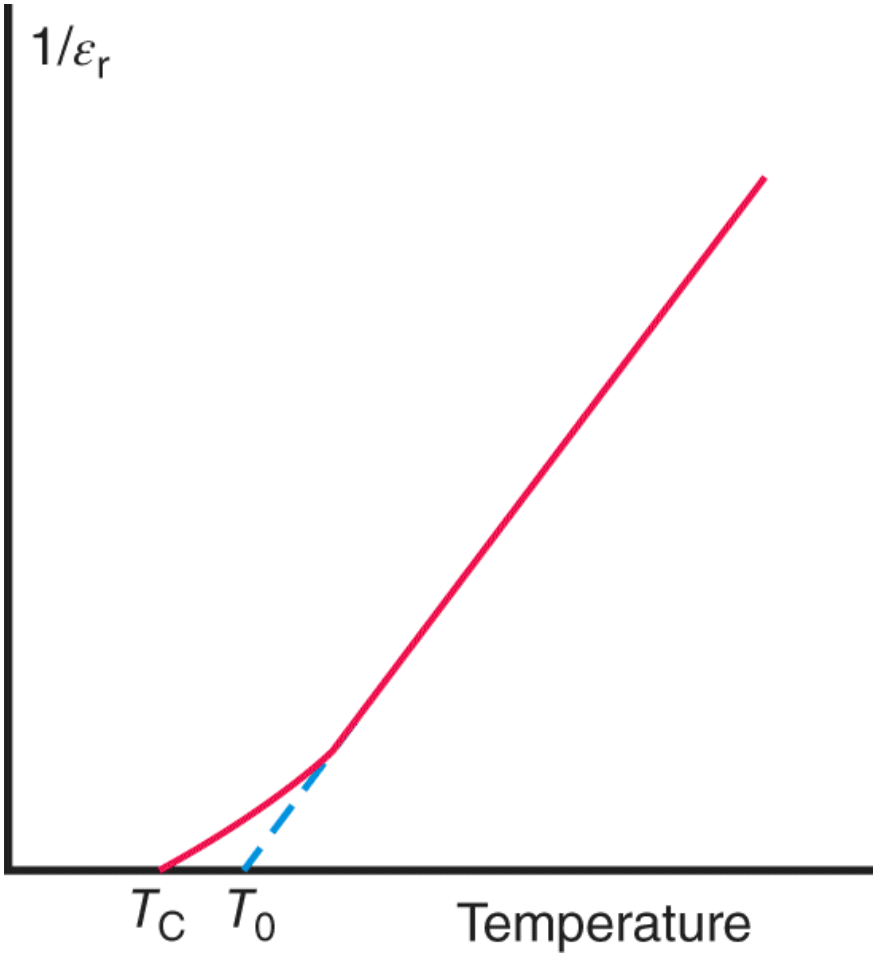


Figure 9.23 The Curie–Weiss behaviour of a ferroelectric above the Curie temperature.

9.3.4 Ferroelectricity Due to Hydrogen Bonds

Hydrogen bonds are formed when a hydrogen atom sits between two electronegative atoms in an off-centre position (Section 1.5). At temperatures below the Curie temperature, the hydrogen atoms sit closer to one atom of the pair, with which it forms a normal chemical bond, while it links to the further atom of the pair by a weak hydrogen bond (Figure 9.24a,b). As the hydrogen in a hydrogen bond can occupy two equally stable positions, it is not difficult to see a possible origin for ferroelectric switching. A sufficiently high electric field will swap the dipole direction by causing the hydrogen ion to jump to the alternative bonding position, reversing the electric polarisation. At temperatures higher than the Curie temperature, atomic vibrations induced by thermal energy overcome the barrier between the two alternatives and the hydrogen atoms will occupy an average position between the two adjacent electronegative atoms, and the polar nature of the solid is lost (Figure 9.24c). The compounds in this group have ordered hydrogen atoms at lower temperatures and these become disordered at higher temperatures in an order–disorder transition (Section 6.4).

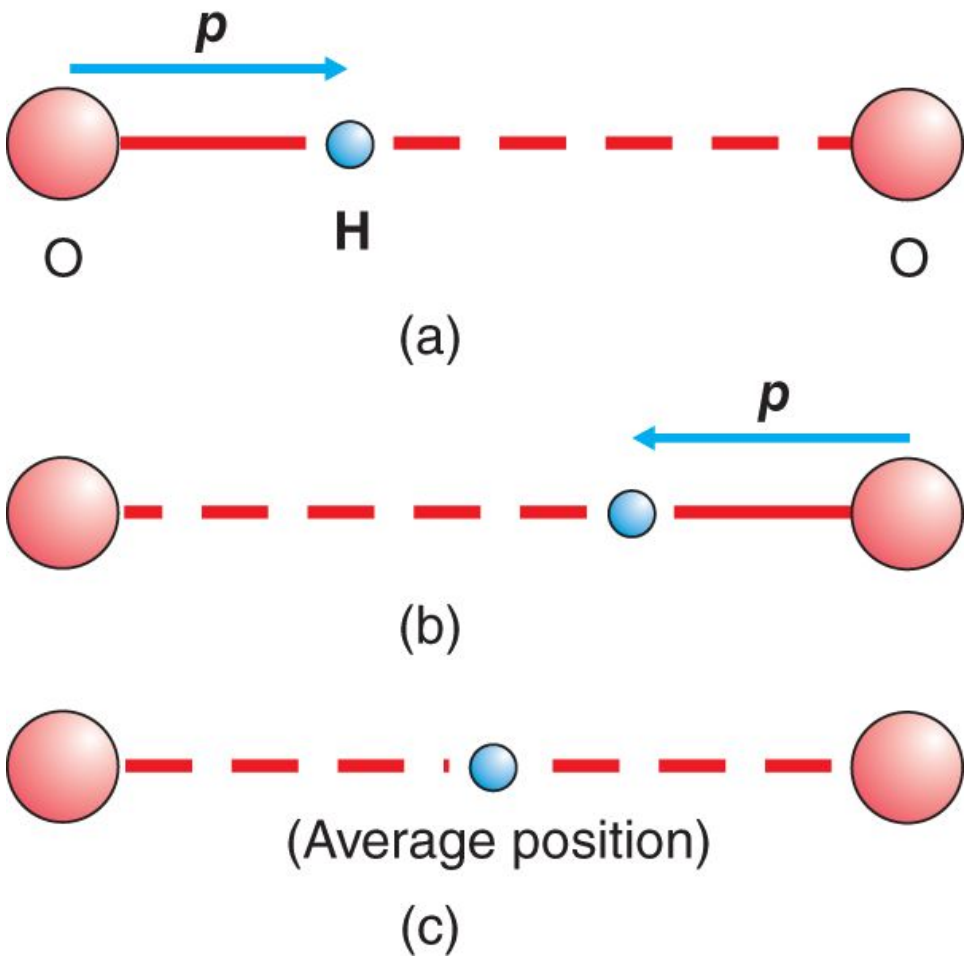


Figure 9.24 Hydrogen bonds (dashed) in ferroelectrics: (a, b) below the Curie temperature; (c) above the Curie temperature.

Hydrogen bonding is the origin of ferroelectricity in potassium dihydrogen phosphate, KH_2PO_4, Rochelle salt (sodium potassium tartrate), Na(COO·CHOH·CHOH·COO)K, and triglycine sulfate $(NH_2CH_2COOH)_3 \cdot H_2SO_4$. However, the interaction of the hydrogen bonds with other features of the crystal structure usually makes each compound unique. This aspect is illustrated by the transition in potassium dihydrogen phosphate. At a temperature above 123 K, this compound is paraelectric. The skeleton of the structure (Figure 9.25a) is made up of regular (PO_4) tetrahedra connected by hydrogen bonds. On average, the hydrogen atoms are found at the centres of the hydrogen bonds and no electric dipoles form. Below 123 K the hydrogen atoms order (Figure 9.25b), so that each (PO_4) tetrahedron in the high temperature form is converted into a $[PO_2(OH)_2]$ tetrahedron (Figure 9.25c). The phosphorus atoms in the tetrahedra are off-centre, pushed away by the hydrogen atoms. The dipoles responsible for ferroelectricity arise in these tetrahedra. They lie along $[00\bar{1}]$ and the **z**-axis is the polar axis. The O□H···O bonds are almost perpendicular to these dipoles, and although hydrogen bonding is the prime cause of ferroelectricity, the hydrogen bonds themselves are not the seat of the dipoles. However, the off-centre positions of the phosphorous

and hydrogen atoms are closely linked. When an external field is applied both the hydrogen and phosphorous atoms switch in concert.

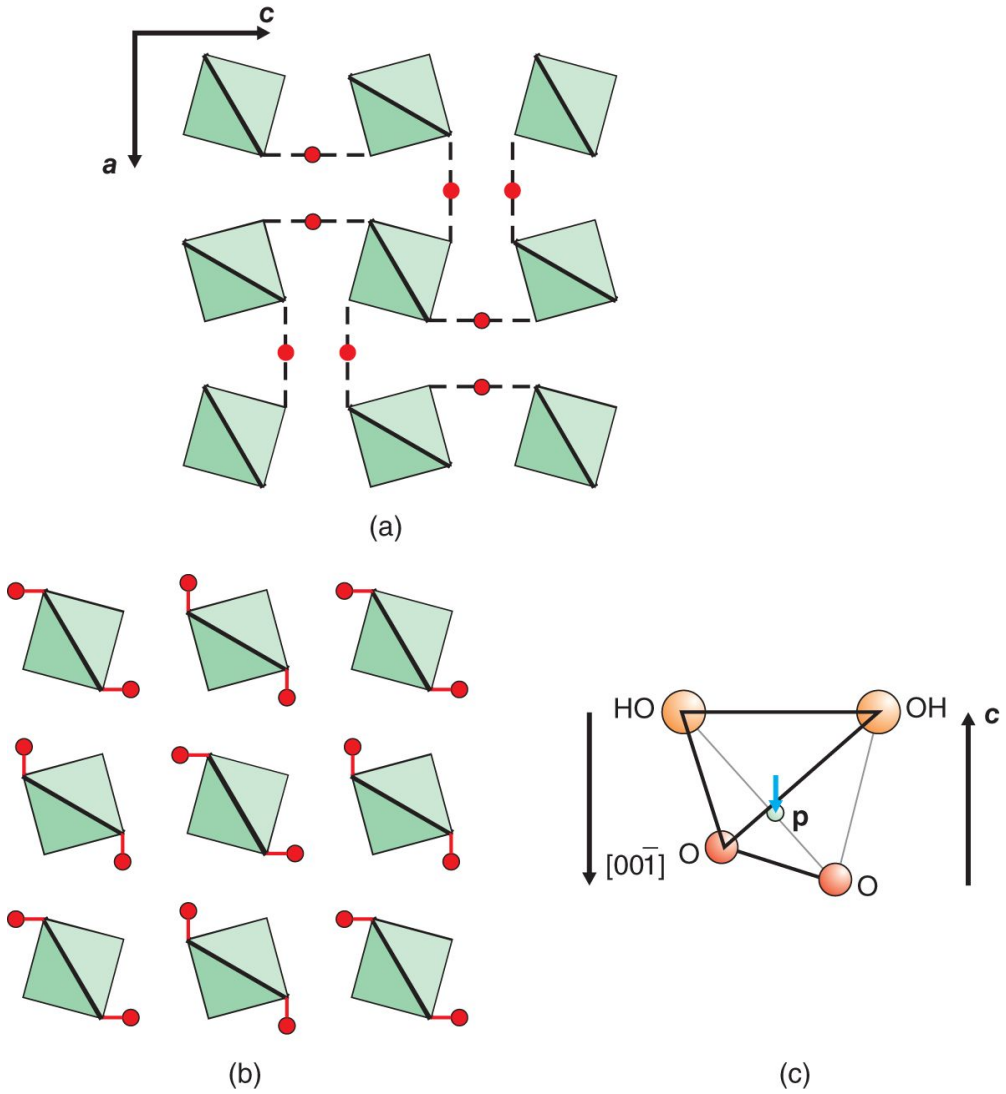


Figure 9.25 Ferroelectric KH_2PO_4: (a) skeleton of the room temperature structure projected down [001] showing PO_4 tetrahedra and H bonds (dashed), with K atoms omitted; (b) low temperature structure with ordered O—H bonds; (c) the displacement of the P atoms in a PO_4 tetrahedron as a result of H-atom ordering and formation of (OH) groups, giving rise to a dipole **p**.

Ammonium dihydrogen phosphate $NH_4H_2PO_4$ is structurally and chemically very similar to ferroelectric potassium dihydrogen phosphate, but the proton ordering is different, and NH_4 H_2PO_4 is an antiferroelectric below the transition temperature.

9.3.5 Ferroelectricity Due to Polar Groups

Compounds with polar groups such as $(NO_3)^-$, which is pyramidal, and nitrite $(NO_2)^-$, which is shaped like an arrowhead, may form ferroelectric phases. At temperatures

below the Curie temperature, these angular units are locked into one position in the solid in an ordered array. In cases where the geometry of the crystal structure will allow, a sufficiently high electric field can reorient such groups, thus causing the dipole to point in a different direction. At temperatures above the Curie temperature, these units may disorder and ferroelectric behaviour is lost in an order–disorder transition.

Sodium nitrite, $NaNO_2$, is an example of this behaviour. The structure is built up from planar NO_2 groups, shaped like blunt arrowheads, the shape of which gives rise to an electric dipole, **p**, pointing towards the N atom (Figure 9.26a). These occupy the Cl positions in the NaCl structure-type adopted by $NaNO_3$. The NO_2 units point along the **b**-axis with their planes perpendicular to the **a**-axis, so that the dipoles in each NO_2^- group are parallel and also point along the **b**-axis (Figure 9.26b). In an applied electric field, these NO_2^- groups can be made to reverse and the material is a ferroelectric. There are two ways in which the dipoles could be imagined to change direction, either the nitrogen atom could flip between the O atoms or the NO_2^- groups could rotate in their own plane. As the NO_2^- ion is fairly rigid, the rotation mechanism operates.

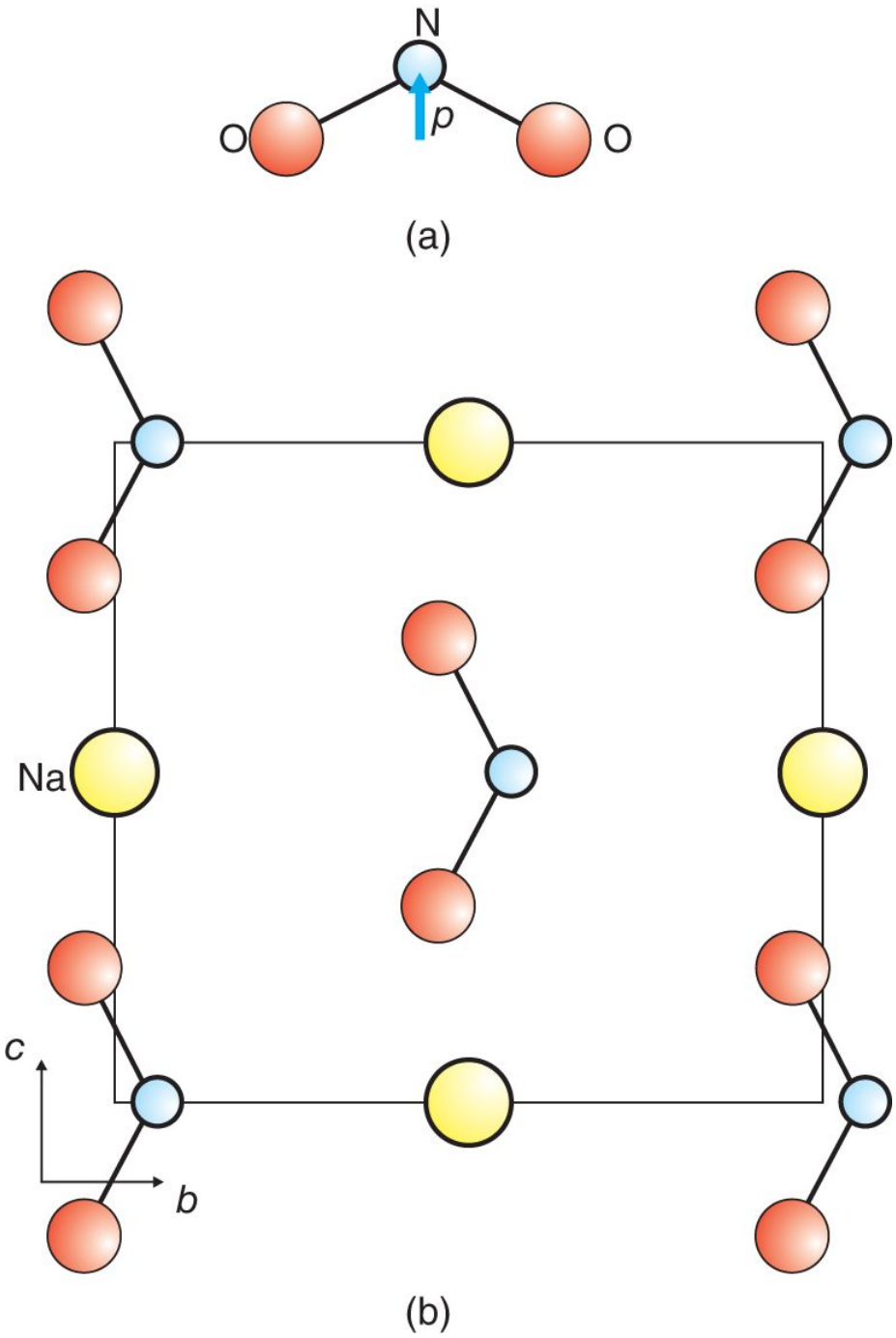


Figure 9.26 Ferroelectricity in sodium nitrite, $NaNO_2$: (a) the planar nitrite (NO_2^-) group; (b) the low temperature structure of $NaNO_2$ projected down [001].

The ferroelectric to paraelectric phase transition occurs at 165 °C, and in the paraelectric phase the net dipole moment has been lost. Again, there are two possibilities, either free rotation of the NO_2^- groups could occur or they could simply directionally disorder. In reality the high temperature structure is disordered with half of the NO_2^- dipoles point along +**b** and half point along –**b**. This transition is again an

order–disorder transition, and its onset is gradual. At 150 °C, 15° below the transition temperature, some 10% of the NO_2^- groups have reversed their orientation.

9.3.6 Ferroelectricity Due to Medium-Sized Transition-Metal Cations

Many oxides with structures containing medium sized cations are important ferroelectrics. From this large body of materials, the perovskite oxides (Figures 3.11 and 3.30) with general formula ABO_3, where A is a large cation such as Ca^{2+} and B is a medium sized cation such as Ti^{4+}, can be used to outline the atomic mechanism leading to ferroelectric behaviour. The medium sized cations in these phases are octahedrally coordinated by six O^{2-} ions. At high temperatures the octahedra are regular with the cation in a central position and the structure is cubic. A displacive transition takes place at the Curie temperature, below which the cations are usually moved from the centre of the surrounding oxygen coordination polyhedron thus creating an electric dipole. They can jump from one off-centre position to another under the influence of an electric field, allowing the dipolar direction to be switched. In general, the transition has been attributed to the fact that the B cation is rather too small to fit into the oxygen octahedron and so 'rattles around' and this is the root of the lower temperature displacement. This explanation is not entirely correct, however, and there is no doubt that covalent bonding between the B cation and surrounding oxygen atoms is also involved.

The situation can be illustrated with reference to the archetypal ferroelectric material, barium titanate, $BaTiO_3$. In this phase, the paraelectric form of $BaTiO_3$ with the cubic perovskite structure ($a \approx 0.4018$ nm, depending on the temperature of measurement) is found above 398 K (Figure 9.27a,b). The large Ba^{2+} cations are surrounded by 12 oxygen ions and the medium sized Ti^{4+} ions are situated at the centre of an octahedron of oxygen ions with equal bond lengths of approximately 0.2 nm. Between 398 and 278 K the unit cell is tetragonal ($a \approx 0.3956$ nm, $c \approx 0.4035$ nm) and ferroelectric. As the crystal cools through the cubic – tetragonal transition temperature, the cubic cell expands slightly along one edge to produce the tetragonal **c**-axis and is slightly compressed along the other two edges to form the tetragonal **a**- and **b**-axes. The change from cubic to tetragonal is accompanied by an off-centre movement of the octahedrally coordinated Ti^{4+} ions along the +**c**-axis, accompanied by a slight change in octahedron dimensions so that two equatorial oxygen atoms move parallel to the +**c**-axis and two move in the opposite direction (Figure 9.27c). The O^{2-}□Ti^{4+} bond lengths parallel to the **c**-axis are now 0.22 and 0.18 nm, while the equatorial bond lengths remain at 0.2 nm. This results in the formation of a dipole pointing along the **c**-axis (Figure 9.27d) with a net dipole moment of the order of 26 μC cm^{-2}. The change in the Ba^{2+} positions is almost negligible. The polar axis is the **c**-axis. The off-centre Ti^{4+} position and the octahedral deformation can be changed in an electric field and hence tetragonal $BaTiO_3$ is a ferroelectric. There is no preference as to which of the original cubic axes becomes the polar direction, and so this can take one of 6 equiv directions, parallel to ± **x**, ± **y**, or ± **z**. On cooling, a ferroelectric domain pattern forms, reflecting this symmetry.

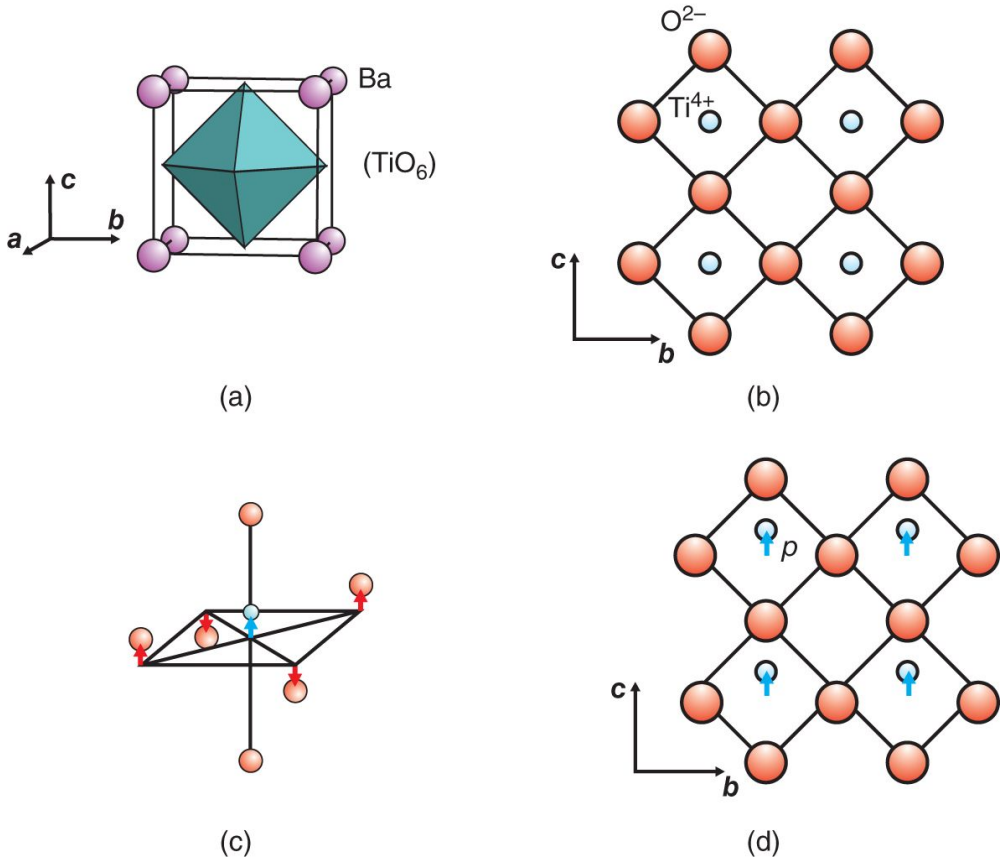


Figure 9.27 The barium titanate, $BaTiO_3$, structure: (a) cubic paraelectric high-temperature structure; (b) cubic structure projected down [100] with Ba^{2+} cations omitted; (c) ion displacements in the tetragonal ferroelectric phase, exaggerated; (d) projection of the tetragonal phase down [100], ion displacements exaggerated.

Between the temperatures 278–183 K, the tetragonal structure undergoes a further displacive transition resulting in an elongation along a face diagonal of the unit cell to give an orthorhombic phase ($a \approx 0.3987$ nm, $b \approx 0.5675$ nm, $c \approx 0.5690$ nm). Below 183 K the structure another displacive transition causes the unit cell to become rhombohedral ($a \approx 0.400$ nm, $\alpha \approx 89.86°$).

These displacements and distortions are rather small – a feature that applies to many ferroelectric perovskites. Structural relationships between the various crystallographic unit cells are then often simplified by reference to a *pseudocubic* structure. For example, if the polarisation **P** lies along an octahedron axis, it is simpler to refer to this as the pseudocubic <100> direction in all structures rather than apparently different directions when referred to the true tetragonal, orthorhombic, or rhombohedral cells.

9.3.7 Modification of Properties

A ferroelectric crystal such as $BaTiO_3$ does not normally show any observable polarisation because the domain structure leads to overall cancellation of the effect. Polycrystalline ceramics would be expected to be similar but the majority of ferroelectric materials used are, in fact, polycrystalline. In order to induce an

observable polarisation in a polycrystalline material the crystals are *poled*. This process involves heating the crystals above the Curie point and then cooling them in a strong electric field. The effect of this is to favourably orient dipoles so that the polycrystalline ceramic shows a ferroelectric effect.

The same is true of polymer piezoelectrics. In these materials, the crystallites that give rise to piezoelectricity are oriented at random within the polymer matrix. Poling can give the dipoles an overall preferred orientation. Poling will not affect a crystallite that does not show a permanent dipole, and so poling only applies to pyroelectric and ferroelectric materials.

Many ferroelectric materials show interesting and potentially useful properties, but not at the temperature or pressure required for a particular application. It is then necessary to *tune* the property to the fit the application. This modification is frequently brought about by the replacement of one or more of the constituents of the compound, or by the deliberate addition of impurities, i.e. *doping*.

Ferroelectric oxides, for example, are of interest as capacitor materials because of their high relative permittivity values, but usually the sharp maximum in dielectric constant at the Curie point must be broadened and moved to room temperature. For instance, $BaTiO_3$ has a high relative permittivity at the Curie temperature, about 393 K. The Curie temperature can be increased by the replacement of some of the Ba^{2+} ions by Pb^{2+} ions. These ions are 'softer' (i.e. more easily polarised) than the Ba^{2+} ions, as they have a lone pair of electrons, and so are more easily affected by an applied electric field. The resultant compound retains the crystal structure of $BaTiO_3$, but has a formula $Ba_{1-x}Pb_xTiO_3$. The compound $Ba_{0.6}Pb_{0.4}TiO_3$ has a Curie temperature of approximately 573 K, an increase of 200 K. In a similar way, the Curie temperature can be lowered by the substitution of Ba^{2+} ions by Sr^{2+}. These ions are smaller than Ba^{2+} ions and can be considered to be 'harder' and more difficult to polarise. The compound $Ba_{0.6}Sr_{0.4}TiO_3$ has a Curie temperature of 0 °C. The Curie temperature can also be lowered by the replacement of some of the Ti^{4+} ions by Zr^{4+} or Sn^{4+} ions. In this type of doping the off-centre M^{4+} ions in MO_6 octahedra are modified.

The consequences of doping ferroelectric materials are best understood in terms of the phase diagrams of the relevant systems, such as that for the $PbZrO_3$–$PbTiO_3$ (PZT) system (Figure 9.28). The majority of the phases are ferroelectric, with the exceptions of the high temperature cubic phase and the orthorhombic phase at the $PbZrO_3$ end of the diagram, which is antiferroelectric. It is found experimentally that the ferroelectric properties of these phases, which vary smoothly across the composition ranges, is often significantly enhanced in the neighbourhood of *morphotropic phase boundaries* (MPBs), which are boundaries within a broad solid solution range where the symmetry changes. These often occur within solid solutions formed between symmetrically dissimilar end members. For example, the boundary between the rhombohedral and tetragonal phases in the $PbZrO_3$–$PbTiO_3$ (PZT) system (Figure 9.28) is an MPB.

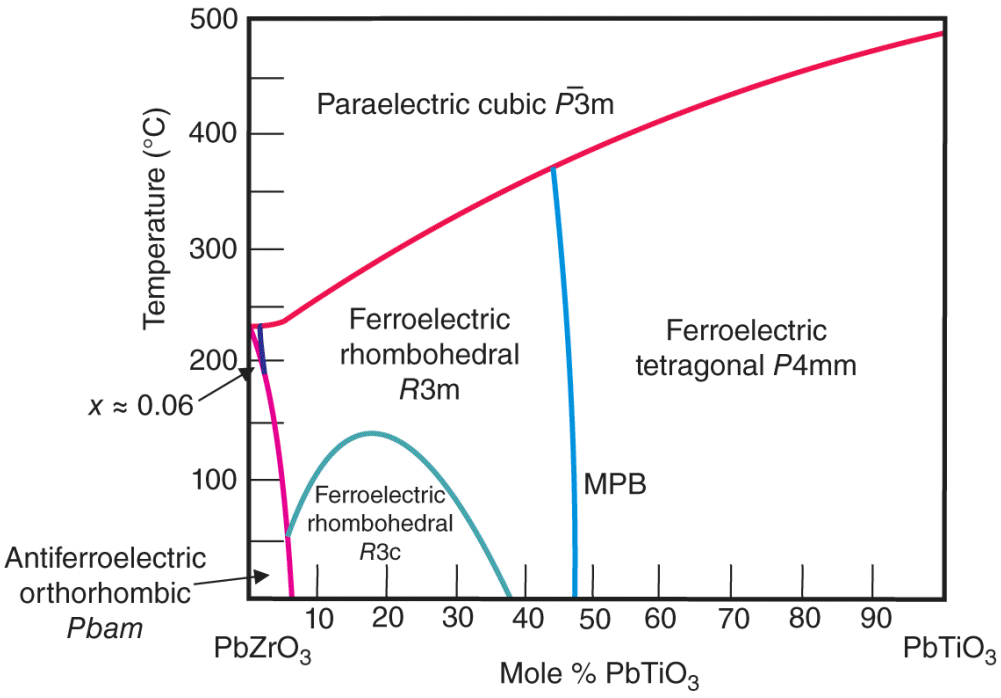


Figure 9.28 The phase diagram of the $PbZrO_3$–$PbTiO_3$ (PZT) system. The space group and ferroelectric status of each phase is marked. MPB: morphological phase boundary.

Although these diagrams display the overall dielectric behaviour of the phases, *measured* properties are always very dependent upon preparation methods, and are strongly influenced by extrinsic variables such as grain size and impurities. Sintering and the careful removal of impurity phases still remain central to the successful fabrication of polycrystalline ceramic samples with desired piezoelectric and ferroelectric characteristics.

9.3.8 Relaxor Ferroelectrics

Relaxor ferroelectrics are a group of ferroelectric perovskite structure materials with a general formula $Pb(B_1\ B_2)O_3$ in which B_1 is typically Mg^{2+}, Zn^{2+}, Ni^{2+}, Fe^{3+}, Sc^{3+}, and In^{3+} and B_2 is typically Nb^{5+}, Ta^{5+}, and W^{6+}. They are represented by $Pb(Mg_{1/3}Nb_{2/3})O_3$ (PMN) and especially the solid solution series $Pb(Mg_{1/3}Nb_{2/3})O_3$–$PbTiO_3$ (PMN–PT) and $Pb(Mg_{1/3}Nb_{2/3})O_3$–$Pb(Zr_{0.52}Ti_{0.48})O_3$ (PMN–PZT). They differ from the ferroelectric materials described earlier in the following ways: (i) the relative permittivity shows a wide *diffuse*, *frequency dependent* transition over a broad temperature interval with a maximum value at a temperature T_m rather than a sharp transition at a temperature T_C (Figure 9.29a); (ii) the polarisation of the solid exhibits a *diffuse* transition near to T_m and does not fall to zero at a temperature T_C (Figure 9.29b). There is no Curie–Weiss behaviour above T_m; (iii) the hysteresis loop tends to be narrow with low value for the remnant polarisation (Figure 9.29c); and (iv)

the crystal structure does *not change* significantly at T_m while in a normal ferroelectric there is a noticeable change of symmetry, easily revealed by X-ray diffraction.

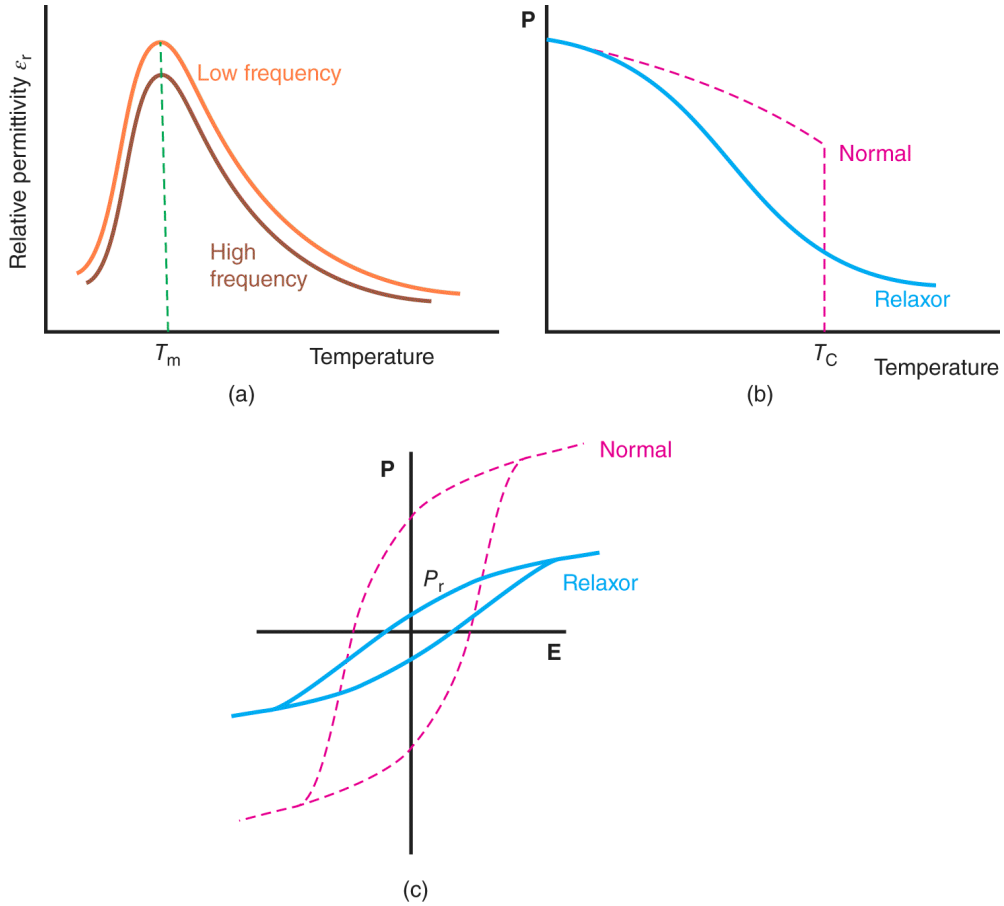


Figure 9.29 Relaxor ferroelectric properties, schematic: (a) variation of relative permittivity with temperature and measurement frequency; (b) variation of polarisation with temperature; (c) variation of polarisation with electric field (hysteresis).

These differences are attributed to the different *microstructure* of relaxor ferroelectrics compared with a normal ferroelectric such as $BaTiO_3$. In normal materials the domain size is large, often identical to the grain size in a ceramic sample, and each domain is uniform. In relaxor ferroelectrics the B sites are occupied by two cations, B_1 and B_2. These are not homogeneously distributed so that an ordinary domain may be made up of regions with slightly differing ordering between the B_1 and B_2 cations, leading to compositional disorder at an atomic scale. These inhomogeneous volumes are known as *nanodomains*, *microdomains*, or *polar microregions* (PMRs). In effect they destroy long-range order and each behaves as a slightly different ferroelectric. Every microdomain will have a slightly different value of T_C to the others and this smears out the sharp transitions seen in normal ferroelectrics and gives rise to the broad maximum centred on T_m. Similarly, because of the spread of T_C values across the microdomains, a certain degree of polarisation will persist above T_m in a relaxor ferroelectric.

Statistical mechanics suggests that the domain sizes will follow a temperature sensitive distribution with high temperatures giving rise to larger numbers of smaller volume microdomains. A decrease in temperature, allowing for equilibrium to establish, will tend to result in fewer and larger microdomains. Thus, relaxor behaviour is to some extent tunable by varying processing temperature and cooling rate.

9.3.9 Ferroelectric Nanoparticles, Thin Films, and Superlattices

In bulk ferroelectrics the polarisation arises from dipoles within the interior of the solid. However, as size decreases surface effects become pronounced and can lead to decrease or increase in the distortions of the polyhedral components of the phase, due to surface energy contributions. This means that distortions such as cation centring in anion polyhedral may be blocked in the bulk but become allowed at surfaces or vice versa. Moreover, it might be suspected that as the particle size of a ferroelectric decreased the spontaneous polarisation would eventually vanish as the long-range order characteristic of the ferroelectric phase is no longer supported. Thus, with nanoparticles or very thin films the question arises as to whether ferroelectricity and domain structures are still preserved and will new features emerge as a result of surface interference. Moreover, it must also be remembered that at these small dimensions, especially with thin films, these are supported on a substrate, and substrate–ferroelectric interactions may also play an important role in the properties measured.

Evidence now suggests that ferroelectric properties persist in small particles prepared by milling of bulk material. It appears, though, that at dimensions below about 5 nm, the value of T_C and polarisation begin to diminish. In the case of thin films, a similar picture emerges. Films of lead titanate appear to remain ferroelectric down to one or two unit cells (0.4–1 nm) in thickness, although the Curie temperature gradually falls for the thinnest films suggesting that below this critical size matters may change. Barium titanate nanoparticles show polarisation and switching in particles down to approximately 5 nm diameter. It is beginning to appear that a dimension of the order of 5 nm represents a critical size for ferroelectric phenomena.

Ferroelectric superlattices are alternating layers, a few unit cells thick, of two or three perovskite structure materials, grown sequentially so that a perfect crystal slab is produced. To achieve this, layers are built up atom by atom, thus allowing control of crystal composition and perfection as well as layer thickness. As anticipated, the ferroelectric properties of these superlattices can be tuned by varying the chemical constituents (ABO_3, $A'B'O_3$) of the layers and also by changing the relative proportions of each layer, m unit cells of ABO_3 and n unit cells of $A'B'O_3$ (Figure 9.30). However, tuning in another, rather unexpected way, using strain, can also be achieved. This possibility arises because the lattice parameters of the perovskite unit cells are close but not identical, so that degrees of strain exist at the interfaces that are related to the difference between the two lattice parameters in the interface plane. Thus, interfacial strain has an important role to play in the properties of such superlattices and a number of unique effects have been attributed to this cause. For example, a superlattice formed of the non-ferroelectric perovskite oxides $SrTiO_3$ and $CaTiO_3$ enclosing thin sheets of ferroelectric $BaTiO_3$ shows a much higher degree of polarisation than $BaTiO_3$

alone, attributed to the strain at the ferroelectric–nonferroelectric interfaces, which produces an enhanced polarisation of the TiO_6 octahedra in the $BaTiO_3$ component.

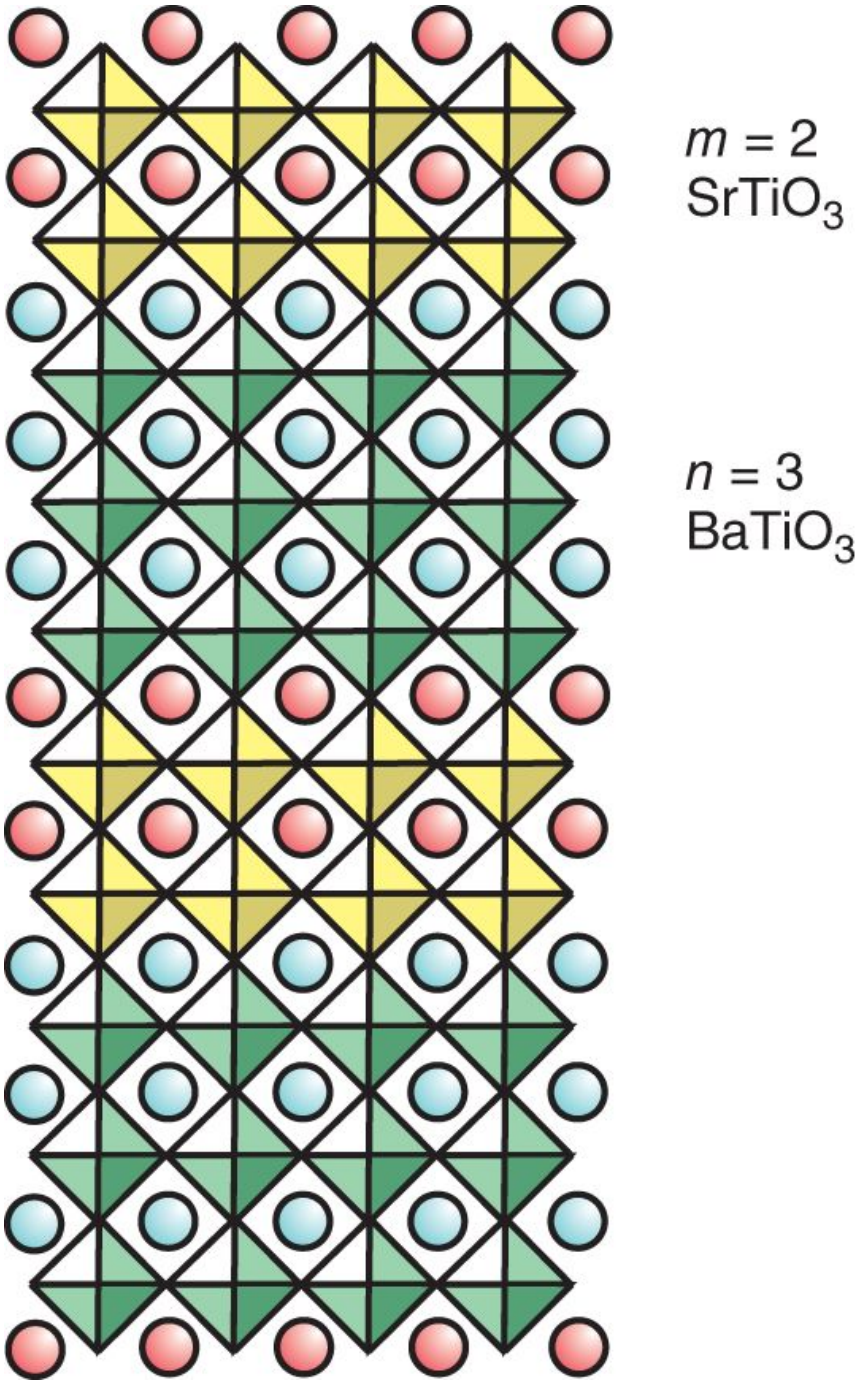


Figure 9.30 A ferroelectric superlattice composed of $m = 2$ unit cells of $SrTiO_3$ and $n = 3$ unit cells of $BaTiO_3$.

Interfacial strain in the superlattice series formed by m unit cells of ferroelectric $PbTiO_3$ and n unit cells of dielectric $SrTiO_3$ also produces curious behaviour. Normal ferroelectric behaviour is found when the layers are relatively thick. This diminishes as

layer thicknesses fall but surprisingly, at the lowest values, ferroelectricity recovers. In bulk $PbTiO_3$ octahedral tilt is supressed and in $SrTiO_3$ oxygen rotation is supressed. However, in superlattices these distortions become possible, creating a strain between the two perovskite slices. As the slabs become thin the strain component of the interfaces becomes relatively greater and ultimately, in the thinnest layers, is able to induce polarisation and an increased ferroelectric response.

A similar strain induced feature is the development of ferroelectricity in superlattices formed by two *non-ferroelectric* components. This has been found, for instance, in the 1 : 1 superlattices $LaGaO_3/YGaO_3$ and $LaAlO_3/YAlO_3$. The bulk perovskite structures of these oxides contain slightly rotated octahedral GaO_6 or AlO_6 groups but they do not lead to permanent polarisation and the solids are simply dielectrics. However, the construction of thin layer superlattices relaxes the bulk constraints so that in the interfacial regions the rotation effects become unbalanced to such an extent that permanent polarisation is introduced, resulting in measurable ferroelectricity in the superlattice.

These last two features are known as *improper ferroelectricity* (perhaps better *extrinsic ferroelectricity*, see Lines and Glass, in Further Reading). In essence, improper ferroelectricity is ferroelectricity which is not due to the normal polarisation of the structure but arises from other interactions. Improper ferroelectricity is rare in bulk materials and is a weak effect, usually rather difficult to detect. The creation of artificial superlattices in which manipulation of interfacial strain energy is possible has allowed improper ferroelectrics to become accessible.

9.3.10 Flexoelectricity in Ferroelectrics

There is a strong link between the polarisation of a ferroelectric and deformation within the crystal structure of the ferroelectric phase. The deformation leads to the production of dipoles that can be switched by the application of an electric field, which, in effect, switches the deformation. Mechanical stress will deform a solid and the question then arises as to whether it is possible to switch polarisation in a ferroelectric by the application of stress. The connection between these physical properties is called *flexoelectricity*, and describes the coupling between strain gradient (the result of the application of a stress) and polarisation. This is of potential importance because the if the polarisation state of a ferroelectric film can be changed mechanically then the process can be used to create a permanent domain array that will serve as a memory store that can be mechanically written but read or erased electrically. In particular, if the switched domains are of nanometre dimensions, the amount of information that can be stored will be significant.

Computation shows that the crucial feature to enable mechanical switching is the creation of a *strain gradient* in the ferroelectric. A homogeneous stress simply shifts the double potential well associated with ferroelectricity (Figure 9.17) without changing the shape of the curve. On the other hand, a strain gradient alters the form of the curve so that one configuration is preferred over the alternative (Figure 9.31a). The switching of the polarisation direction can then be envisaged to occur in a thin film if a force can be applied to a point location on the film surface. This is because, in a thin film, the strain gradient will be sufficiently extended to penetrate he film and hence completely switch the strained volume. (In a bulk sample the strain field will only penetrate a small

way into the crystal and on removal of the strain the surroundings will force the polarisation to revert to its original direction.) Flexoelectric domain writing has been achieved in $BaTiO_3$ films approximately 4.8 nm (12 unit cells) thickness, oriented with the polarisation vector perpendicular to the surface. The initial polarisation was reversed by forcing the probe of a scanning probe microscope into the crystal surface, creating a domain with dimensions of the order of 30 nm diameter (Figure 9.31b). The resultant mechanically written domain pattern can be both read and erased electrically, thus generating a mechanically written permanent memory store.

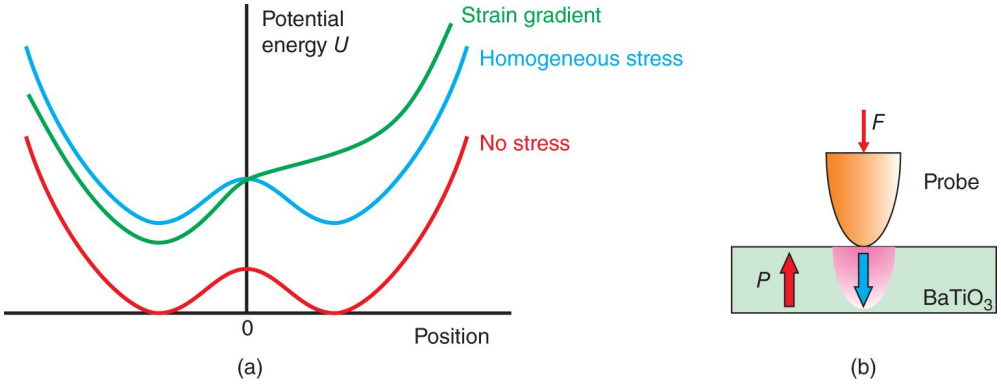


Figure 9.31 The flexoelectric effect: (a) evolution of the potential energy curve for a ferroelectric under a homogeneous stress and strain gradient; (b) switching via mechanical stress imposed by a tip.

Source: Adapted from Lu et al. (2012).

Reference

Flexoelectric Effect

Lu, H., Bark, C.-W., de los Ojos, D.E. et al. (2012). Mechanical writing of ferroelectric polarization. *Science***336**: 59–61.

Further Reading

General:

Lines, M.E. and Glass, A.M. (2001). *Principles and Applications of Ferroelectrics and Related Materials*, Oxford Classics. Oxford: Oxford University Press.

Introductory Crystallography with Respect to Dielectric Properties:

Tilley, R.J.D. (2020). *Crystals and Crystal Structures*, 2nde. Chichester: John Wiley & Sons Chapter 5.

The Dielectric, Piezoelectric and Ferroelectric Properties of Perovskite Structures are Detailed in:

Tilley, R.J.D. (2016). *Perovskites*. Chichester: John Wiley & Sons, Chapter 6.

Biomolecular Materials are Described in:

Kim, D., Han, S.A., Kim, J.H. et al. (2020). Biomolecular piezoelectric materials: from amino acids to living tissues. *Adv. Mater.***32**: 1906989.

Nanoparticles, Thin Films and Superlattices:

Ahn, C.H., Rabe, K.M., and Triscone, J.-M. (2004). Ferroelectric superlattices. *Science***303**: 488–491.

Baek, S.H., Park, J., Kim, D.M. et al. (2011). Giant piezoelectricity on Si for hyperactive MEMS. *Science***334**: 958–961.

Bousquet, E., Dawber, M., Stucki, N. et al. (2008). Improper ferroelectricity in perovskite oxide artificial superlattices. *Nature***452**: 732–736.

Cheema, S.S., Kwon, D., Shanker, N. et al. (2020). Enhanced ferroelectricity in ultrathin films grown directly on silicon. *Nature***580**: 478.

Dawber, M., Stucki, N., Lichtensteiger, C. et al. (2007). Tailoring the properties of artificially layered ferroelectric superlattices. *Adv. Mater.***19**: 4153–4159.

Liu, H., Wu, H., Ong, K.P. et al. (2020). Giant piezoelectricity in oxide thin films with nanopillar structure. *Science***369**: 292.

Nelson, C.T., Gao, P., Jokisaari, J.R. et al. (2011). Domain dynamics during ferroelectric switching. *Science***334**: 968–971.

Rondinelli, J.M. and Fennie, C.J. (2012). Octahedral rotation-induced ferroelectricity in cation ordered perovskites. *Adv. Mater.***24**: 1961–1968.

Rørvik, P.H., Grande, T., and Einarsrud, M.-A. (2011). One-dimensional nanostructures of ferroelectric perovskites. *Adv. Mater.***23**: 4007–4034.

Hwang, H.Y., Iwasa, Y., Kawasaki, M. et al. (2012). Emergent phenomena at oxide interfaces. *Nat. Mater.***11**: 103–113.

Problems and Exercises

Quick Quiz

9.1 The relative permittivity of a material is also called the:

a. Dielectric constant.

b. Dielectric permittivity.

c. Dielectric susceptibility.

9.2 The polarisation of a solid in an electric field is due to:

a. The formation of charges on the surface.

b. The flow of charges from one surface to the other.

c. The formation of electric dipoles.

9.3 The dielectric susceptibility relates:

a. Polarisation of a solid to the charges in a solid.

b. Polarisation to the capacitance of a solid.

c. Polarisation of a solid to the electric field.

9.4 Which of the following contributes to the polarisability of a solid in a very high frequency electric field?

a. Ionic polarisability.

b. Electronic polarisability.

c. Space-charge polarisability.

9.5 The relative permittivity of a crystal with tetragonal symmetry is:

a. The same along the **a**-, **b**-, and **c**-axes.

b. The same along **a**- and **b**-, and different along **c**-.

c. Different along the **a**-, **b**-, and **c**-axes.

9.6 The relative permittivity of a crystal with orthorhombic symmetry is:

a. The same along the **a**-, **b**-, and **c**-axes.

b. The same along **a**- and **b**-, and different along **c**-.

c. Different along the **a**-, **b**-, and **c**-axes.

9.7 In the direct piezoelectric effect:

a. An applied voltage causes a dimensional change.

b. A dimensional change produces a voltage.

c. An applied voltage produces a temperature change.

9.8 A spontaneous polarisation is NOT found in:

a. Piezoelectric solids.

b. Ferroelectric solids.

c. Pyroelectric solids.

9.9 A solid that has a switchable spontaneous polarisation is:

a. A pyroelectric.

b. A piezoelectric.

c. A ferroelectric.

9.10 Electrets are:

a. Polymer sheets with a permanent surface charge.

b. Polymer sheets with permanent internal dipoles.

c. Polymer sheets that are charged easily.

9.11 The crystal structure of pyroelectric crystals must contain:

a. A polar axis.

b. A centre of symmetry.

c. Switchable permanent dipoles.

9.12 Ferroelectric crystals must possess:

a. Hydrogen bonds.

b. Switchable dipoles.

c. Polar groups.

9.13 Hysteresis is characteristic of:

a. Pyroelectric crystals.

b. Piezoelectric crystals.

c. Ferroelectric crystals.

9.14 The dipoles in an antiferroelectric crystal are:

a. In parallel rows.

b. In antiparallel rows.

c. In randomly aligned rows.

9.15 The paraelectric state of a ferroelectric is:

a. The high temperature phase.

b. The low temperature phase.

c. The antiferroelectric phase.

9.16 The Curie temperature is NOT the temperature at which:

a. A ferroelectric transforms to a paraelectric state.

b. A paraelectric transforms to a ferroelectric state.

c. A ferroelectric transforms to an antiferroelectric state.

9.17 The temperature dependence of the relative permittivity of many ferroelectric crystals obeys the Curie–Weiss law:

a. In the ferroelectric state.

b. In the paraelectric state.

c. At low temperatures.

9.18 Hydrogen bonding is NOT the cause of ferroelectricity in:

a. Triglycine sulfate.

b. Rochelle salt.

c. Sodium nitrite.

9.19 The cause of ferroelectricity in perovskite materials is often due to:

a. The presence of medium sized cations in octahedral coordination.

b. The presence of hydrogen bonds.

c. The presence of polar groups.

9.20 The process by which suitable polycrystalline solids can be made ferroelectric is:

a. Annealing.

b. Sintering.

c. Poling.

Calculations and Questions

9.1 1 The plates on a parallel plate capacitor are separated by 0.1 mm and filled with air.

a. What is the capacitance if the plates have an area of 1 cm^2?

b. If the space between the plates is filled with a polyethylene sheet, with a relative permittivity of 2.3, what is the new capacitance?

9.2 A parallel plate capacitor is connected to a battery and acquires a charge of 200 μC on each plate. A polymer is inserted and the charge on the plates is now found to be 750 μC. What is the relative permittivity of the polymer?

9.3 A parallel plate capacitor has a capacitance of 8 nF and is to be operated under a voltage of 80 V. What is the relative permittivity of the dielectric if the maximum dimensions of the capacitor are 1 mm × 1 cm × 1 cm?

9.4 The dipole moment of the molecule nitric oxide, NO, is 0.5×10^{-30} C m. The N□O bond length is 0.115 nm.

a. What is the charge on the atoms?

b. Which atom is more positive?

9.5 The dipole moment of a water molecule is 6.2×10^{-30} C m and the H□O bond length is 95.8 pm and the H□O□H bond angle is 104.5°. Determine the charge on

each atom.

9.6 The dipole moment of the molecule HCN is 9.8×10^{-30} C m and the charges, which reside on the H and N atoms (with carbon central), are measured to be 3.83×10^{-20} C.

a. What is the dipole length and hence the approximate length of the molecule?

b. The molecules are packed into a cubic structure, with a lattice parameter of 0.512 nm; each unit cell containing one molecule at 0, 0, 0. Determine the bulk polarisation of the solid when all dipoles are aligned.

9.7 The dipole moments of the following molecules are:

carbon monoxide, CO (linear) 0.334×10^{-30} C m; nitrous oxide, N_2O (linear) 0.567×10^{-30} C m; ammonia, NH_3 (tetrahedral, N at one vertex, H–N–H = 106.6°), 4.837×10^{-30} C m; sulfur dioxide, SO_2 (angular, O–S–O = 119.5°), 5.304×10^{-30} C m.

The covalent radii of the atoms involved are: C, 0.077 nm; O, 0.074 nm; N, 0.074 nm; H, 0.037 nm; S, 0.104 nm. Calculate the nominal charges on the atoms as a fraction of the electron charge.

9.8 Derive the Clausius–Mossotti equation using equations:

$$\mathbf{P} = (\varepsilon_r - 1)\,\varepsilon_0 \mathbf{E}_0$$

$$\mathbf{P} = N\,\alpha \mathbf{E}_{loc}$$

$$\mathbf{E}_{loc} = \mathbf{E}_0 + \mathbf{P}/3\,\varepsilon_0$$

9.9 The following data is for a single crystal of MgO; relative permittivity, 9.65; refractive index, 1.736; unit cell, cubic, a = 0.4207 nm, Z = 4 formula units of MgO. Estimate:

a. the electronic and ionic polarisability volumes;

b. the electronic and ionic polarisabilities of MgO.

9.10 The following data is for a single crystal of α-quartz; average relative permittivity, 4.477; average refractive index, 1.5485; unit cell, hexagonal, a = 0.49 136, c = 0.54 051 nm, Z = 3 formula units of SiO_2. Although the Clausius-Mossotti and Lorentz-Lorenz relations apply only approximately for this symmetry, estimate:

a. the electronic polarisability volume and corresponding polarisability,

b. the ionic polarisability volume and the corresponding polarisability.

9.11 Use the additivity rule to estimate (a) the ionic polarisability, (b) the electronic polarisability, (c) the total polarisability of forsterite, Mg_2SiO_4, given the information in Questions (9.9) and (9.10), assuming that only ionic and electronic polarisations are important.

9.12 The crystallographic and optical data for forsterite are: orthorhombic, a = 0.4758 nm, b = 1.0214 nm, c = 0.5984 nm, Z = 4 formula units of Mg_2SiO_4; the three principle refractive indices are 1.635, 1.651, 1.670. Estimate the electronic polarisability of the mineral.

9.13 Estimate the polarisability and relative permittivity of the garnet $Ca_3Ga_2Ge_3O_{12}$, using the following data: $Ca_3Ga_2Ge_3O_{12}$, cubic, a = 1.2252 nm, Z = 8; CaO, polarisability volume $\alpha' = 5.22 \times 10^{-30}$ m^3; Ga_2O_3, polarisability volume $\alpha' = 8.80 \times 10^{-30}$ m^3; GeO_2, polarisability volume $\alpha' = 5.50 \times 10^{-30}$ m^3.

9.14 The relative permittivity of the garnet $Y_3Fe_5O_{12}$ was measured as 15.7. Calculate the polarisability volume and polarisability of Y_2O_3, if the polarisability volume of Fe_2O_3 is 10.5×10^{-30} m^3. The unit cell is cubic, a = 1.2376 nm, Z = 8 formula units of $Y_3Fe_5O_{12}$.

9.15 Use the additivity rule to estimate (a) the polarisability volume; (b) the polarisability of mullite, Al_2SiO_5. The following polarisability volumes were found in the literature. SiO_2, 4.84×10^{-30} m^3, Al_2O_3, 7.70×10^{-30} m^3. (c) The experimental value is 15.22×10^{-30} m^3. Comment on the accuracy of the method.

9.16 Use the additivity rule to estimate the (a) polarisability volume; (b) the polarisability of diopside, $CaMgSi_2O_6$. The following polarisability volumes were found in the literature. Magnesium oxide, MgO, 3.32×10^{-30} m^3; SiO_2, 4.84×10^{-30} m^3; CaO, 5.22×10^{-30} m^3. (c) The experimental value is 18.78×10^{-30} m^3. Comment on the accuracy of the method.

9.17 The polarisability for the oxide $Mn_3Al_2Si_3O_{12}$ was estimated to be 35.83×10^{-40} C m^2 V^{-1}. Determine the value of the polarisability for Mn^{2+} ions in this oxide given the data: α (Al^{3+}) 0.32×10^{-40} C m^2 V^{-1}; α (Si^{4+}) 0.11×10^{-40} C m^2 V^{-1}, α (O^{2-}) 2.64×10^{-40} C m^2 V^{-1}.

9.18 Show that the units for the direct piezoelectric coefficient, d, of C N^{-1} and m V^{-1} are equivalent.

9.19 The value of the piezoelectric coefficient for quartz is given as 2.3 pC N^{-1}. Calculate the polarisation of a plate of dimensions 10 cm × 5 cm × 0.5 mm when a mass of 0.5 kg is placed on it.

9.20 An electret film has a piezoelectric coefficient of 170 pC N^{-1}. Calculate the change in thickness when 500 V are applied across a film 0.1 mm thick.

9.21 The semiprecious gemstone tourmaline, with an approximate formula $CaLi_2Al_7(OH)_4(BO_3)_3Si_6O_{18}$, has a pyroelectric coefficient, π_i, of 4×10^{-6} C m^{-2} K^{-1}. The unique polar axis is the crystallographic **c**-axis. What is the change in polarisation caused by a change of temperature of 100 °C?

9.22 The measured relative permittivity, ε_r, of a ceramic sample of $PbZrO_3$ as a function of temperature, T °C, is given in the following table. Determine (a) the

Curie temperature, (b) the Curie constant for this sample.

ε_r	T (°C)
130	50
142	100
166	150
222	200
360	225
420	230
472	234
556	235
775	236
3200	238
3000	240
2840	242
2440	250
1620	275
1240	300
840	350

9.23 The measured relative permittivity, ε_r, of a ceramic sample of $Cd_2Nb_2O_7$ as a function of temperature, T °C, is given in the following table. Determine the extrapolated Curie temperature and the Curie constant, C, for this sample.

ε_r	T (°C)
4500	−80
4125	−75
3750	−70
3500	−65
3225	−60
3000	−55
2800	−50
2600	−45
2465	−40
2280	−35
2115	−30
2000	−25
1860	−20
1750	−15
1630	−10

1560	−5

9.24 The measured relative permittivity, ε_r, of a crystal of triglycine sulfate as a function of temperature, T °C, is given in the following table. Determine the Curie temperature and the Curie constant for triglycine sulfate.

ε_r	T (°C)
120	40
190	41
280	42
400	43
540	44
730	45
1300	46
~7000	47
1100	48
830	49
700	50
590	51
520	52
460	53
420	54
380	55
250	60
180	65
130	70
110	75

9.25 Zinc oxide, which is isostructural with wurtzite, Figure 9.8, has a hexagonal unit cell with a = 0.3250 nm, c = 0.5207 nm, Z = 2. The atom positions are:

Zn: ⅓, ⅔, 0; ⅔, ⅓, ½

O: ⅓, ⅔, 0.3825; ⅔, ⅓, 0.8825

Estimate the maximum spontaneous polarisation of ZnO, assuming that the structure is ionic.

9.26 Calculate the maximum spontaneous polarisation of a crystal of sodium nitrite given that the unit cell is orthorhombic, with a = 0.360 nm, b = 0.575 nm, c = 0.535 nm, Z = 2 formula units of $NaNO_2$. The dipole moment of each N□O bond is 0.5×10^{-30} C m^{-1} and the O–N–O angle is 115°.

9.27

a. Calculate the dipole moment of a TiO_6 octahedron in tetragonal $BaTiO_3$, a = 0.3997 nm, c = 0.4031 nm, assuming that the compound is fully ionic and the Ti^{4+} ions are displaced by 0.012 nm along the **c**-axis of the unit cell.

b. Determine the maximum spontaneous polarisation under these conditions.

9.28

a. Calculate the dipole moment of an NbO_6 octahedron in tetragonal $KNbO_3$, a = 0.4002 nm, c = 0.4064 nm, assuming that the compound is fully ionic and the Nb^{5+} ions are displaced by 0.017 nm along the **c**-axis of the unit cell.

b. Determine the maximum spontaneous polarisation under these conditions.

9.29

a. Calculate the dipole moment of a TiO_6 octahedron in tetragonal $PbTiO_3$, a = 0.3899 nm, c = 0.4153 nm, assuming that the compound is fully ionic and the Ti^{4+} ions are displaced by 0.030 nm along the **c**-axis of the unit cell.

b. Determine the maximum spontaneous polarisation under these conditions.

9.30 Both silica glass and quartz, SiO_2, are composed of SiO_4 tetrahedra and neither material possesses a centre of symmetry. Why is silica glass not a piezoelectric, as is quartz?

Note

1 The SI units of α are C m^2 V^{-1}. The most commonly quoted units for polarizability in the literature are non-SI polarizability volumes, α', m^3, cm^3, or $Å^3$. To convert these units note that:

α (C m^2 V^{-1}) = 1.11 265 × 10^{-10} α' (m^{-3}) = 1.11 265 × 10^{-16} α' (cm^{-3}) = 1.11 265 × 10^{-30} α' ($Å^3$).

10
Magnetic Solids

- What is a paramagnetic material?
- What causes magnetic hysteresis?
- What materials are used for magnetic data storage?

Magnets, sometimes called *permanent* magnets, pervade everyday life. In reality permanent magnets are examples of *ferromagnetic* materials. If a small bar magnet (i.e. a small rod of a ferromagnetic substance) is freely suspended, it will align (approximately) north–south. The end pointing north is called the north pole of the magnet and the end pointing south, the south pole of the magnet. It is found that opposite magnetic poles attract each other and similar magnetic poles repel each other and this fact proves that the Earth acts as a magnet. (Because of this, the end of a freely suspended magnet that points towards the north should be labelled as a south pole, but it is too late to change things now!) A ferromagnetic solid behaves as if surrounded by a magnetic field, and will attract or repel other ferromagnets via the interactions of the magnetic fields.

Not only solids, but also wires carrying an electric current also give rise to magnetic fields, the strength of which is proportional to the current flowing. Most solids, however, including wires not carrying a current, and ferromagnetic materials above a certain temperature, are loosely termed non-magnetic. Strictly speaking, this is inaccurate, as these materials simply exhibit extremely weak magnetic effects.

Magnetism in a solid arises from interactions between the electrons present. For many purposes, electron–electron interactions can simply be regarded as a generalised repulsion, so that the behaviour of electrons in a solid can be regarded as something equivalent to an electron gas in which electron–electron interactions are minimal. In explaining magnetism and other phenomena, such as superconductivity (Section 11.5), the electrons are no longer more or

less independent, but the behaviour of the electrons is *strongly correlated*. That is, the behaviour of one electron is closely linked to the behaviour of others, giving rise to ferromagnetism, antiferromagnetism, ferrimagnetism, and so on, described in the following text.

10.1 Magnetic Materials

10.1.1 Characterisation of Magnetic Materials

The weak magnetic properties of most solids can be measured using a *Gouy balance* (Figure 10.1). In this equipment, the sample is suspended between the poles of an electromagnet from a sensitive balance. The vast majority of solids show only miniscule magnetic effects. Of these, most weigh slightly less when the electromagnet is on than when the electromagnet is turned off. These materials, which are weakly repelled by the magnetic field, are the *diamagnetic* materials. The rest of the 'non-magnetic' group weigh slightly more when the electromagnet is on than when it is off. These substances are drawn weakly into a magnetic field and are called *paramagnetic* materials. Ferromagnetic materials are strongly attracted to one or other of the pole pieces of the magnet and the technique does not give a result.

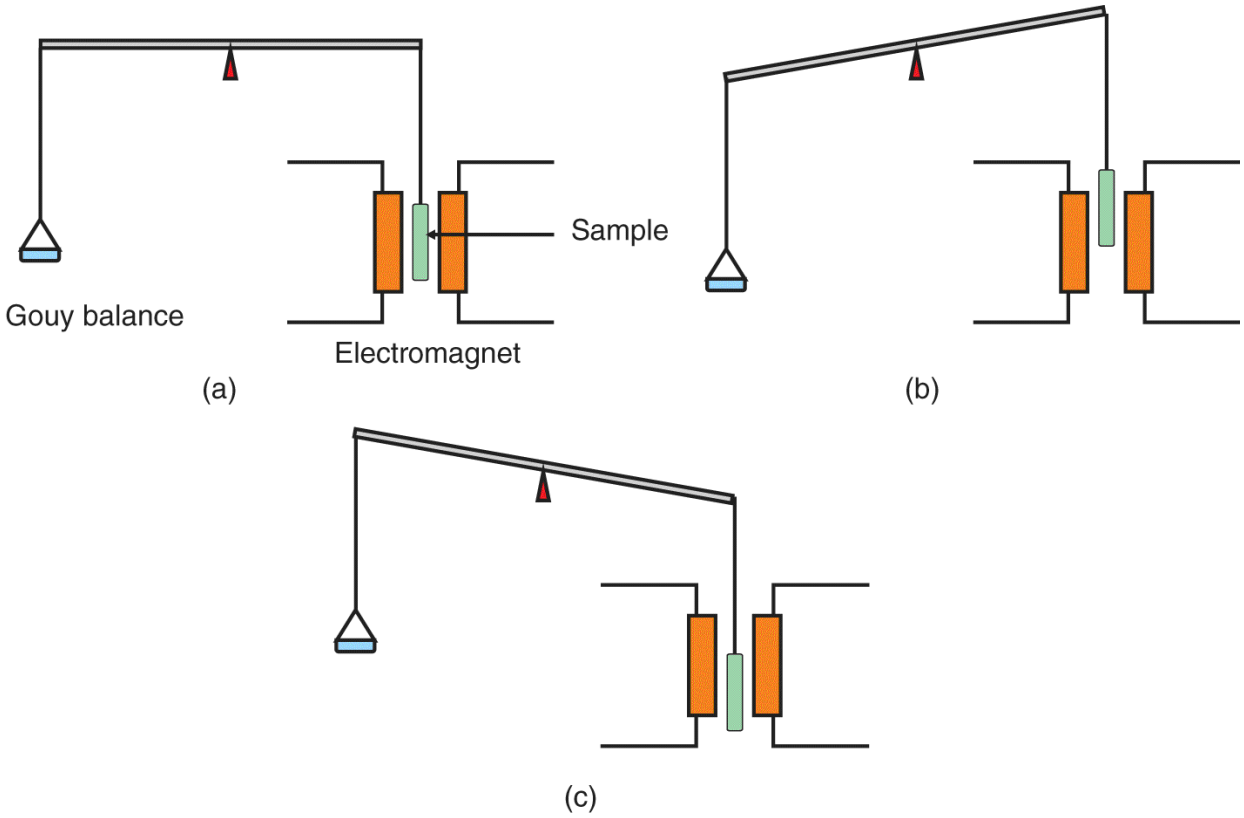


Figure 10.1 Magnetic materials in a Gouy balance: (a) electromagnet off; (b) a diamagnetic substance; (c) a paramagnetic substance.

Diamagnetic and paramagnetic substances are characterised by their *susceptibility* to the magnetic field. For materials that are isotropic, a group that includes gases and liquids as well as glasses, cubic crystals, and polycrystalline solids, the *magnetic susceptibility*, χ, is defined by:

$$\mathbf{M} = \chi \mathbf{H}$$

where **M** is the *magnetisation* of the sample and **H** is the *magnetic field strength*. In non-isotropic solids, **M** and **H** are not necessarily parallel, and are usually defined as vectors.

These weak magnetic materials can also be characterised by the extent to which an external magnetic field is able to penetrate into the sample, the *magnetic permeability*, μ, defined by:

$$\mu = \mu_0(1 + \chi)$$

where μ_0 is a fundamental constant, the *permeability of free space* or vacuum permeability. Diamagnetic materials have μ less than μ_0, while paramagnetic materials have μ greater than μ_0. The magnetic field strength **H** decreases inside a diamagnetic solid and increases inside a paramagnetic solid (Figure 10.2).

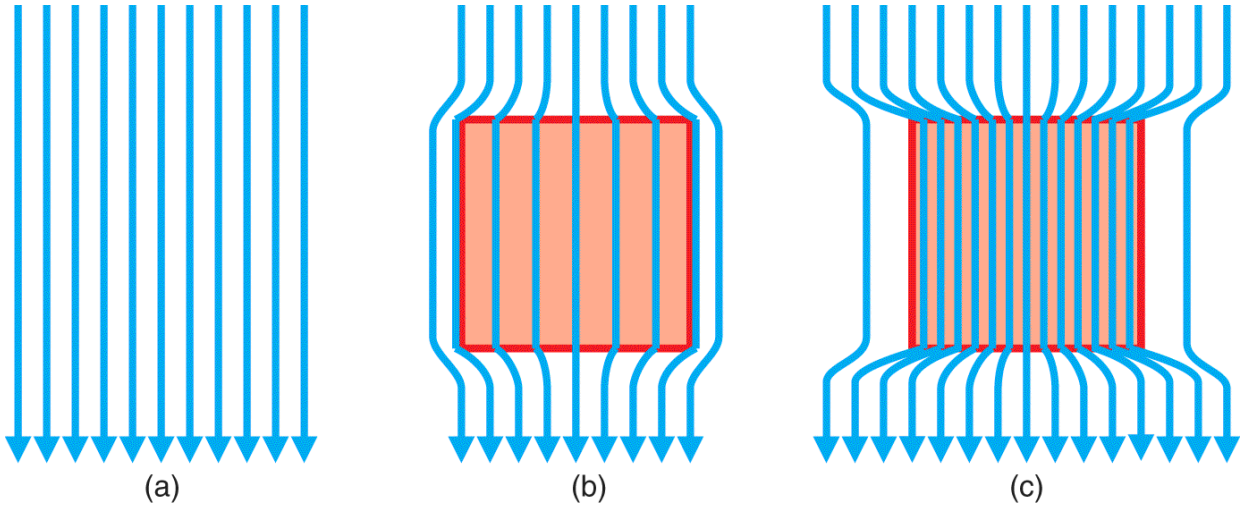


Figure 10.2 Weak magnetic materials in an external field: (a) no solid present; (b) a diamagnetic material; (c) a paramagnetic material.

10.1.2 Magnetic Dipoles and Magnetic Flux

Magnetic units are defined in terms of electric current. A small closed loop of current is called a *magnetic dipole*. It has a magnetic field similar to that of a small bar magnet and the two can be regarded as equivalent. The magnetic dipole moment of the current loop is defined by:

$$\begin{aligned}\mu &= [\text{current in loop}] \times [\text{area of loop}] \\ &= I\,\pi\,r^2\end{aligned}$$

where I is the current and r is the radius of the loop. The Earth acts as magnetic dipole, with a dipole moment of approximately 8×10^{22} A m^2. A magnetic dipole is frequently represented by an arrow, with the arrowhead at the 'north seeking' or north end of the dipole. A number of animals are able to detect the magnetic field of the Earth and use it for navigation (see Further Reading).

Two other important magnetic quantities, the *magnetic flux density* (also called *magnetic induction*) a vector quantity, **B**, and the magnetic field strength, **H**, are defined by reference to a solenoid, which is a cylindrical coil carrying an electric current. In a vacuum, the magnetic flux density and magnetic field strength within the solenoid, due to the current in the windings, are given by:

$$\mathbf{B}_0 = \mu_0\,\mathbf{H}$$

where μ_0 is the vacuum permeability. The magnetic field within the solenoid is given by:

$$\mathbf{H} = I\,n$$

where I is the current in the solenoid and n is the number of turns of the coil per metre. In the case when the solenoid is filled with a material, the magnetic flux density and the magnetic field are given by:

$$\mathbf{B} = \mu\mathbf{H}$$

where μ is the *permeability* of the material. The magnetic flux density is now due to two parts, the vacuum value, $\mathbf{B}_0$, and a part due to the magnetic behaviour of the sample. It is possible to write:

$$\mu = \mu_r \mu_0$$

where μ_r is the *relative permeability* of the material.

The contribution of the material to the magnetic flux density can also be defined by the equation:

$$\begin{aligned} \mathbf{B} &= \mathbf{B}_0 + \mu_0 \mathbf{M} \\ &= \mu_0 \mathbf{H} + \mu_0 \mathbf{M} \\ &= \mu_0 (\mathbf{H} + \mathbf{M}) \end{aligned} \qquad (10.1)$$

where $\mathbf{M}$ is the magnetisation of the sample. If $\mathbf{M}$ opposes $\mathbf{B}_0$, $\mathbf{B}$ is less than $\mathbf{B}_0$ and the sample is diamagnetic. If $\mathbf{M}$ is in the same direction as $\mathbf{B}_0$, $\mathbf{B}$ is increased and the substance is a paramagnetic or ferromagnetic. Equation (10.1) is also written:

$$\mathbf{B} = \mu_0 \mathbf{H} + \mathbf{J}$$

where $\mathbf{J}$ is the *magnetic polarisation*. For many paramagnetic materials $\mathbf{M}$ is small and it is possible to write:

$$\mathbf{B} \approx \mu_0 \mathbf{H}$$

For many ferromagnetic materials, $\mathbf{B}$ is much greater than $\mathbf{H}$, and it is possible to write:

$$\mathbf{B} \approx \mu_0 \mathbf{M}$$

For *isotropic materials*, the magnetisation is related to the magnetic field, $\mathbf{H}$, by the expression:

$$\mathbf{M} = \chi \mathbf{H} \qquad (10.2)$$

where the constant of proportionality, χ, is the magnetic susceptibility.

Combining Eqs. (10.1) and (10.2) gives:

$$\mathbf{B} = \mu_0(1 + \chi)\,\mathbf{H}$$

hence

$$\mu = \mu_0(1 + \chi)$$

and

$$\mu_r = (1 + \chi)$$

10.1.3 Atomic Magnetism

Magnetic properties reside in the subatomic particles that make up atoms. Of these, electrons make the biggest contribution, and only these will be considered here. Each electron has a magnetic dipole moment, which can be thought of as a minute bar magnet linked to the electron. There are two contributions to the magnetic dipole of an electron bound to an atomic nucleus, which, in semi-classical models, are attributed to orbital motion and spin, related to the two quantum numbers l and s. A spinning electron has a magnetic dipole moment, the *Bohr magneton*, μ_B, given by:

$$\mu_B = eh/4\pi m$$

where e is the charge on the electron, h is Planck's constant, and m the mass of the electron. Values of atomic magnetic moments are often expressed in terms of the Bohr magneton. The magnitude of the magnetic dipole moment of a single electron due to the orbital quantum number, l, the *orbital component*, is given by:

$$\mathbf{m}_{\text{orbital}} = \mu_B\,[l(l+1)]^{1/2}$$

This magnetic dipole can only take certain quantised directions in an applied magnetic field, so that the component of the magnetic dipole moment along the applied field direction is:

$$\mathbf{m} = -\mu_B m_l$$

where m_l is the magnetic quantum number that can take values 0, ±1, ±2, … ±l. Note the negative sign in this latter equation.

The magnitude of the magnetic dipole moment of a single electron due to the spin quantum number, s, the *spin component*, is given by:

$$\mathbf{m}_{spin} = g\mu_B[s(s+1)]^{1/2}$$

where g is *the free electron g-value*, equal to 2.002 319, and s is ½. As with the magnetic dipole due to the orbital angular momentum, the magnetic dipole due to the spin is only allowed to take restricted values in an applied magnetic field. The projection of the spin magnetic moment along the applied magnetic field is given by:

$$\mathbf{m} = -g\mu_B m_s$$

where m_s is analogous to m_l and can take the values ±½. Note the negative sign in this latter equation. These m_s values have been represented by ↑ and ↓ in earlier chapters, and in the present context, it is possible to think of an electron with spin 'up', ↑, as a magnetic dipole (or bar magnet) in one orientation and an electron with spin 'down', ↓, as a magnetic dipole (or bar magnet) in the opposite orientation.

The orbital and spin components are linked, or *coupled*, on isolated atoms or ions to give an overall magnetic dipole moment for the atom. The commonest procedure for calculating the resultant magnetic dipole moment is called the *Russell–Saunders coupling*. In summary, individual electron spin quantum numbers, s, are added to give a many electron total spin quantum number, S, the individual orbital quantum numbers, l, are added to give a many electron total orbital angular momentum quantum number, L. The many electron quantum numbers L and S are further combined to give a total angular momentum quantum number, J. (See Appendix B, for the method of calculating S, L, and J.)

The total magnetic dipole moment of the atom is given by:

$$\mathbf{m}_{\text{atom}} = g_J \, \mu_B \, [J(J+1)]^{1/2} \tag{10.3}$$

where g_J is the *Landé g-factor*, given by:

$$g_J = 1 + \{[J(J+1) - L(L+1) + S(S+1)]/2J(J+1)\} \tag{10.4}$$

The total magnetic dipole can only take a restricted number of directions with respect to an applied magnetic field, leading to values along the applied field direction of:

$$\mathbf{m} = -g_J \, \mu_B \, M_J$$

where M_J, which can take values of 0, ±1, ±2, ... ±J, is the quantum number representing the projection of the total angular momentum vector, J, onto the field axis. Note the negative sign in this latter equation.

A completely filled orbital contains electrons with opposed spins and the value of S is zero. Similarly, a completely filled s, p, d, or f orbital set has L equal to zero. This means that atoms with filled closed shells have no magnetic moment. The only atoms that display a magnetic moment are those with incompletely filled shells notably the transition metals, with incompletely filled d shells, and the lanthanoids and actinoids, with incompletely filled f shells.

10.1.4 Overview of Magnetic Materials

Magnetic materials can be classified in terms of the arrangements of magnetic dipoles in the solid. These dipoles can be thought of, a little imprecisely, as microscopic bar magnets attached to the various atoms present. Materials with no elementary magnetic dipoles at all are diamagnetic. The imposition of a magnetic field generates weak magnetic dipoles that are only present for as long as the field persists (Figure 10.3). The induced dipole is opposed to the magnetic flux density, **B**. The magnetic susceptibility of a diamagnetic substance is negative and very slightly less than 1. There is no appreciable variation of diamagnetism with temperature.

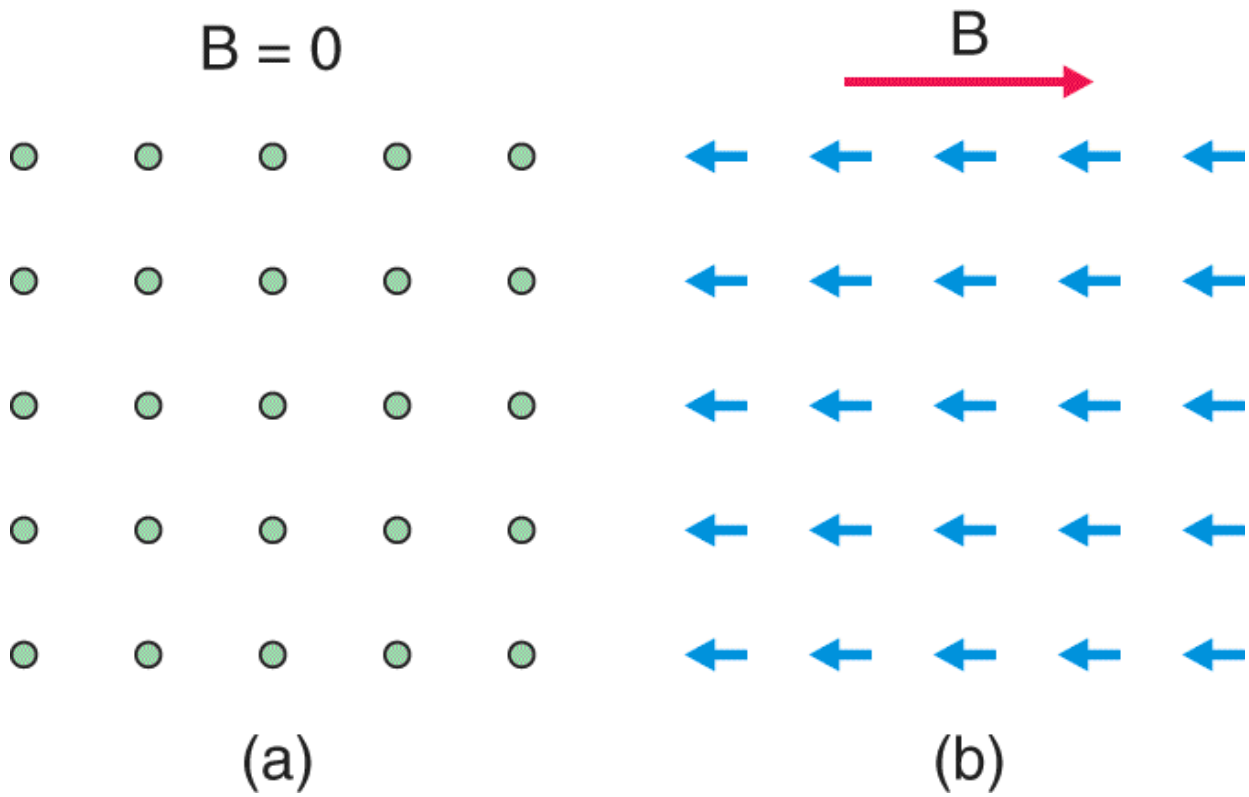


Figure 10.3 Behaviour of a diamagnetic solid (a) in no field; (b) in an applied magnetic field.

Paramagnetic solids are those in which some of the atoms, ions, or molecules making up the solid possess a permanent magnetic dipole moment. These dipoles are isolated from one another. The solid, in effect, contains small, non-interacting atomic magnets. In the absence of a magnetic field, these are arranged at random and the solid shows no net magnetic moment. In a magnetic field, the elementary dipoles will attempt to orient themselves parallel to the magnetic flux density in the solid, and this will enhance the internal field within the solid and give rise to the observed paramagnetic effect (Figure 10.4a,b). The alignment of dipoles will not usually be complete, because of thermal effects and interaction with the surrounding atoms in the structure and the dipoles continually change orientation because of this jostling. Thus, the disposition of the dipoles at any instant, t_1, will be different from that at any other instant, t_2. The magnetic effect is much greater than diamagnetism, and the magnetic susceptibility of a paramagnetic solid is positive and slightly greater than 1.

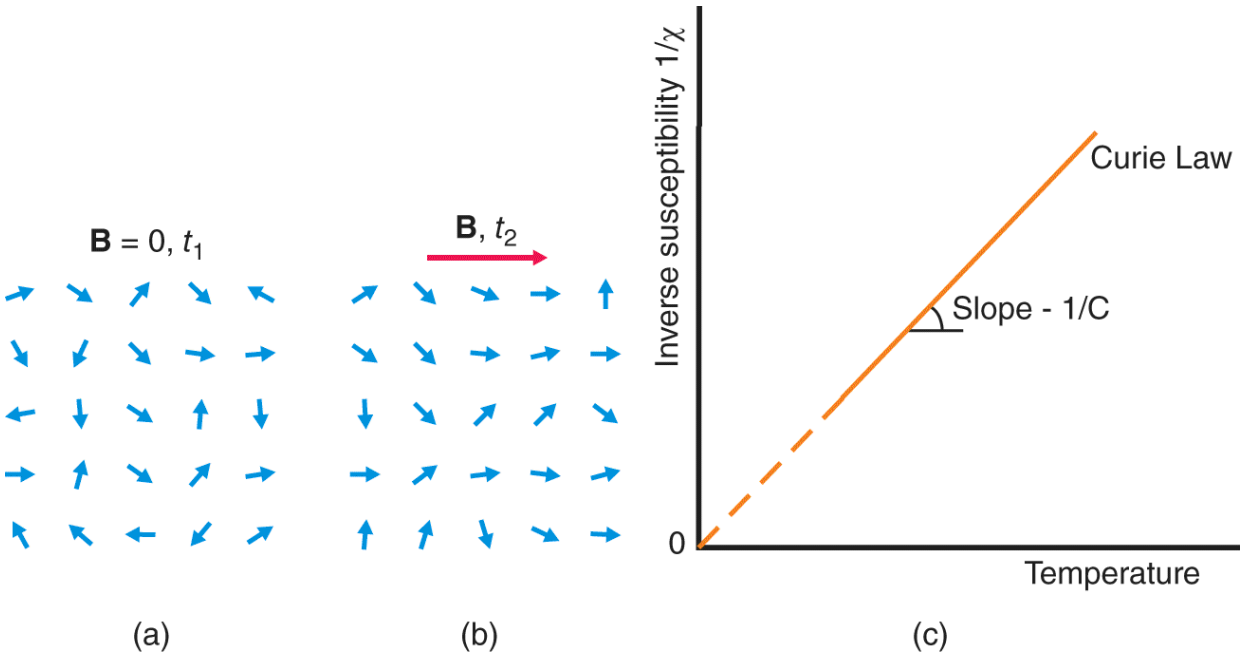


Figure 10.4 (a, b) Behaviour of a paramagnetic solid in an applied magnetic field; (c) temperature dependence of reciprocal magnetic susceptibility.

Because of thermal agitation it would be expected that paramagnetic susceptibility would vary with temperature. This variation (Figure 10.4c) is given by the *Curie law*:

$$\chi = C/T$$

where χ is the magnetic susceptibility, T is the temperature (K), and C is the *Curie constant*. Curie law dependence in a solid is indicative of the presence of isolated paramagnetic ions or atoms in the material.

Ferromagnetic materials are those in which the magnetic dipoles align parallel to each other over considerable distances in the solid (Figure 10.5a). An intense external magnetic field is produced by this alignment. Ferromagnetism is associated with the transition elements that possess unpaired d electrons, and the lanthanoids and actinoids with unpaired f electrons.

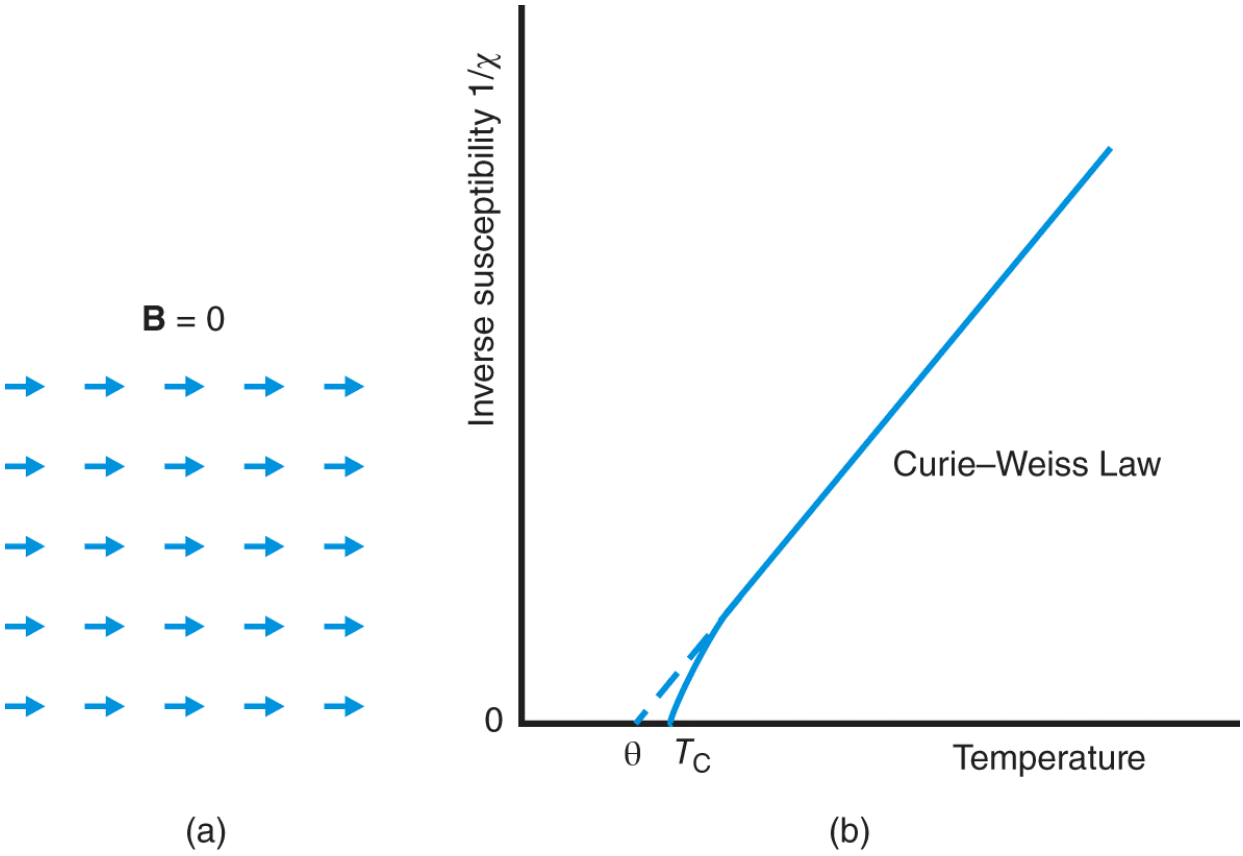


Figure 10.5 (a) Magnetic dipole ordering in a ferromagnetic solid; (b) temperature dependence of reciprocal magnetic susceptibility.

All ferromagnetic materials become paramagnetic above the *Curie temperature*, T_C. The transition to a paramagnetic state comes about when thermal energy is greater than the magnetic interactions, and causes the dipoles to disorder. The transition is reversible, and on cooling, ferromagnetism returns when the magnetic dipoles align parallel to one another as the temperature drops through the Curie temperature. *Well above* the Curie temperature, ferromagnetic materials obey the *Curie–Weiss law*:

$$\chi = C/T - \theta$$

The Curie–Weiss constant, θ, is positive, and has the dimensions of temperature, and a value usually close to, but not quite identical to, the Curie temperature, T_C (Figure 10.5b).

It is energetically favourable in some materials for the elementary magnetic dipoles to align in an antiparallel fashion. These are called

antiferromagnetic compounds. Above a temperature called the *Néel temperature*, T_N, this arrangement disorders and the materials revert to paramagnetic behaviour. Cooling the sample through the Néel temperature causes the antiferromagnetic ordering to reappear. *Well above* the Néel temperature, antiferromagnetic materials obey the Curie–Weiss law as earlier. In this case, the Curie–Weiss constant, θ, is negative (Figure 10.6).

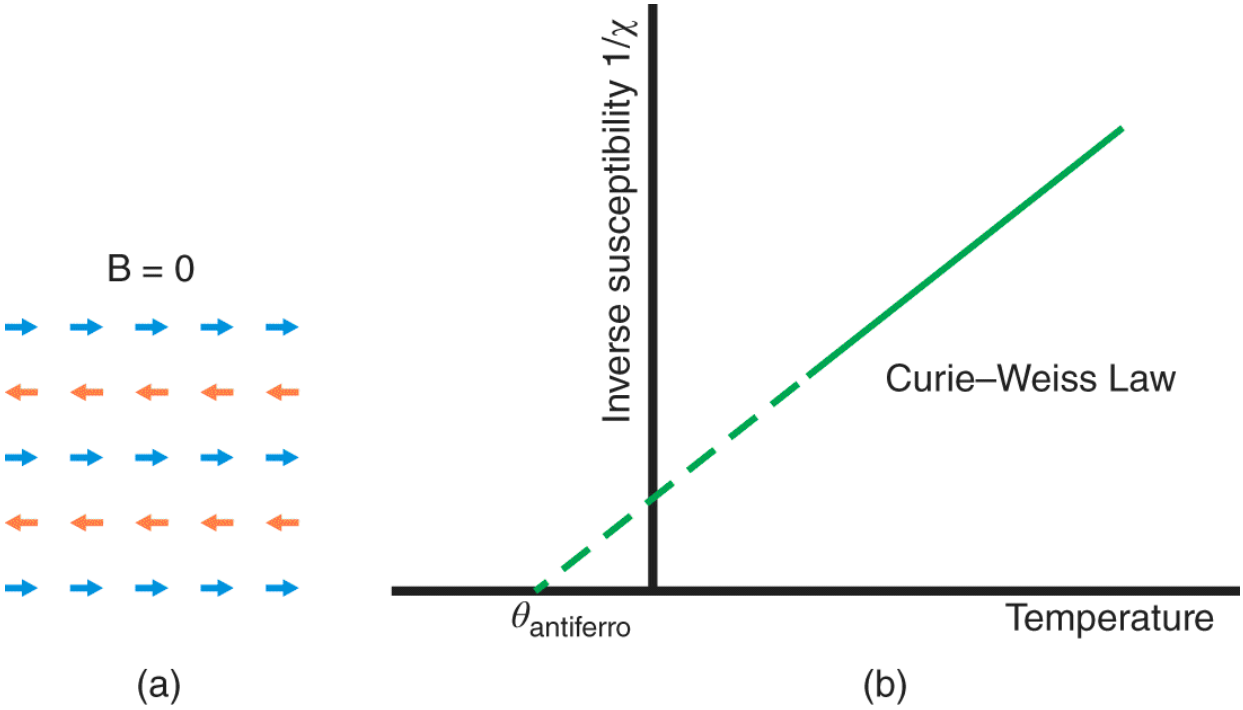


Figure 10.6 (a) Magnetic dipole ordering in an antiferromagnetic solid; (b) temperature dependence of reciprocal magnetic susceptibility.

An important group of solids have two different magnetic dipoles present, one of greater magnitude that the other. When these align in an antiparallel arrangement, they behave rather like ferromagnetic materials. They are called *ferrimagnetic* materials. A ferrimagnetic solid shows complex temperature dependence because the distribution of the magnetic ions over the available sites is sensitive to both temperature and the spin interactions. The behaviour can be approximated by the equation:

$$1/\chi = T/C + 1/\chi_0 + \xi/[T - \theta]$$

where the parameters χ_0, θ, and ξ depend upon the population of the available cation sites and the spin interactions. These are often taken as constants and a graph of $1/\chi$ vs. temperature is a straight line except near the Curie temperature T_C (Figure 10.7).

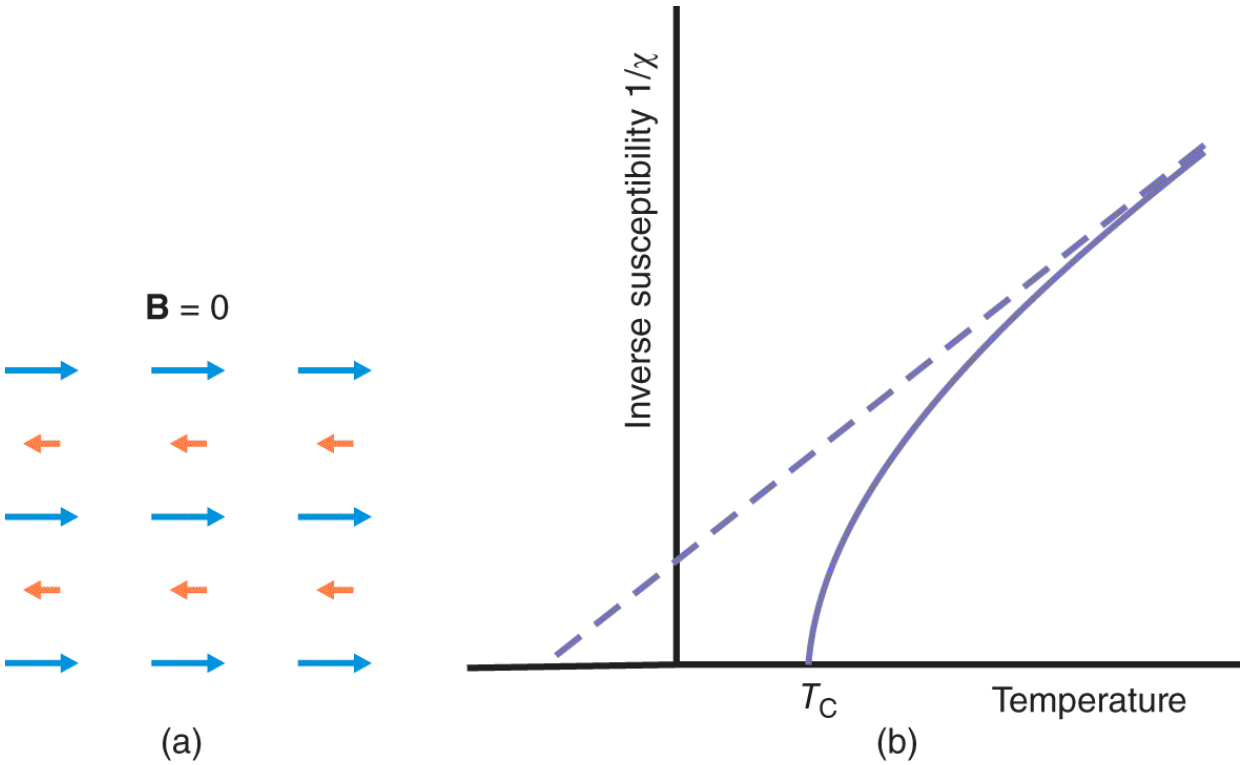


Figure 10.7 (a) Magnetic dipole ordering in a ferrimagnetic solid; (b) temperature dependence of reciprocal magnetic susceptibility.

Ferromagnetic ordering and antiferromagnetic ordering represent extremes of dipole orientation. In a number of solids, neighbouring magnetic dipoles, although ordered, are not aligned parallel to one another, but at an angle, referred to a *canted* arrangement. There are a number of possible arrangements of magnetic dipoles that fall into this category, the simplest of which can be thought of as an intermediate configuration between ferromagnetic and antiferromagnetic states (Figure 10.8).

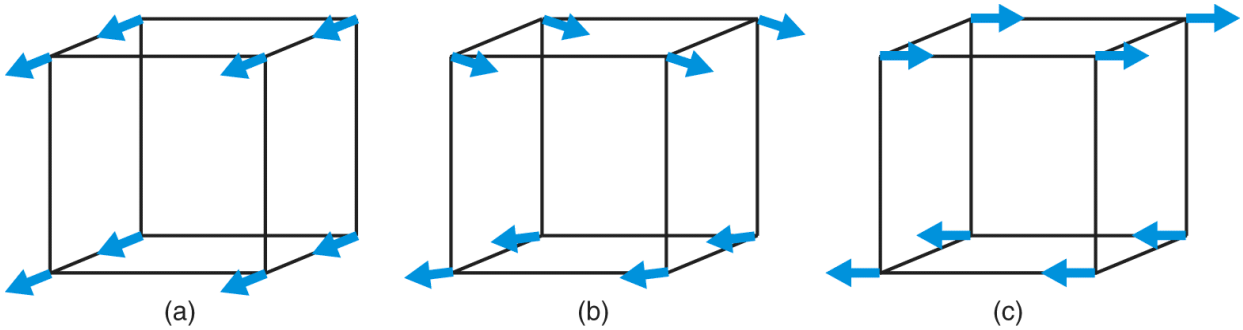


Figure 10.8 Ordered arrays of magnetic dipoles arranged on a cubic lattice: (a) ferromagnetic configuration; (b) canted antiferromagnetic configuration; (c) antiferromagnetic configuration.

Interacting magnetic dipoles can take on a variety of other configurations in a solid. One of these is the *spin glass* state. The structural characteristics of this group of solids are that at low temperatures, ordered spin states exist (such as ferromagnetic alignment, or helical spin spirals), but these are confined to small regions, each of which is magnetically incoherent with respect to neighbouring regions, in an analogous way to the structure of a silica glass, which has ordered regions of SiO_4 tetrahedra that are not structurally aligned with neighbours so as to form a long-range crystal structure. Spin glass states are formed at low temperatures, when the random re-orientation of the dipoles is restricted, and changes only slowly with time. The directions of the spins are said to become frozen below a freezing temperature, T_f, to produce the spin glass state. Spin glass behaviour occurs in a wide variety of solids at low temperatures, including elemental Nd, metal alloys in the Cu–Mn system, and in compounds such as Prussian blue, $KFe_2(CN)_6$, the ferrites $RFeTi_2O_7$, in which R is a lanthanoid cation and the paramagnetic $Bi_4V_{1.7}Ni_{0.3}O_{10.55}$. In the main, the spin glass state is characterised by the way in which the magnetic properties change with time (age) at low temperatures, but these materials still remain poorly understood.

10.2 Paramagnetic Materials

10.2.1 The Magnetic Moment of Paramagnetic Atoms and Ions

Paramagnetic atoms and ions are those with unpaired electrons, most importantly the transition metals and the lanthanoids. These endow the atoms with a magnetic dipole that is oriented at random with respect to neighbouring dipoles. The magnetic dipole moment of a paramagnetic solid containing paramagnetic atoms or ions is given by Eqs. (10.3) and (10.4).

Lanthanoid ions have a partly filled 4f shell, and these orbitals are well shielded from any interaction with the surrounding atoms by filled 5s, 5p, and 6s orbitals, so that (with the notable exceptions, Eu^{3+} and Sm^{3+}), they behave like isolated ions. The calculated magnetic dipole moments of these species, assuming that the magnetism arises solely from the 4f electrons, is in good agreement with measurements (Table 10.1).

Table 10.1 Calculated and observed magnetic dipole moments for the lanthanoids.

Ion	Configuration	S	L	J	g_J	m_{calc}[a]	m_{meas}[a]
La^{3+}, Ce^{4+}	f^0	0	0	0	0	0	Diamagnetic
Ce^{3+}, Pr^{4+}	f^1	1/2	3	5/2	6/7	2.54	~2.5
Pr^{3+}	f^2	1	5	4	4/5	3.58	3.5
Nd^{3+}	f^3	3/2	6	9/2	8/11	3.62	3.5
Pm^{3+}	f^4	2	6	4	3/5	2.68	—
Sm^{3+}	f^5	5/2	5	5/2	2/3	0.84	1.5
Sm^{2+}, Eu^{3+}	f^6	3	3	0	0	0	3.4 (Eu^{3+})
Eu^{2+}, Gd^{3+}, Tb^{4+}	d^7	7/2	0	7/2	2	7.94	~8.0
Tb^{3+}	f^8	3	3	6	3/2	9.72	9.5
Dy^{3+}	f^9	5/2	5	15/2	4/3	10.63	10.6
Ho^{3+}	f^{10}	2	6	8	5/4	10.61	10.4
Er^{3+}	f^{11}	3/2	6	15/2	6/5	9.59	9.5
Tm^{3+}	f^{12}	1	5	6	7/6	7.57	7.3
Yb^{3+}	f^{13}	1/2	3	7/2	8/7	4.54	4.5
Yb^{2+}, Lu^{3+}	f^{14}	0	0	0	0	0	Diamagnetic

[a] The units are Bohr magnetons. The calculated values are from Eq. (10.3).

For the transition metals, especially those of the 3d series, interaction with the surroundings is considerable. This has two important consequences. One is the curious fact that 3d-transition metal ions in paramagnetic solids often have magnetic dipole moments corresponding only to the electron spin contribution, given by the quantum number S. The orbital moment, L, is said to be *quenched*. In such materials, Eq. (10.3) reduces to:

$$g_J = 1 + \{[S(S+1) + S(S+1)]/$$
$$\times 2S(S+1)\} = 2$$

To a good approximation, Eqs. (10.3) and (10.4) can then be replaced by a *spin only formula*:

$$\mathbf{m} = 2\mu_B [S(S+1)]^{1/2}$$

where S is the total spin quantum number, equal to $s_1 + s_2 + s_3 + \cdots$ for the unpaired electrons, and μ_B is the Bohr magneton. In the case of n unpaired electrons on each ion, each of which has s equal to ½,

$$S = s_1 + s_2 + s_3 \cdots s_n = 1/2 + 1/2 + 1/2 \cdots$$
$$= n/2$$

The spin only formula can then be written:

$$\mathbf{m} = [n(n+2)]^{1/2} \mu_B \qquad (10.5)$$

The magnetic dipole moments calculated in this way are generally in good agreement with observation (Table 10.2). In cases where the experimental value is considerably different to the spin only value, orbital effects are presumed to be significant.

Table 10.2 The magnetic properties of the 3d-transition metal ions.

Ion	Configuration	S	m_{calc}[a]	m_{meas}[a]
Sc^{3+}, Ti^{4+}, V^{5+}	d^0	0	0	Diamagnetic
Ti^{3+}, V^{4+}	d^1	1/2	1.73	1.7–1.8
Ti^{2+}, V^{3+}	d^2	1	2.83	2.8–2.9
Cr^{3+}, Mn^{4+}, V^{2+}	d^3	3/2	3.87	3.7–4.0
Cr^{2+}, Mn^{3+}	d^4	2	4.9	4.8–5.0
Mn^{2+}, Fe^{3+}	d^5	5/2	5.92	5.7–6.1
Co^{3+}, Fe^{2+}	d^6	2	4.9	5.1–5.7
Co^{2+}	d^7	3/2	3.87	4.3–5.2
Ni^{2+}	d^8	1	2.83	2.8–3.5
Cu^{2+}	d^9	1/2	1.73	1.7–2.2
Cu^{+}, Zn^{2+}	d^{10}	0	0	Diamagnetic

[a] The units are Bohr magnetons. The calculated values are from Eq. (10.5).

10.2.2 High and Low Spin: Crystal Field Effects

An important feature of 3d transition metal ions is that those ions with configurations between $3d^4$ and $3d^7$ have two apparent spin states, *high spin* (HS) and *low spin* (LS). For example, Fe^{3+}, with a $3d^5$ electron configuration, sometimes appears to have five unpaired spins and sometimes only one. To understand this, it is necessary to take into account the effect that the surrounding atoms in the solid have on the energy of the d orbitals. The simplest approach – *crystal field theory* – replaces the atoms in the surrounding structure by point electric charges. The interactions are electrostatic in nature and split the five d orbitals into groups with differing energies, depending upon the arrangement of the point charges.

The energy level splitting comes about in the following way. The d orbitals on the transition metal ion point along or between the *x*-, *y*-, and *z*-axes (Figure 1.5). In the d_{xy}, d_{xz}, and d_{yz} groups, the lobes of

electron density lie between the axes, while in the other pair, $d_{x^2-y^2}$ and d_{z^2}, the electron density is concentrated along the axes. In an isolated atom or ion, the d orbitals all have the same energy. Suppose that the cation is now placed into an oxygen octahedron, made up from by six negative O^{2-} ions (Figure 10.9a). Firstly, the energy of the whole d group will be raised considerably from that of an isolated ion. Secondly the energy will split into two parts. The d orbitals pointing directly towards the O^{2-} ions ($d_{x^2-y^2}$ and d_{z^2}) will be raised in energy most. This group, labelled e_g, contains two orbitals of equal energy. The d orbitals pointing between the O^{2-} ions (d_{xy}, d_{xz}, and d_{yz}) are favourably located and their energy increase will be less than he first pair. This lower energy group, labelled t_{2g}, contains three orbitals of equal energy.

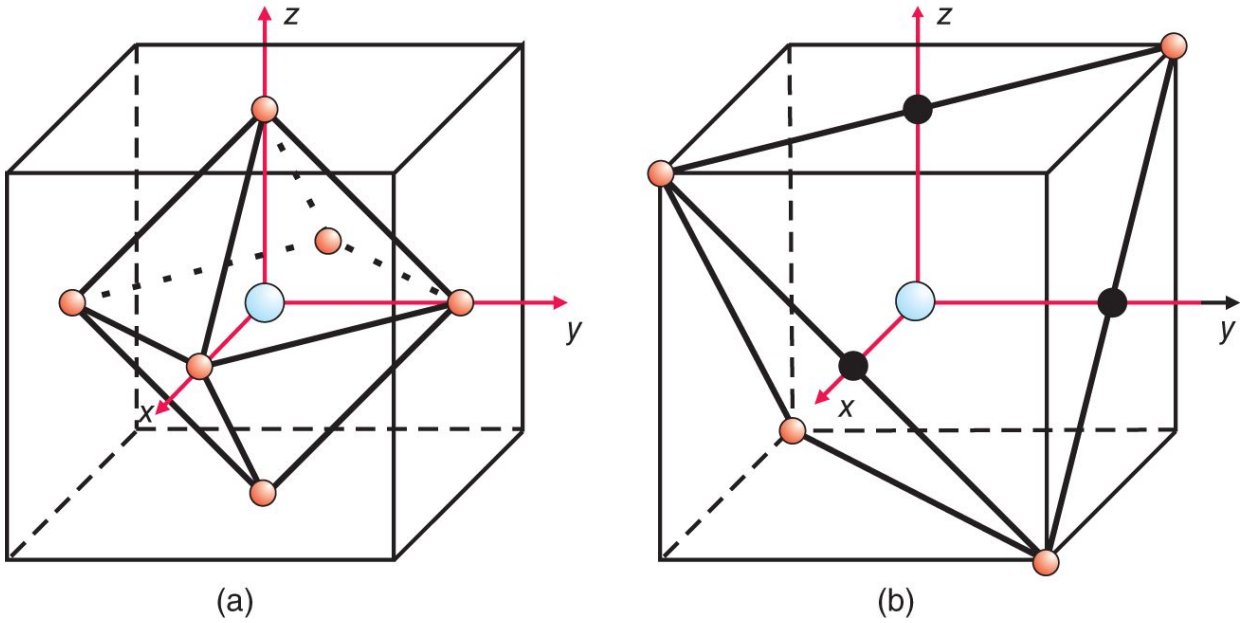


Figure 10.9 Cations in (a) an anion octahedron; (b) an anion tetrahedron.

The energy gap between the t_{2g} and the e_g orbitals, called the *crystal field splitting*, is written as Δ or 10Dq. The magnitude of the crystal field splitting will depend on the anion–cation spacing and charges. In a *strong crystal field*, produced when the surrounding charges are large and close to the cation, the crystal field splitting is large. In a *weak crystal field*, produced when the surrounding charges are small

and further away from the cation, the splitting is smaller (Figure 10.10).

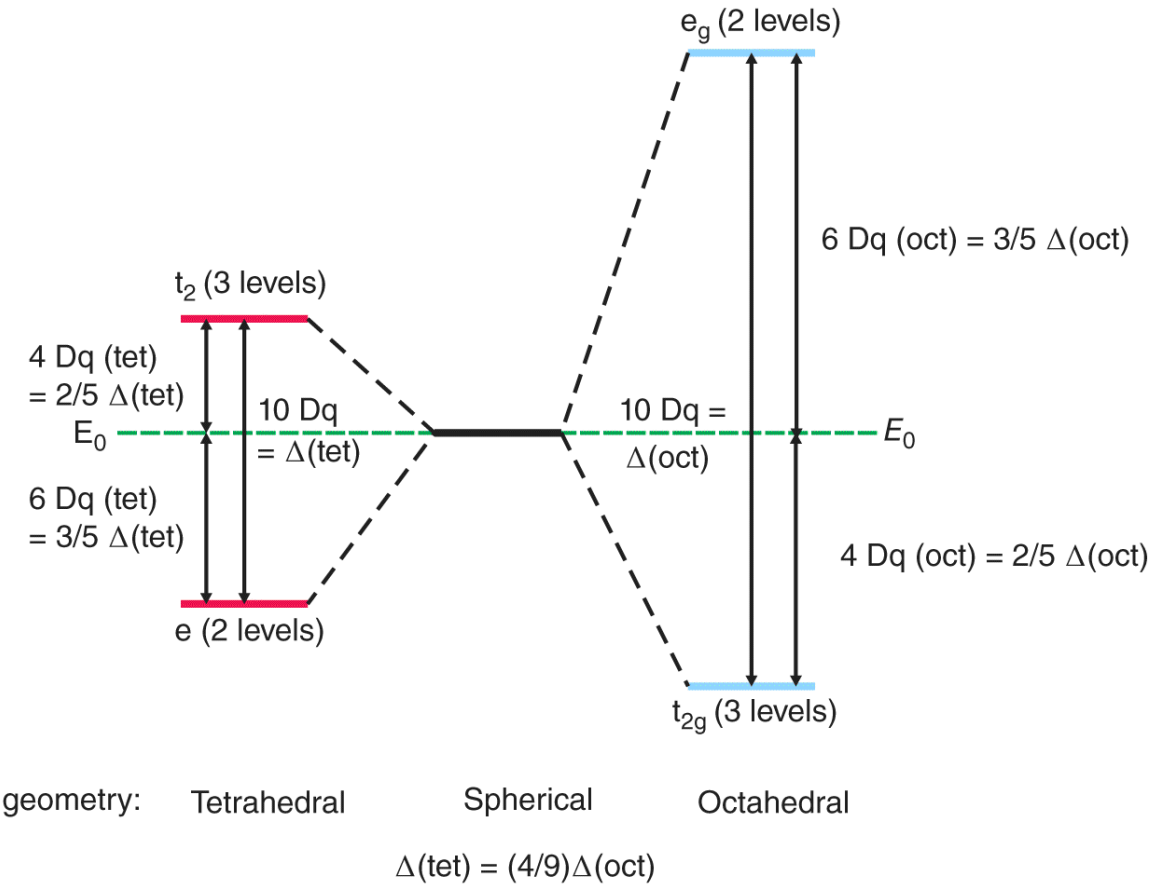


Figure 10.10 The crystal field splitting, Δ (tet) or 10 Dq (tet) and Δ (oct) or 10 Dq (oct), of the energy of the d orbitals in tetrahedral and octahedral crystal fields with respect to a spherical distribution of the surrounding charges.

The form of the crystal field splitting clearly depends on the geometry of the surrounding anions. When a transition metal ion is surrounded by a tetrahedron of oxygen ions (Figure 10.9b), the crystal field splitting is reversed compared with that in an octahedral geometry. In this case, the d_{xy}, d_{xz}, and d_{yz} orbitals, now called the t_2 group, are raised in energy relative to $d_{x^2-y^2}$ and d_{z^2}, the e group. The magnitude of the splitting for cations in a tetrahedron will be less than that for cations in an octahedron because there are only four anions instead of six and because they are further away from the central ion. Calculations show that the tetrahedral crystal field splitting is 4/9 of the octahedral splitting.

In order to build up the electron configurations of the 3d transition metal ions, electrons are placed into the now split d orbitals. Take, as an example, the sequence for an ion in an octahedral crystal field (Figure 10.11). In the case of the ions $3d^1$–$3d^3$, the electrons keep apart and retain parallel spins, in accordance with Hund's rules. However, two options are possible for a $3d^4$ ion. The fourth electron can keep apart from the others, and enter an empty e_g orbital, or it can spin pair with one of the electrons in the t_{2g} orbitals. Which alternative is chosen will depend upon the strength of the crystal-field splitting. When this is small, the *weak field* case, separation is preferred. When it is large, the *strong field* case, spin pairing is the energetically favoured option. These are the HS and LS configurations, respectively. LS and HS alternatives exist for the ions from d^4 to d^7. Only one configuration is again possible once the d^8 and d^9 ions are reached. The magnetic properties of d^4–d^7 ions will thus depend upon whether they are in an HS or LS configuration.

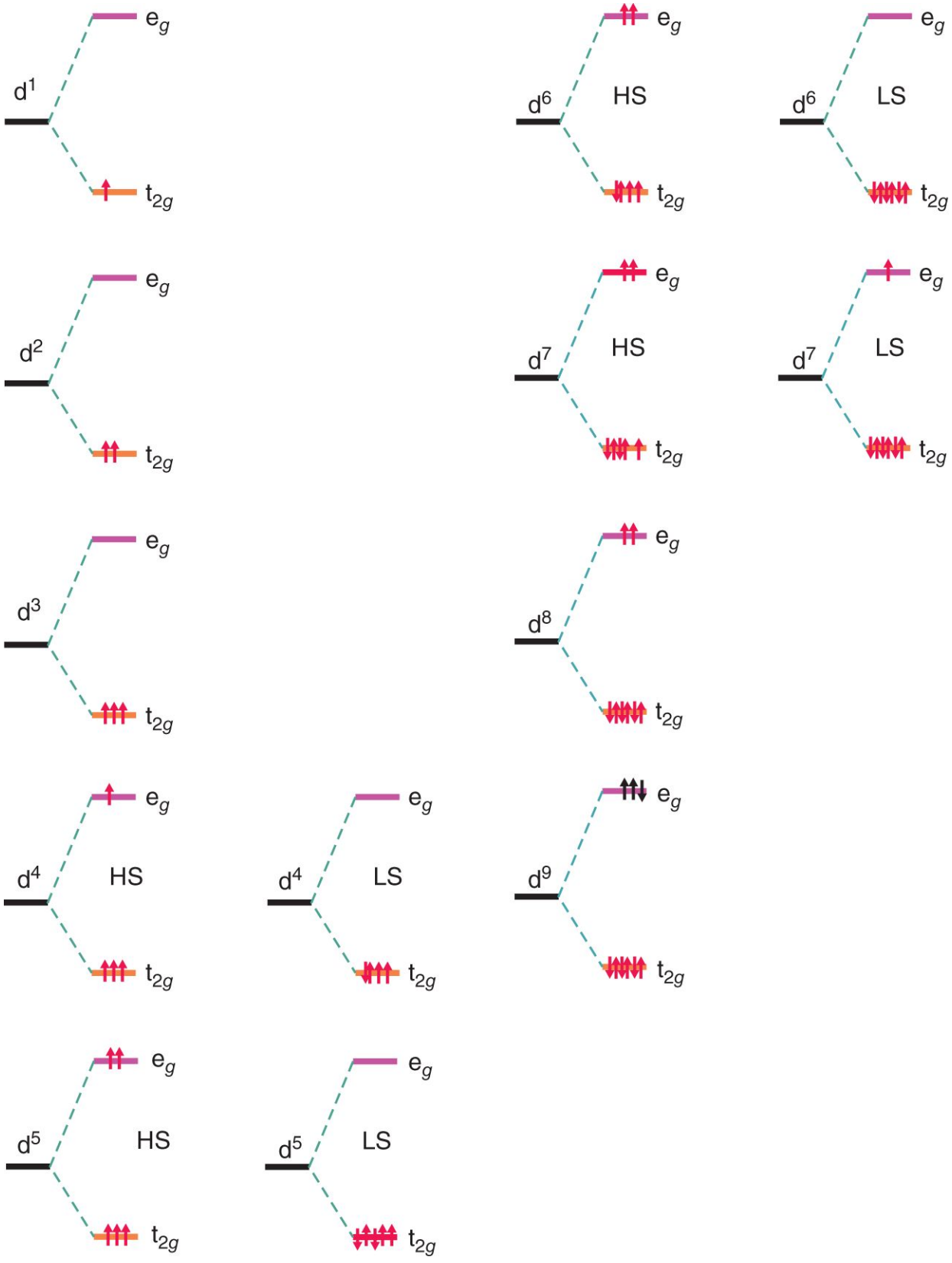

e_g
t_{2g}
d^1
d^2
d^3
d^4 HS
d^4 LS
d^5 HS
d^5 LS
d^6 HS
d^6 LS
d^7 HS
d^7 LS
d^8
d^9

Figure 10.11 The electron configurations possible for d^n cations in an octahedral crystal field.

This procedure can be repeated for ions in a tetrahedral geometry. However, only HS configurations are generally found because the crystal field splitting is always much smaller than that in octahedral geometry.

10.2.3 Temperature Dependence of Paramagnetic Susceptibility

As described earlier, the projection of the magnetic moment vector of a paramagnetic atom upon the magnetic field direction is restricted to values of M_J given by:

$$M_J = J, J-1, J-2, \ldots -J$$

The $2J + 1$ values of M_J give rise to a set of energy levels of separation:

$$\Delta E = g_J \, \mu_B \, B$$

where g_J is the Landé g-factor, Eq. (10.4), and B is the magnitude of the magnetic flux density (see Figures 1.6 and 10.12). The average value of the magnetic dipole moment, $\langle \mathbf{m} \rangle$, will depend upon the population of these levels when at thermal equilibrium. The bulk magnetisation will then be given by:

$$\mathbf{M} = N\langle \mathbf{m} \rangle$$

where N is the number of magnetic dipoles per unit volume.

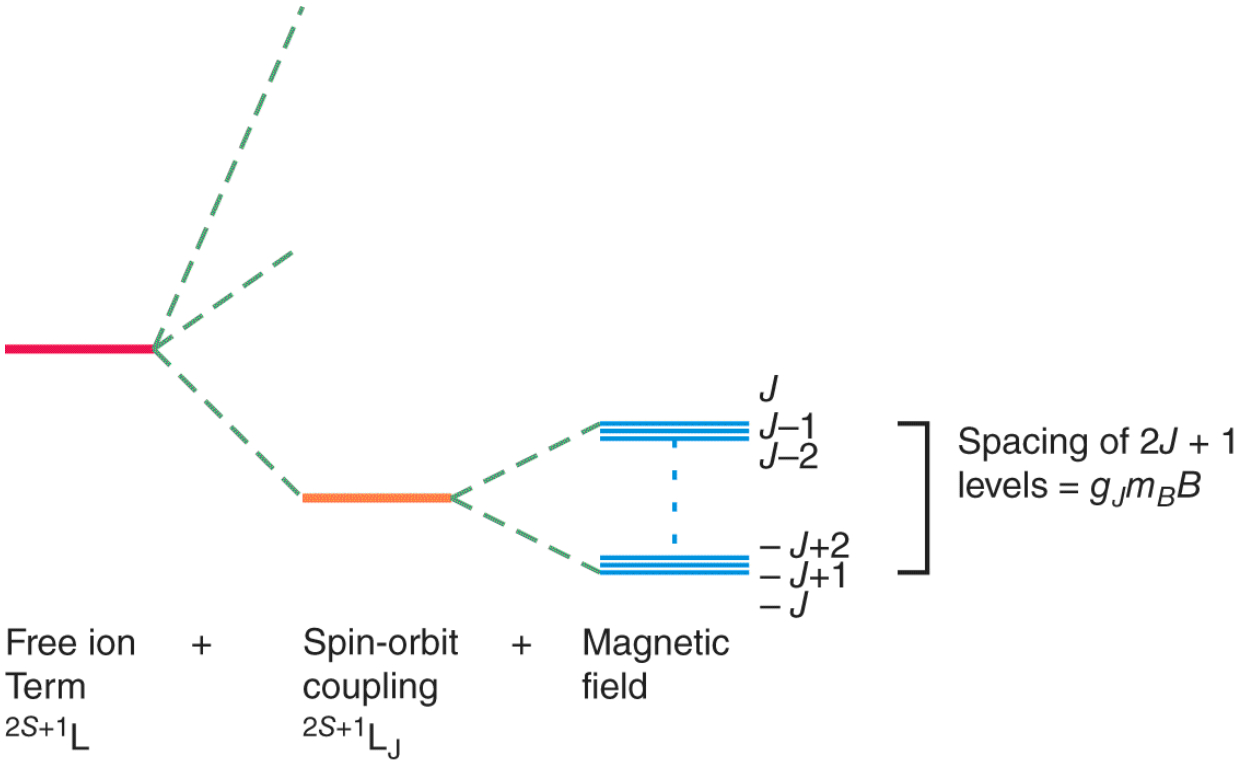


Figure 10.12 Splitting of atomic energy levels in a magnetic field.

It is found that the value of **M** is:

$$\mathbf{M} = N g_J \mu_B J B_J(x) \tag{10.6}$$

where $B_J(x)$ is the *Brillouin function*:

$$B_J(x) = \frac{2J+1}{2J} \coth\left\{\frac{2J+1}{2J}x\right\} - \frac{1}{2J}\coth\left\{\frac{x}{2J}\right\}$$

and

$$x = g_J \mu_B J B / k_B T$$

(The classical case, in which the magnetic dipole moments can rotate freely, is equivalent to a continuum of energy levels with J tending to infinity. In this case, the magnetisation is given by:

$$\mathbf{M} = N\mathbf{m}L(x) = N\mathbf{m}\,[\coth\, x - 1/x]$$

where

$$x = mB/k_{\mathrm{B}}T$$

and the function $L(x)$ is the *Langevin function*.)

The rather complex equation for the magnetisation of a paramagnetic substance can be simplified for the case of $x \ll 1$, which holds in the case of higher temperatures and lower magnetic fields. (In fact, even when the magnetic flux density is equal to 1 T, a temperature of 300 K is sufficient to make $x \ll 1$). In this case, the magnetisation is given by:

$$\mathbf{M} = [N\, {g_J}^2 {\mu_{\mathrm{B}}}^2\, B J(J+1)]/3k_{\mathrm{B}}T$$

Substituting $\mathbf{B} = \mu_0\mathbf{H}$ for paramagnetic materials and rearranging gives:

$$\mathbf{M}/\mathbf{H} = \chi = [N\mu_0 {g_J}^2 {\mu_{\mathrm{B}}}^2\, J(J+1)]/3k_{\mathrm{B}}T$$

where χ is the magnetic susceptibility. This has the same form as the Curie equation if the Curie constant is given by:

$$C = [N\mu_0 {g_J}^2 {\mu_{\mathrm{B}}}^2\, J(J+1)]/3k_{\mathrm{B}}$$

The value of χ is proportional to the number of magnetic dipoles present and is thus given in several forms: when N is the number of dipoles per unit volume, the volume susceptibility, χ_{v} m^{-3}, is appropriate; when N is the number of magnetic dipoles per unit mass, the mass susceptibility, χ_{mass}, is appropriate; with $\chi_{\mathrm{mass}} = \chi_{\mathrm{v}}/\rho$, where ρ is the density. The susceptibility of 1 mol of the substance, the molar susceptibility, χ_{m}, is given by $\chi_{\mathrm{v}}\, V_{\mathrm{m}}$, where V_{m} is the molar volume of the material. The units of molar susceptibility are $\mathrm{m}^3\ \mathrm{mol}^{-1}$. Molar susceptibility is also obtained by substituting $N = N_{\mathrm{A}}$ in the formula.

This analysis of paramagnetism is widely used to obtain the spin states and valence of cations, especially 3d transition metal cations, in solids. Assuming that the atomic magnetism, **m**, is given by Eqs. , the expression for the Curie constant can be written:

$$C_{\text{molar}} = [N_A \mu_0 g_J{}^2 \mu_B{}^2 J(J+1)]/3k_B$$
$$= [N_A \mu_0 \mu_B{}^2/3k_B]\,\mathbf{m}^2$$

where **m** is given in Tables 10.1 and 10.2. Inserting values for the constants:

$$C_{\text{molar}} = 1.5721 \times 10^{-6}\,\mathbf{m^2}$$

For example, a plot of the inverse paramagnetic susceptibility against temperature for the perovskite $EuTiO_{2.93}H_{0.7}$ yields a value of 9.211 $\times 10^{-5}$ K mol^{-1} for the Curie constant. The magnetic moment of the paramagnetic cation is then:

$$9.211 \times 10^{-5}\ \text{K mol}^{-1} = 1.5721 \times 10^{-6}\,\mathbf{m}^2$$

$$\mathbf{m}^2 = 9.211 \times 10^{-5}/1.5721 \times 10^{-6}$$

$$\mathbf{m} = 7.65$$

This corresponds to Eu^{2+} (Table 10.1)

10.2.4 Pauli Paramagnetism

Most metals show weak paramagnetic behaviour. It is rather small and independent of temperature, and is quite different to the Curie law behaviour of the paramagnetic solids described in Sections 10.2.1, 10.2.2, and 10.2.3, associated with unpaired electrons in the d and f orbitals. These orbitals do not interact greatly with the surroundings and the magnetic situation is quite well described in terms of localised electrons placed on a particular atom. However, the properties of metals are best described in terms of band theory, in which electrons are free to move throughout the bulk (Section 1.4).

The paramagnetic behaviour of metals can be explained on this basis and is known as *Pauli paramagnetism*.

Electrons in a normal metal occupy the (partly filled) uppermost conduction band. (All electrons not in the conduction band will be in filled shells, paired, and not magnetically active.) The electrons are spin paired, with two electrons being allocated to each energy state. Equal numbers of spin up and spin down electrons are present, which would then make the metal diamagnetic (Figure 10.13a). In a magnetic field those electrons with a spin parallel to the field have a slightly lower energy than those with a spin opposed to the field. Those electrons near the very top of the conduction band, i.e. those at the Fermi surface, can reorient themselves in the applied magnetic field. This produces an imbalance in the numbers of spin up and spin down electrons, and the metal becomes paramagnetic (Figure 10.13b).

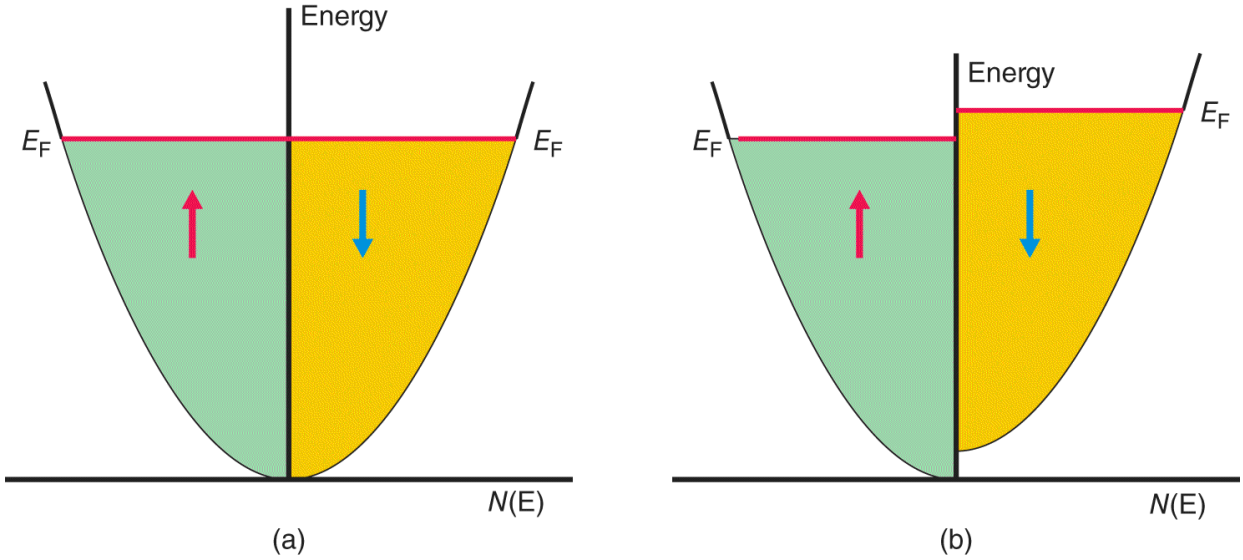


Figure 10.13 The density of states for electrons in a metal: (a) in the absence and (b) in the presence of a magnetic field.

Calculation shows that the number influenced is very small so that the paramagnetic magnetic susceptibility is proportional to the density of states at the Fermi level, $N(E_F)$:

$$\chi \propto \mu_0 \, \mu_B \, N(E_F)$$

where μ_0 is the permittivity and μ_B the Bohr magneton. Thus, a measurement of the Pauli paramagnetism of a normal metal provides

a method of assessing the density of states at the Fermi level. Temperature does not change the numbers of electrons at the Fermi surface appreciably, so the paramagnetism is virtually temperature independent for the same reason that the specific heat contribution of the electrons is very small.

10.3 Ferromagnetic Materials

10.3.1 Ferromagnetism

Ferromagnetic materials are of importance in many industries and the fabrication of better and more powerful permanent magnets is an important research area. Historically the first magnet known (and the origin of the word) was the iron oxide *magnetite* or *lodestone*, Fe_3O_4, which is, in fact, a ferrimagnetic material. Ferromagnetism is found in only four elements at room temperature, Fe, Co, Ni, and Gd. In addition, two lanthanoids, Tb and Dy, are ferromagnetic below room temperature (Table 10.3). In the period up to 1950, only two types of ferromagnetic materials were in widespread use, some steels and an alloy of Al, Ni, and Co called Alnico. The latter half of the twentieth century produced the Rare Earth (i.e. lanthanoid) magnets found in the systems Sm–Co (1960s), Nd–Fe–B (1980s), and Sm–Fe–Ni (1990s). (Other permanent magnets made during these years, notably the ferrites, are ferrimagnets, Section 10.5.)

Table 10.3 Ferromagnetic and antiferromagnetic compounds.

Compound	T_C (K)	T_N (K)
Ferromagnetic solid		
Fe	1043	
Co	1388	
Ni	627	
Gd	293	
Tb	220–230	
Dy	87–176	
CrO_2	386	
$SmCo_5$	973	
$Nd_2Fe_{14}B$	573	
Antiferromagnetic solid		
NiO		523
CoO		293
FeO		198
MnO		69
MnF_2		67
CuF_2		40
α-Fe_2O_3		950
Cr_2O_3		318
$NiCr_2O_4$		65
Co_3O_4		116
K_2NiF_4		97
$LaFeO_3$		750

Ferromagnetic solids, like paramagnetic materials, contain atoms and ions with unpaired electrons, which endow the atoms with a magnetic dipole that is oriented parallel to neighbouring dipoles. The

saturation magnetisation of a ferromagnetic compound, $\mathbf{M}_S$, when all of the dipoles are aligned, can be measured and used to estimate the effective magnetic moment on the atoms in the structure when the structure is known.

$$\mathbf{M}_s = N\,\mathbf{m}_{\text{eff}}\,\mu_B$$

where N is the number of dipoles per unit volume and $\mathbf{m}_{\text{eff}}\,\mu_B$ is the effective magnetic moment on each dipole.

The first theory to account for the existence of ferromagnetic solids was proposed by Weiss. He suggested that an internal 'molecular' field existed inside the magnetic compound and this acted to align the magnetic dipoles of neighbouring atoms. The result was that magnetisation was present even when the magnetic field was zero. Equation (10.1) is replaced by:

$$\mathbf{B} = \mu_0\,(\mathbf{H} + \lambda\mathbf{M})$$

where λ is the *molecular field constant* that indicates the strength of the molecular field. The resulting magnetisation of the solid is still given by Eq. (10.6):

$$\mathbf{M} = N\,g_J\,\mu_B\,J\,B_J(x)$$

but the Brillouin function, $B_J(x)$, is now a function of $\mathbf{M}$:

$$\begin{aligned} x &= g_J\,\mu_B\,J\,\mathbf{B}/k_B T \\ &= g_J\,\mu_B\,J\,\mu_0\,(\mathbf{H} + \lambda\mathbf{M})/k_B T \end{aligned}$$

This equation cannot be solved analytically, and graphical solutions are often utilised (Figure 10.14).

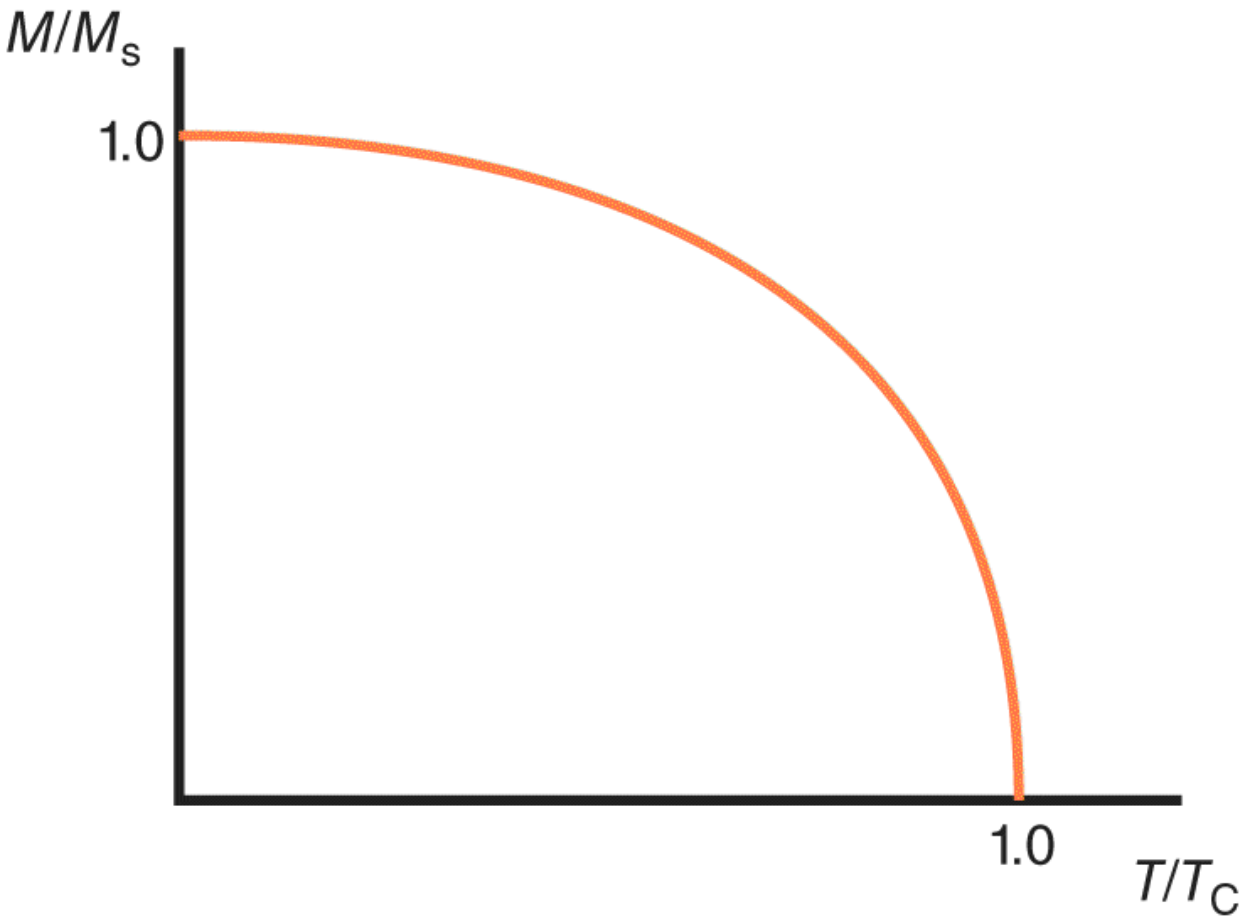


Figure 10.14 Variation of the relative spontaneous magnetisation $\mathbf{M}/\mathbf{M}_S$ as a function of relative temperature, T/T_C.

There are three special cases to note.

1. $T = 0$ K, $\mathbf{M} = \mathbf{M}_S = Ng_J \mu_B J$
2. $T = T_C$, $\mathbf{M} = 0$

$$T_C = \{N\mu_0 {g_J}^2 {\mu_B}^2 J(J+1)\lambda\}/3k_B = \lambda C \tag{10.7}$$

3.

$$T \gg T_C, \chi = \{N\mu_0 {g_J}^2 {\mu_B}^2 J(J+1)\}/ \\ 3k_B(T - \lambda C) = C/(T - T_C) \tag{10.8}$$

The theory gives good agreement with the observations. In particular Eq. (10.7) shows that a high molecular field corresponds to a high Curie temperature, as one would expect. When there is a high interaction between the magnetic dipoles, it will be harder to disrupt the ordering by temperature effects alone. Equation (10.8) shows that ferromagnetic compounds obey the Curie–Weiss law well above a transition temperature, T_C, at which the material loses its ferromagnetic properties.

10.3.2 Exchange Energy

Three 3d transition metals, iron, cobalt, and nickel, are ferromagnetic metals (Table 10.4). The band model (Section 1.4) leads to the expectation that all metals would show Pauli paramagnetism so that the Weiss model is not really applicable to this group and the simple band picture must be expanded to account for this complication. The interaction that leads to ferromagnetic metals is related to the electron interactions that lead to chemical bonding. It is called the *exchange interaction*, giving rise to the *exchange energy*. To illustrate this concept, we will focus upon the important 3d orbitals.

Table 10.4 Magnetic moments and electron configurations of the ferromagnetic 3d-transition metals.

Element	Electron configuration	Magnetic moment (μ_B)
Iron	[Ar] $3d^6\ 4s^2$	2.22
Cobalt	[Ar] $3d^7\ 4s^2$	1.72
Nickel	[Ar] $3d^8\ 4s^2$	0.6

The electron distribution in an atom or an ion with several d electrons results from electrostatic repulsion between these electrons. This interaction is equivalent to the classical Coulomb repulsion between like charges, and is called the *Coulomb repulsion*. The total contribution to the energy is obtained by summing all of the various electron–electron interactions, and is summarised as the *Coulomb integral*. This energy term is decreased if the electrons avoid each other as much as possible. Now electrons with opposite spins tend to occupy overlapping regions of space while electrons with parallel

spins tend to avoid the same regions. Thus, the electrostatic energy is decreased if the electrons all have parallel spins. In this case, as we know, they occupy different d orbitals, far as possible. The energy decrease due to the preference for parallel spins is called the *exchange energy*. This idea has been encountered as Hund's first rule: in a free atom the state of lowest energy (the ground state) has as many electrons as possible with parallel spins. This is the same as saying that the ground state of a free atom has a maximum value of the spin multiplicity, or the equivalent, quantum number *S*.

However, when the atom is introduced into a solid or a molecule, another interaction, *chemical bonding*, is important. Chemical bonding is the result of placing electrons from neighbouring atoms into bonding orbitals. For bonding to occur, and the energy of the pair of atoms to be lowered, the electron spins must be *antiparallel*. However, the exchange interaction between electrons on neighbouring atoms is still present, tending to make the electron spins adopt a parallel orientation. In general, bonding energy is greater than exchange energy, and paired electrons are more often found in solids and molecules. However, if the bonding forces are weak, the exchange energy can dominate and unpaired electrons with parallel spins are favoured.

The 3d transition metals are notable in that the d orbitals do not extend far from the atomic nucleus, and so bonding between d orbitals is weak and the exchange energy is of greater importance. As one moves along the 3d elements, it is found the d orbitals become more compact, making bonding less favourable and increasing the exchange energy. The interplay between bonding energy and exchange energy can be pictured in terms of the ratio of D/d, where D is atomic separation of the interacting atoms and d is the diameter of the interacting d orbitals expressed as a *Bethe–Slater diagram* (Figure 10.15). If the value of the exchange interaction is positive, then ferromagnetism is to be expected. Of all of the 3d-transition metals, only Fe, Co, and Ni have positive values of D/d, and these are the only ferromagnetic 3d-transition elements. The lanthanoids have slightly positive values of D/d and are expected to be weakly ferromagnetic.

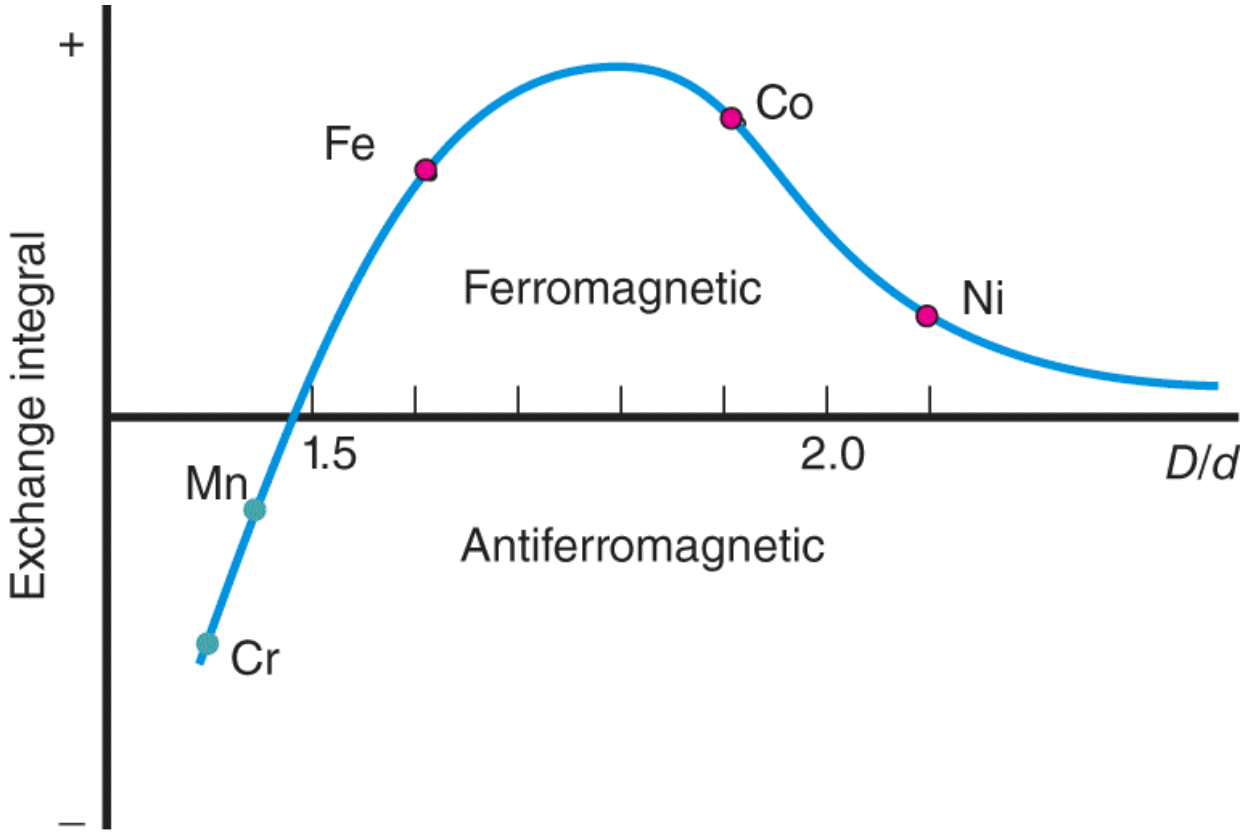


Figure 10.15 The Bethe–Slater curve for the magnitude of the exchange integral as a function of D/d, where D is the separation of the atoms in a crystal and d the extent of the d orbital.

How does this chemical bonding viewpoint link with the band approach? The 3d band is narrow and is overlapped by broad outer bands from the 4s and 4p orbitals (Figure 10.16). Both the s and d electrons will be allocated to this composite band. As electrons are added to the band from the elements K, Ca, Sc, and so on, they occupy the broad mainly s–p low energy part. The number of spin up and spin down electrons would be identical and there is a low density of states at the Fermi level, so that relatively small numbers of electrons are promoted in a magnetic field, exchange energy is low and Pauli paramagnetism results. As more electrons are added, moving from one transition metal to the next (Ti, V, Cr, Mn), the electrons now occupy the narrow d portion of the band. The density of states at the Fermi level starts to rise. Large numbers of electrons can now be promoted and if energetically favourable, reverse spin, leading to greater exchange energy. The tipping point comes with iron. In this

case the density of states at the Fermi level is high enough for the exchange energy to dominate and ferromagnetic ordering occurs. When the density of states at the Fermi surface is calculated with a high precision, it is found that only the three metals iron, cobalt, and nickel will have sufficient exchange energy to retain a ferromagnetic state. The experimental value of the magnetic moment (Table 10.4) is calculated to be equal to the number of holes in the d band. Thus, the d band of iron contains 7.78 electrons (2.22 holes), that of cobalt contains 8.28 electrons (1.72 holes), and that of nickel contains 9.4 electrons (0.6 holes).

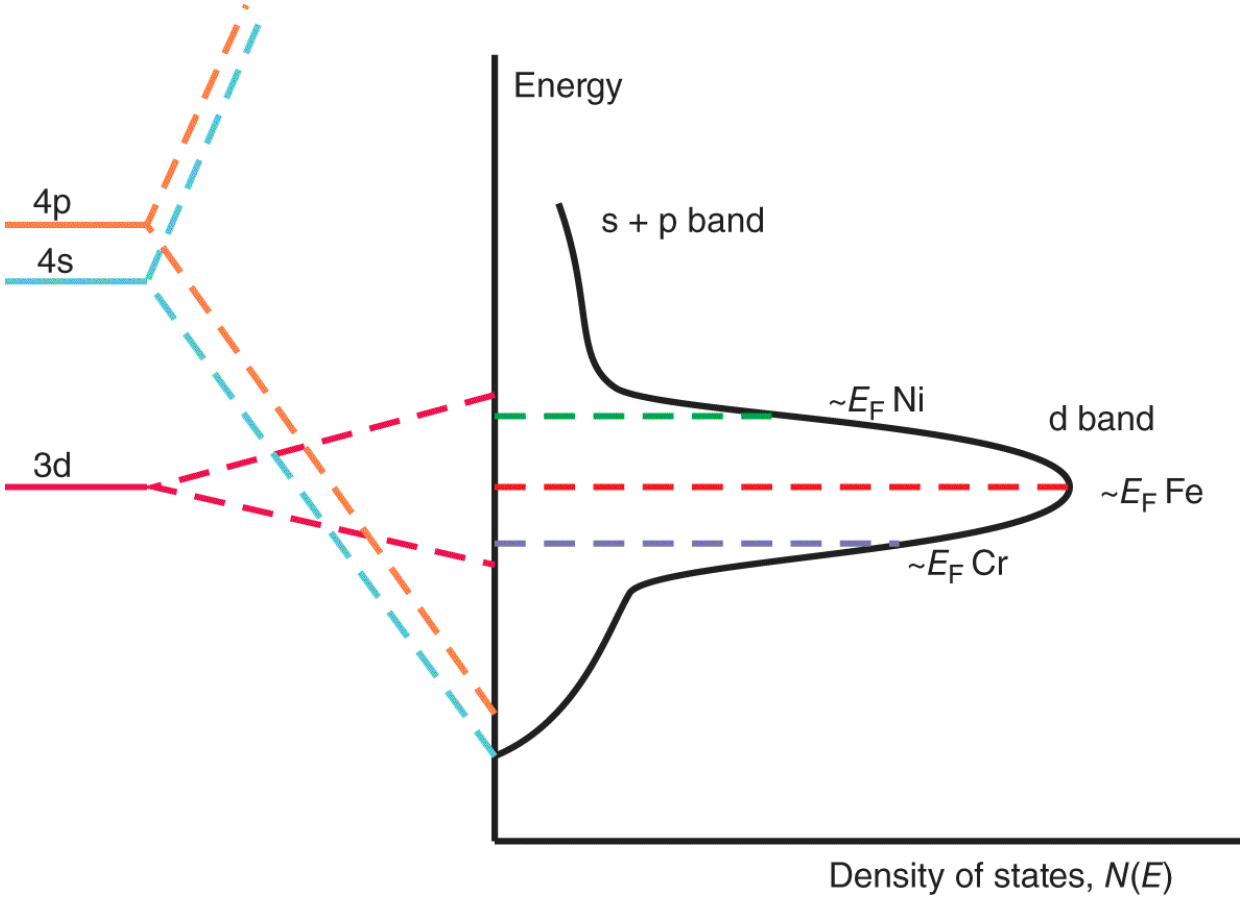


Figure 10.16 Schematic diagram of energy band overlap of 3d, 4s, and 4p orbitals for the 3d transition metals.

The magnetic properties of alloys of these ferromagnetic metals with non-magnetic metals confirm this interpretation. For example, both copper and zinc have 10 d electrons. As they are alloyed with a ferromagnetic metal, the d band is filled and the ferromagnetic properties diminish. These vanish when the d band is just filled with

electrons. A good example is provided by the nickel–copper system, because the alloys all have the same face-centred cubic crystal structure as the parent Ni and Cu phases. The electron configuration of copper is [Ar] $3d^{10}$ $4s^1$, and in this metal, the 3d band is filled. As copper is added to nickel to form the alloy, the excess copper electrons are added to the d band to gradually fill it. At 20 °C, the ferromagnetic–paramagnetic change occurs at a composition of 31.5 wt% copper, approximately Ni_{235} Cu_{100}.

Exact details of the magnetic properties of most alloys depend upon the crystal structures of the alloys and the form of the density of states curve at the Fermi level, as well as the relative proportions of the metals present.

10.3.3 Domains

A characteristic feature of all ferromagnetic and ferrimagnetic[1] substances is that they remain magnetised after the external field is removed and some retain a state of magnetisation almost indefinitely. These are called *magnetically hard* materials and are used to make permanent magnets. However, many ferromagnetic materials appear to lose most of their magnetisation rather easily. These are called *magnetically soft* materials.

If the ordering forces between the atomic magnetic moments are strong enough to lead to ferromagnetism, it is reasonable to ask why soft ferromagnetic materials often show no obvious magnetic properties. The answer lies in the microstructure of hard and soft magnetic solids. On this scale, all such crystals are permanently magnetised but are composed of *magnetic domains* or *Weiss domains*, which are regions that have all of the elementary magnetic dipoles aligned to give an overall magnetisation. In an apparently non-magnetic sample, the alignment directions in neighbouring domains are different, so that the external magnetic field is very small (Figure 10.17a). (Note that domains are distinct from the grain structure of a metal and occur in single crystals as well as polycrystalline ferromagnetic solids.)

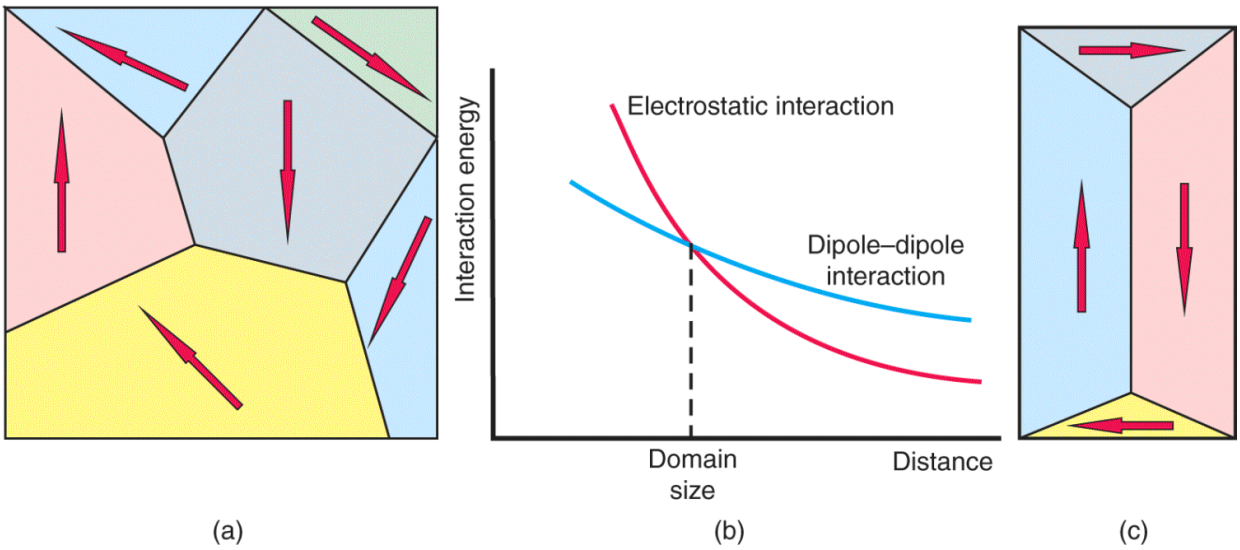


Figure 10.17 Ferromagnetic domains: (a) Weiss domains, schematic; (b) domain size reflecting the balance between dipole–dipole and electrostatic interactions; (c) domain closure.

Two main interactions need to be considered in order to understand domain formation, an *electrostatic* interaction and a *dipole–dipole* interaction. The parallel ordering is caused by the exchange energy. This is effectively electrostatic and *short range*, decreasing approximately as the reciprocal of the distance between atoms. On the other hand, the interaction between the magnetic dipoles is a *long-range* weak bonding force that falls approximately as the cube of the distance between atoms. It leads to a preferred antiparallel arrangement of the dipoles.

When the two interactions are compared, it is found that at short ranges, the electrostatic force is the most important, and an arrangement of parallel spins is of lowest energy. As the distance from any dipole increases, the short-range interaction falls below that of the dipole–dipole interaction. At this distance, the system can lower its energy by reversing the spins (Figure 10.17b). The domains form as a balance between these two competing effects. The balance is achieved for domains of 1–100 μm size. The system can also reduce energy by reducing the external magnetic field arising from the parallel dipoles, the *magnetostatic energy*. This is minimised if the external magnetic field generated by the dipole alignment is reduced by forming the dipoles into *closure domains* with antiparallel orientations (Figure 10.17c). The alignment from one domain to the

adjoining one is not abrupt, but depends on the balance between the electrostatic and dipole interactions. The dipole orientation is found to change gradually over a distance of several hundred atomic diameters. This structure is called a *domain* or *Bloch wall* (Figure 10.18).

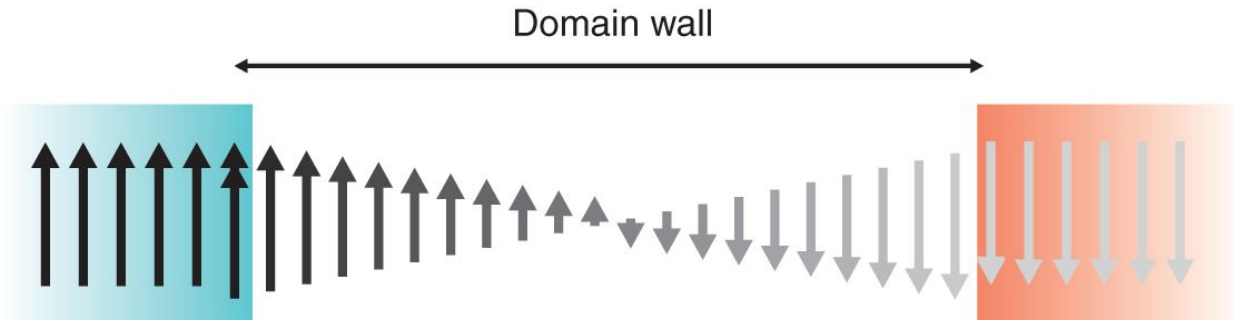


Figure 10.18 A domain (Bloch) wall between two magnetic domains, schematic.

The number of different domain orientations that can occur is determined by the crystal structure and symmetry of the parent phase, and the geometry of the domains varies from one ferromagnetic solid to another. In addition, most ferromagnetic solids are magnetically anisotropic. This means that the magnetic moments align more readily in some crystallographic directions than others. These are known as *easy* and *hard* directions, respectively.

10.3.4 Hysteresis

In general, a ferromagnetic crystal will be composed of an equal number of domains oriented in all the equivalent directions allowed by the crystal symmetry. The overall magnetisation of the crystal will be zero. If we now apply a magnetic field, **H**, in a nominally positive direction, the magnetic dipoles will attempt to re-orient themselves in a direction parallel to the applied field. However, for small values of **H** this tendency is not great enough to overcome the energy barrier between alternative configurations and little change is registered in the magnetic flux density, **B**. This corresponds to the segment 0–a on the graph of the magnetic flux density vs. magnetic field strength (Figure 10.19). As **H** increases, dipoles will start to gain sufficient energy to overcome this energy barrier and will be able jump from one orientation to the other and the dipole direction will start to

switch. Gradually all of the domains will start to change orientation and the observed flux density will now increase rapidly, corresponding to section a–b. Ultimately all of the magnetic dipoles will be aligned parallel and the crystal will (in principle) consist of a *single domain*. This is the state of *saturation*, b–c. On reducing and then reversing the applied magnetic field, the converse takes place. Once again, this will be opposed by the internal energy of the solid, as new domain walls have to be created and domain walls must move. The **B–H** path will therefore not follow the original path, but will trace a new path that lags behind the old one, following c–d–e, to reach saturation with dipoles pointing in the opposite direction, at f (Figure 10.19). When the applied magnetic field strength reaches zero, the value of **B** is still positive, and this is called the *remanent* or *residual flux density* (*remanence*) B_r. The magnetic flux density, **B**, will be reduced to zero at a value of the applied field called the *coercive field* or *coercivity*, H_c. (A magnet taken through the path to saturation can only be demagnetised by subjecting it to a negative magnetic field equal in magnitude to the coercive field.) Ultimately domain alteration will cease when all the magnetic moments are now parallel to the reversed applied field and saturation is again reached. Reversal of the magnetic field again causes a reversal of dipole direction, and the curve will follow the path f–g–c. This closed circuit is called a *hysteresis loop* and the phenomenon is that of *hysteresis* (the definition of hysteresis is 'to lag behind'). Repeated cycles now follow this outer pathway. Solids that show hysteresis and a domain structure are termed *ferroic* materials (also see Section 9.3).

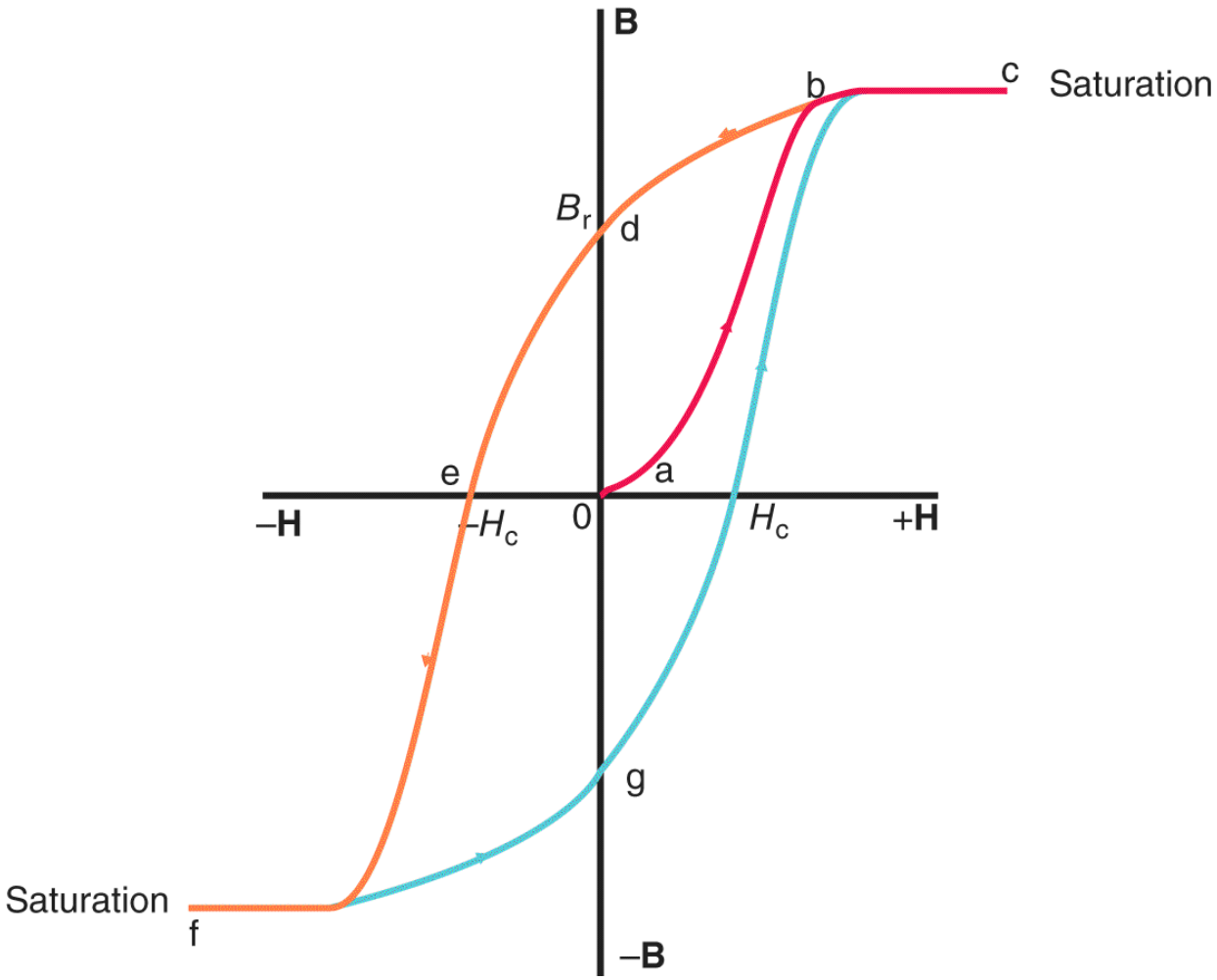


Figure 10.19 The **B**–**H** (hysteresis) loop of a ferromagnetic solid.

10.3.5 Hard and Soft Magnetic Materials

The shape of the hysteresis loop for a ferromagnetic material is of importance. The area of the loop is a measure of the energy required to complete a hysteresis cycle. A soft magnetic material is one that is easily magnetised and demagnetised. The energy changes are low and the hysteresis loop has a small area (Figure 10.20). Soft magnetic materials are used in applications such as transformers, where the magnetic behaviour must mirror the variations of an electric current without large energy losses. Ferrites with the cubic spinel structure (Section 10.5.1) are generally soft magnetic materials. The value of the saturation magnetisation is sensitive to the composition of the material and the formulae of cubic ferrites are carefully tailored to give an appropriate figure. However, the shape of the hysteresis loop

is strongly influenced by the microstructure of the solid. Grain boundaries and impurities hinder domain wall movement, and so change the coercivity and remanence of a sample. Ferrites for commercial applications have to not only have carefully chosen compositions, but also to be carefully fabricated in order to reproduce the desired magnetic performance.

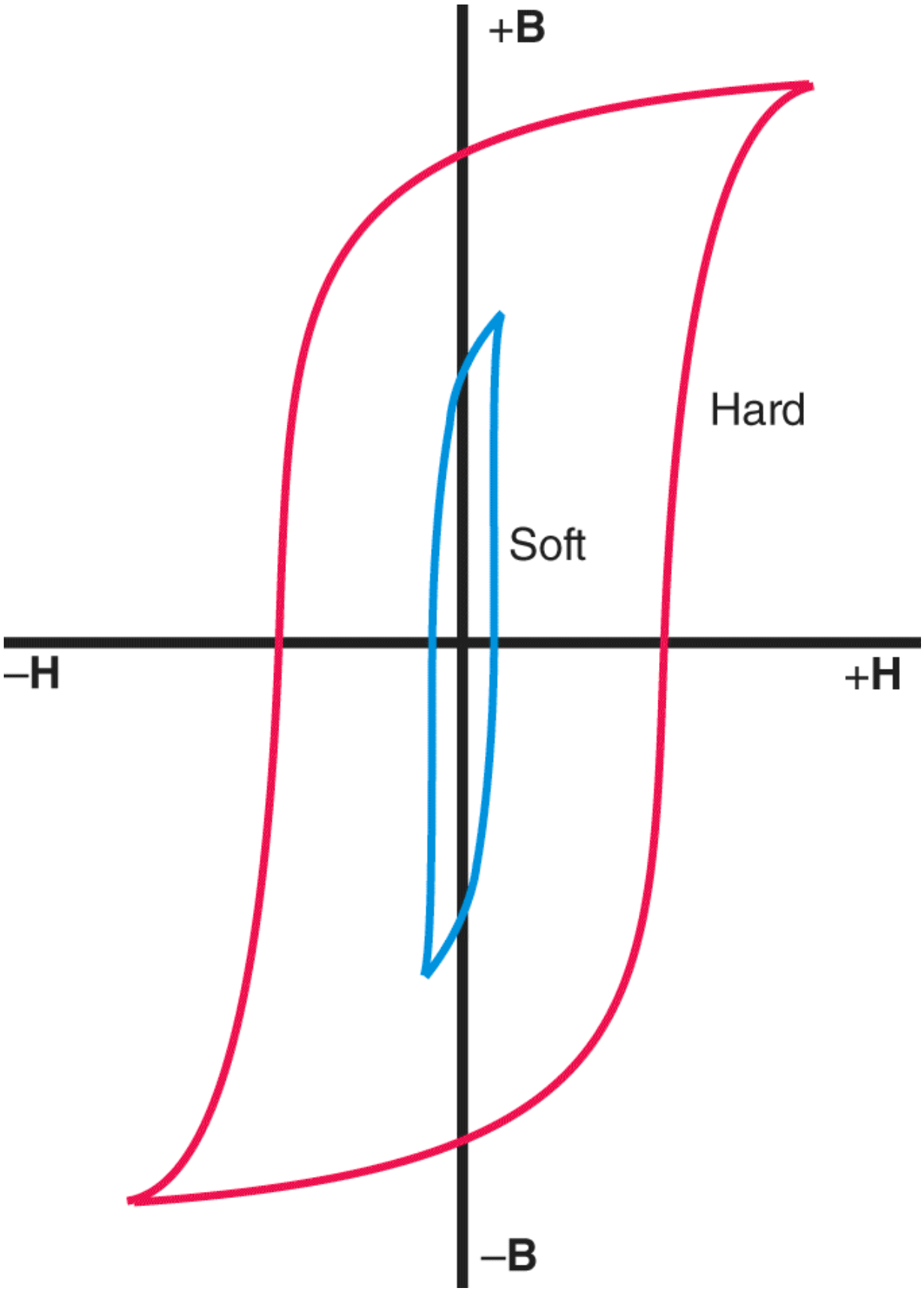

+B
Hard
Soft
–H
+H
–B

Figure 10.20 Schematic **B**–**H** loops for soft and hard magnetic solids.

Hard magnetic materials have rather rectangular broad hysteresis curves (Figure 10.20). These materials are used in permanent magnets, with applications from door catches to electric motors. The hexagonal ferrites (Section 10.5.3) are hard magnetic materials, as are alloys such as $SmCo_5$ and $Nd_2Fe_{14}B$.

Naturally, there is a continuum of magnetic materials between soft and hard, and the characteristics of any particular solid are tailored to the application. Of these, magnetic storage of information is of great importance. The data is stored by magnetising small volumes of material, which may be a thin film deposited on a solid surface or a flexible tape. The magnetic material that receives the signal to be stored must be soft enough to respond quickly to small energising electric signals, but be hard enough to retain the information for long periods.

10.4 Antiferromagnetic Materials and Superexchange

Most antiferromagnetic materials are compounds in which magnetic transition metal ions are separated by non-magnetic atoms. Among the earlier materials so characterised were the alloys CuMnAs, Mn_3Pt, Mn_3Ir, and Mn_3Sn. In these materials the magnetic moment resides on the Mn sublattice only. In Mn_3Pt, for example, the room temperature structure consists of a face-centred cubic atom array with Mn atoms lying on (111) planes. The magnetic moments are in a canted antiferromagnetic arrangement on each plane, which changes to an antiferromagnetic structure in which the spins are all antiparallel at approximately 360 K (Figure 10.21).

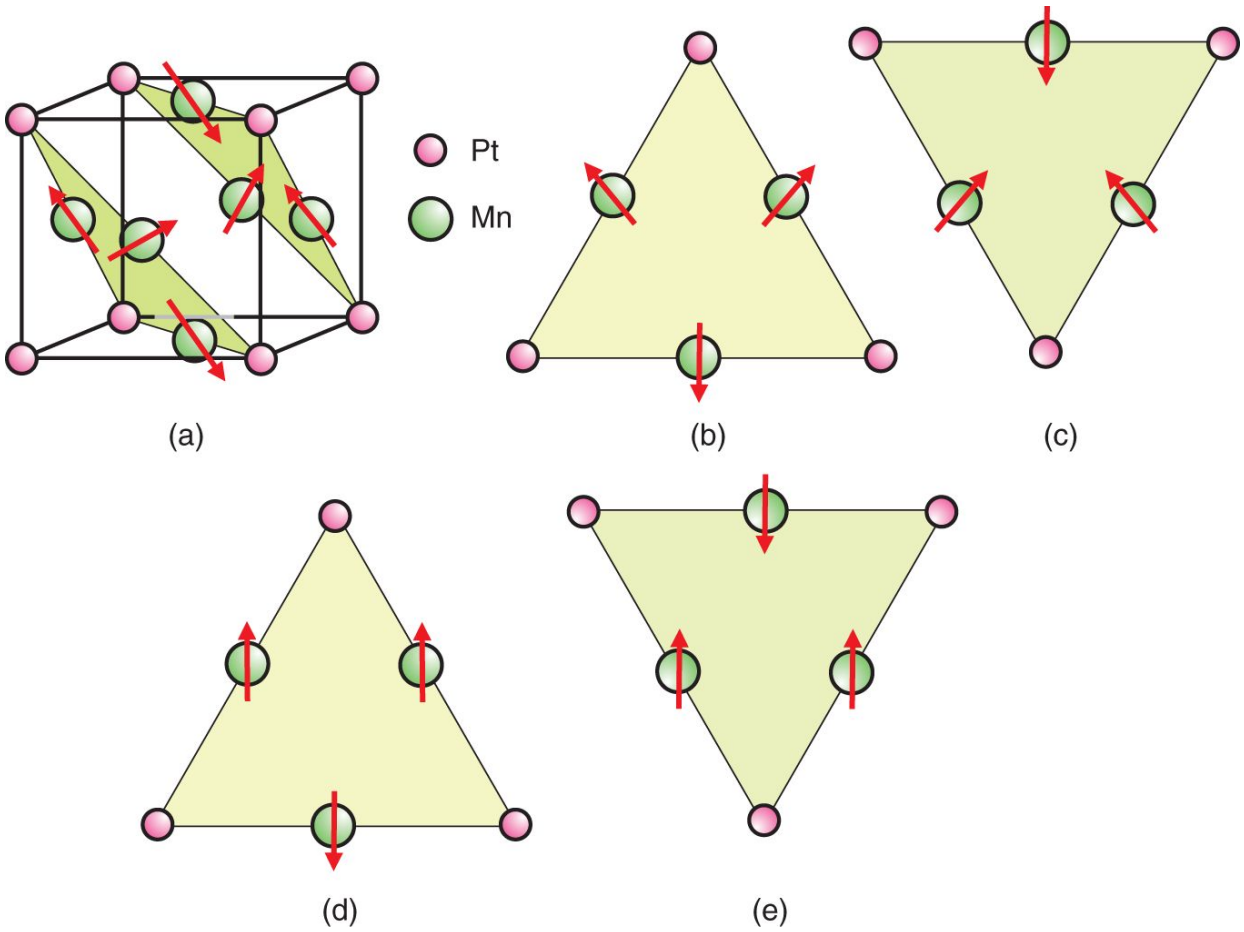


Figure 10.21 (a) The room temperature face-centred cubic structure of Mn_3Pt; (b, c) (111) planes showing a canted spin arrangement on the Mn atoms; (d, e) antiferromagnetic structure above approximately 360 K.

Antiferromagnetism is widespread in oxides containing 3d transition metal cations, including simple oxides such as NiO, as well as more complex structures such as the cubic ferrites AFe_2O_4 and oxide perovskites such as $LaMnO_3$ and $LaCoO_3$. These oxide antiferromagnetic compounds are usually insulators. The interaction of the d orbitals on the cations that produce dipole alignment now takes place via an intermediate anion, in a process called *superexchange*.

The effect can be described with respect to the transition metal oxide nickel oxide, NiO, a typical antiferromagnetic structure. In NiO the Ni^{2+} ions have eight d electrons, two of which, the $d_{x^2-y^2}$ and d_{z^2}

orbitals, contain one unpaired electron due to the crystal field of the surrounding octahedron of oxide anions. The oxide would be expected to be paramagnetic, as the Ni^{2+} ions are separated from each other by non-magnetic oxygen ions. Instead, antiferromagnetic ordering occurs and the magnetic susceptibility increases with temperature up to 250 °C, before showing paramagnetic behaviour. The coupling between the ions leading to an antiferromagnetic ordering is related to a degree of covalent bonding between the Ni^{2+} and O^{2-} ions. An unpaired d electron in the $3d_{x^2-y^2}$ orbital, for example, can have a covalent interaction with an electron in a filled p orbital, but only if the electrons have opposed spins. The oxygen spin down electron will interact with a d orbital in which the electron is spin up. This is true for both Ni^{2+} ions either side of an oxygen ion. In this case, as shown, the arrangement leads to an antiferromagnetic alignment of the unpaired d electrons on the Ni^{2+} ions (Figure 10.22). This type of electron-spin coupling occurs frequently and superexchange tends to lead to an antiferromagnetic ordering in a solid. Note that in superexchange the electrons remain in their respective orbitals and no electron transfer takes place.

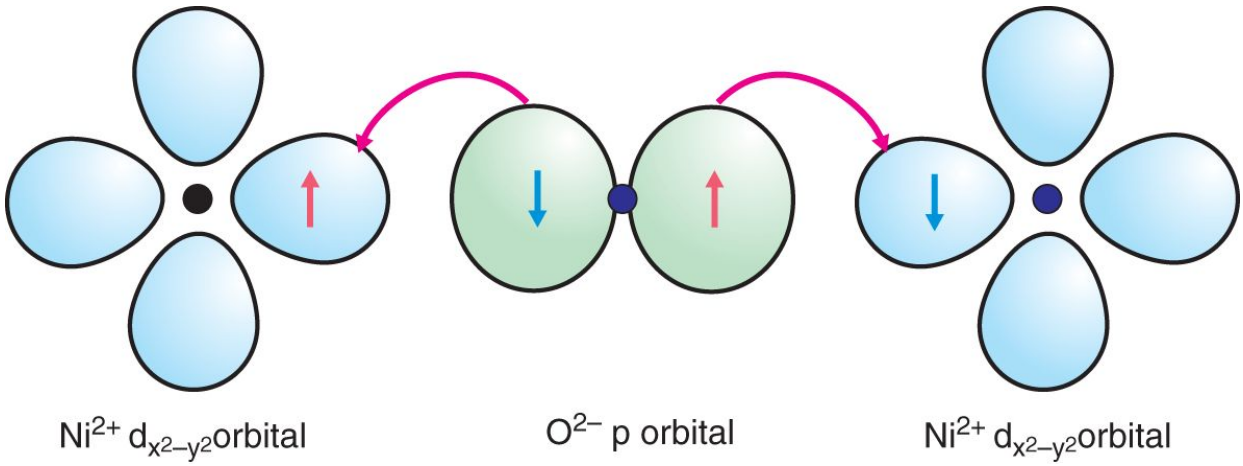


Figure 10.22 Superexchange between Ni^{2+} and O^{2-} (schematic) leading to antiferromagnetic alignment of cation spins.

It would be expected that the Néel temperature would increase with the strength of the covalent interaction between the cations and anions and this is found to be so. The Néel temperature for the series of 3d transition-metal oxides MnO, FeO, CoO, and NiO increases in

the direction from MnO to NiO, as the covalent nature of the solids increase (Table 10.3). Additionally, the dumbbell shape of the p orbitals suggests that the interaction should be greatest for a linear M–O–M configuration and minimum for a 90° M–O–M configuration. The oxides MnO, FeO, CoO, and NiO all have the halite (NaCl) structure in which the linear M–O–M configuration holds, with metal ions at the corners of the cubic unit cell and oxygen ions at the centre of the cell edges.

Antiferromagnetic materials are currently being explored for magnetic storage devices, as these compounds have a low sensitivity to external magnetic fields and very fast spin switching dynamics.

10.5 Ferrimagnetic Materials

Ferrimagnetic materials are characterised by two subsets of magnetic dipoles, with one of the subsets in an antiparallel arrangement with respect to the first (Figure 10.7). The observed magnetism is the due to the difference between the two sets of magnetic moments. In this section attention is focussed upon oxide materials. As with oxide antiferromagnetic compounds, these can only be understood in terms of both the crystal structures and the chemical bonding in the compounds.

10.5.1 Cubic Spinel Ferrites

The importance of crystal structure can be illustrated with reference to the cubic ferrites, $A^{2+}Fe_2^{3+}O_4$ in which A is typically a medium sized cation such as Ni^{2+}. These compounds comprise an important group of soft magnetic materials widely used in electronic circuitry. They adopt the inverse spinel structure in which the cations are distributed between tetrahedral sites and octahedral sites to give a formula $(Fe^{3+})[A^{2+}Fe^{3+}]O_4$, where (Fe^{3+}) represents cations in tetrahedral sites, and $[A^{2+}Fe^{3+}]$ represents cations in octahedral sites (Section 3.2.2.8). Thus nickel ferrite would be written (Fe^{3+}) $[Ni^{2+}Fe^{3+}]O_4$. Lodestone, or magnetite, Fe_3O_4, is an example in which the cations are Fe^{2+} and Fe^{3+}, and the cation distribution is $(Fe^{3+})[Fe^{2+}Fe^{3+}]_2O_4$.

In all these materials the magnetic moment of the Fe^{3+} ions in the tetrahedral sites is opposed to that of the Fe^{3+} ions in the octahedral sites, so that the net magnetic moment due to Fe^{3+} is zero. Nickel ferrite could be written schematically as $Fe^{3+}\uparrow[Fe^{3+}\downarrow Ni^{2+}\uparrow]O_4$ and the structure is ferrimagnetic with the net magnetic moment due to the Ni^{2+} ions alone. The oxide Fe_3O_4 is written $Fe^{3+}\uparrow[Fe^{3+}\downarrow Fe^{2+}\downarrow]O_4$ and the overall magnetic moment of the compound is due to the Fe^{2+} contribution.

The saturation magnetisation, $\mathbf{M}_S$, of these spinels can be estimated by assuming the spin only magnetic moments, m_{ion}, listed in Table 10.2 apply. If all the excess A^{2+} moments are completely aligned:

$$\mathbf{M}_s = N\, m_{ion}\, \mu_B$$

where N is the number of unpaired dipoles per unit volume and m_{ion} μ_B is the magnetic moment on each dipole.

The magnetic properties of cubic ferrites can be altered by forming solid solutions to produce compounds with a precise and unique magnetic signature. The structure gives four degrees of freedom to explore, magnetic or non-magnetic impurity cations on tetrahedral sites, and magnetic or non-magnetic impurities on octahedral sites. Complex mixtures of cations, for example, $Ni_aMn_bZn_cFe_dO_4$, are often used to tailor quite specific magnetic behaviour. However, no matter how complex the formula, the total must be in accord with the spinel formula, $A^{2+}B^{3+}{}_2O_4$, both with respect to the numbers of ions but also with respect to the charges.

The flexibility of the system can be illustrated by the solid solution between the normal spinel $ZnFe_2O_4$ and the inverse spinel $NiFe_2O_4$. In $ZnFe_2O_4$ the magnetic moments of the individual Fe^{3+} ions are opposed even though they both occupy the octahedral sites, and the formula can be written $Zn^{2+}[Fe^{3+}\uparrow Fe^{3+}\downarrow]O_4$. As expected, the phase is paramagnetic. Suppose that this material is now reacted to form a solid solution with $NiFe_2O_4$ thus:

$$(1-x)\,NiFe_2O_4 + x\,ZnFe_2O_4 \rightarrow Zn_xNi_{1-x}Fe_2O_4$$

The rather straightforward formula hides an intriguing situation. This is made clear if we write out the formula as:

$$x\,Zn^{2+}[Fe^{3+} \downarrow Fe^{3+} \uparrow]O_4$$
$$+(1-x)Fe^{3+} \uparrow [Fe^{3+} \downarrow Ni^{2+} \uparrow]O_4$$
$$\rightarrow Zn_x^{2+}Fe_{1-x}^{3+} \uparrow [Fe^{3+} \downarrow Ni_{1-x}^{2+} \uparrow Fe_x^{3+} \uparrow]O_4$$

When x is 0 the material has a magnetism due only to the Ni^{2+} ions. When x is 1.0 the material is paramagnetic. In between, the magnetism is a steady function of the Ni^{2+} concentration and Ni^{2+} is akin to a *magnetic defect*. In fact, the distribution of cations in spinels is rarely perfectly normal or inverse, and the distribution tends to vary with temperature. Moreover, as the magnetic properties depend sensitively upon the cation distributions, processing conditions are important if the desired magnetic properties are to be obtained. Experimentally it is found that the defect interactions are complex and the observed magnetism rises to a maximum when x is near to 0.5. Nickel–zinc ferrites of about this composition are used in recording heads for audio and video recorders.

10.5.2 Garnet Structure Ferrites

Garnet is a general name for a group of minerals with a formula $A_3B_2Si_3O_{12}$ in which A^{2+} is a large cation occupying eight-coordinate sites, B^{3+} is a medium sized cation occupying octahedral sites, and Si^{4+} occupies tetrahedral sites. In the garnet ferrites A^{2+} is replaced by a lanthanoid cation Ln^{3+} and the B^{3+} and Si^{4+} are replaced by Fe^{3+} to give a general formula $Ln_2Fe_5O_{12}$. As with the spinels, the Fe^{3+} ions are distributed between two sites, octahedral and tetrahedral, and the spins in the two are in opposition to each other. However, in the garnets there are three tetrahedral to two octahedral cations, so that there is still some residual magnetism due to the Fe^{3+} component. The materials can then be given stronger magnetic properties by correct choice of lanthanoid cation with unpaired 4f electrons and by partial substitution of the Fe^{3+} ions. For example, the garnet $EuEr_2Ga_{0.7}Fe_{4.3}O_{12}$ was once a contender for magnetic memory storage.

10.5.3 Hexagonal Ferrites

In contrast to the soft magnetic spinel ferrites, another group of ferrites is known that have hard magnetic properties. They are derived from the *magnetoplumbite* ($Pb^{2+}Fe^{3+}_{12}O_{19}$) structure, containing Fe^{3+} ions. The unit cells of these phases are hexagonal; hence they are commonly called *hexagonal ferrites* and are also known as *ceramic magnets*. The variety of ferrites derived from magnetoplumbite and the fine-tuning of magnetic properties is achieved by cation substitution using M^{2+} and M^{3+} ions in place of Pb^{2+} and Fe^{3+} cations.

The two simplest hexagonal ferrites, $BaFe_{12}O_{19}$, known as *ferroxdure*, and $SrFe_{12}O_{19}$, are widely used as permanent magnets in many applications, especially in electric motors. The structure of both, the magnetoplumbite type, is built from a close packing of oxygen and oxygen plus barium BaO_3 layers (Figure 10.23). There are 10 of these layers in the unit cell, which contains two $BaFe_{12}O_{19}$ formula units. The relative positions of the close-packed layers are: ...A B′ A B C A C′A C B..., where the symbols B′ and C′ represent BaO_3 layers, and the other symbols represent oxygen only (O_4) layers. These 10 sheets are thought of as arranged in a sequence of four alternating slices, two that resemble spinel, called S blocks, and two containing the BaO_3 layer, called R blocks. Half a unit cell consists of one R block and one S block. (In the lower half of the cell the R and S blocks are inverted with respect to those shown, but the site geometries remain the same.) The Fe^{3+} cations occupy nine octahedral sites, two tetrahedral sites, and one trigonal pyramidal (five-coordinate) site per $BaFe_{12}O_{19}$ unit. The arrangement of the spins on the Fe^{3+} ions in these sites is such that four point in one direction parallel to the hexagonal **c**-axis and the other eight point in the opposite direction. As with the garnets, the spins are all of the same magnitude and unbalanced. Because the spins are aligned along the crystallographic **c**-axis, the materials show a high degree of anisotropy.

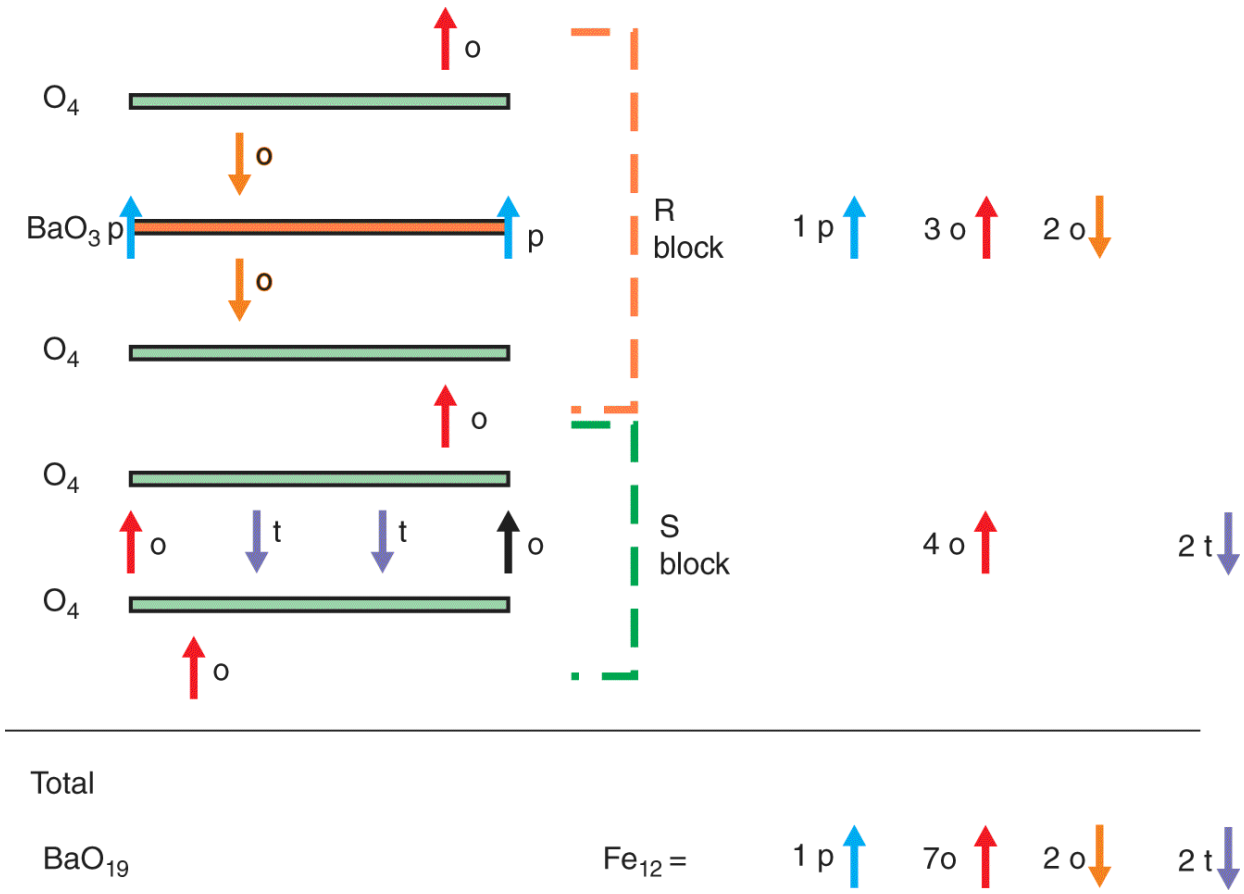


Figure 10.23 Schematic representation of the magnetic spin arrangement in $BaFe_{12}O_{19}$, with the **c**-axis vertical. The Fe ions are in octahedral (o), tetrahedral (t), and trigonal prismatic (p) sites. Note that the (o) sites on the boundaries between the R and S blocks count double, as one site obscures a second, and that the (p) sites count as ½ as these lie on the unit cell edge.

In the hexagonal ferrites hundreds of structures occur. This vast multiplicity arises because a large number of arrangements of R and S blocks are possible. The resulting phases, known as *polytypes*, often have enormous unit cells.

10.5.4 Double Exchange

A number of ferrimagnetic ferrites are good electrical conductors and the same is true of the cobaltites and manganites. Appreciable electrical conductivity implies that electrons are fee to move through the structure. The magnetic properties of insulators (such as the antiferroelectrics) can be explained by ionic models invoking

superexchange. The difficulty is how to explain ferromagnetic alignment in electrical conductors where electron movement through the solid occurs. In the case of metals, exchange energy was invoked. In the conducting oxides a mechanism called *double exchange* is required.

The idea will be illustrated for magnetite, Fe_3O_4. Magnetite crystals appear metallic and show quite good electronic conductivity; approximately $2.5 \times 10^4\ \Omega^{-1}\ m^{-1}$, poorer than most metals (Ti, $2.38 \times 10^6\ \Omega^{-1}\ m^{-1}$) but better than elemental semiconductors (Ge, 2.17 $\Omega^{-1}\ m^{-1}$). In Fe_3O_4, the cations that are ferromagnetically aligned are restricted to those on the octahedral sites, equal numbers of Fe^{2+} and Fe^{3+}. (The population of Fe^{3+} ions in the tetrahedral sites is not involved, but see the following text.) The coupling between these ions must link the spins on two ions separated by an anion such as oxygen and also allow for easy electron movement. In double exchange, an Fe^{2+} ion *transfers* an electron to an adjacent O^{2-} ion, but as this has filled p orbitals, to make this possible the O^{2-} ion simultaneously transfers one of its electrons to a neighbouring Fe^{3+} ion. In essence, the electron hops from Fe^{2+} to Fe^{3+} via the intermediate oxygen ion (Figure 10.24). However, the charges on the Fe ions are reversed in the process. Because all of the orbitals on the oxygen ion are full, and electrons are spin paired, a spin down electron moving from Fe^{2+} will displace a spin down electron from O^{2-} onto Fe^{3+}. This double exchange is only favourable if there are parallel spins on the two cations involved, in which case it leads to ferromagnetic alignment. Note that in double exchange *electron transfer* from one cation to another takes place (quite unlike the situation in superexchange), and ferromagnetic alignment always results.

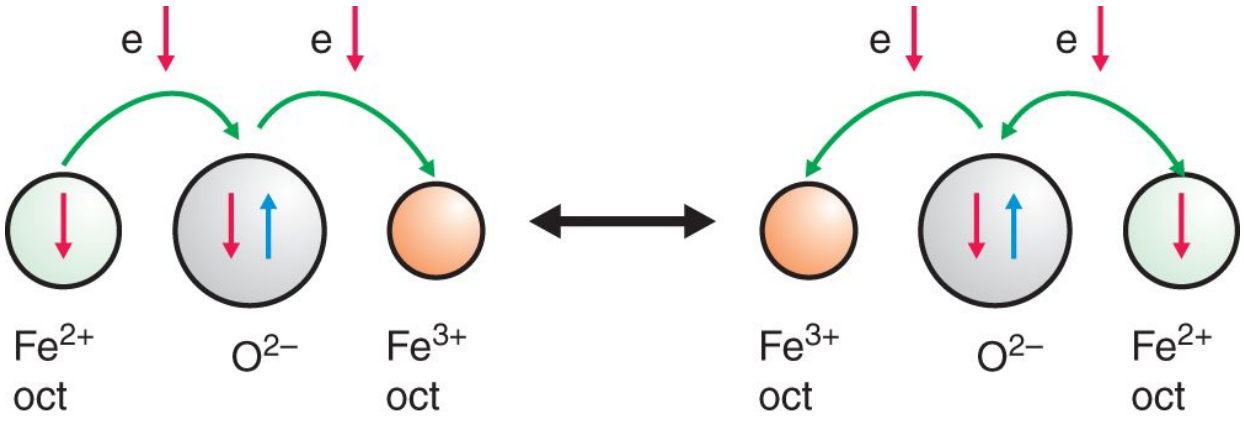


Figure 10.24 Double exchange (electron transfer) between Fe^{2+} and Fe^{3+} (schematic).

To return to magnetite, double exchange is found not to operate between cations in tetrahedral sites and those in octahedral sites. The reason for this separation lies with the d orbital energies, which are controlled by the crystal field splitting. The d orbital energy of a cation in an octahedral site is quite different from that of a cation in a tetrahedral site and this difference prevents double exchange between Fe^{2+} ions in octahedral sites and Fe^{3+} ions in tetrahedral sites. However, superexchange *is* possible between the occupants of the octahedral and tetrahedral sites, and this produces antiferromagnetic ordering between these cations.

In general, double exchange can operate in compounds that have a cation in two valence states. Many crystalline structures potentially fulfil these requirements, including the ion pairs (Mn^{4+}, Mn^{3+}) and (Co^{2+}, Co^{3+}) (Section 10.7.2).

10.6 Nanostructures

10.6.1 Small Particles and Data Recording

The magnetic properties of very small particles have been of interest for a considerable period. Because ferromagnetic properties rely on long-range interactions, it is of interest to follow these as particle size decreases. It has been found, for example, that nickel nanoparticles with diameters in the range of 20–60 nm remain ferromagnetic. However, the saturation magnetisation decreases at the smallest dimensions. (Similar effects are noted with ferroelectric particles, Section 9.3.9.)

The earliest application of small magnetic particles was for magnetic recording media. The essence of magnetic data storage is the existence of two easily distinguished magnetic states. Magnetic data storage uses a thin magnetisable layer laid down on a tape or disc. A *write head* generates an intense magnetic field that changes the direction of magnetisation in the surface material. The induced magnetisation pattern in the layer is the stored data. The data is read by sensing the magnetic field changes using a *read head*.

The data storage layers are mostly composed of small magnetic particles in a polymer film. The magnetic response of the film depends critically on the domain structure of the particles and the particle shape. Ideally, small single domain particles are used, each of which has only two directions of magnetisation, directed along the + and – directions of a single crystallographic axis. The direction is switched by the write head and sensed by the read head.

The commonest magnetic particles in use at present are γ-Fe_2O_3 (maghemite), cobalt doped γ-Fe_2O_3, and chromium dioxide, CrO_2, all of which have acicular (needle-shaped) crystals. The crystal structure of γ-Fe_2O_3 is curious. Despite its formula, it has the spinel structure, and is closely related to the inverse spinel magnetite $(Fe^{3+})[Fe^{2+}Fe^{3+}]O_4$. γ-$Fe_2O_3$ is a ferrimagnetic spinel that has a formula $(Fe^{3+})[Fe^{3+}_{5/3}\ V_{1/3}]O_4$, where V represents vacancies on octahedral Fe^{2+} sites. It is thus a ferrite with an absence of M^{2+} ions. Chromium

dioxide is a ferromagnetic oxide with the rutile structure. The direction of magnetisation in all of these compounds corresponds to the needle axis. The crystallites are aligned in a collinear array with the needle axis parallel to the direction of motion of the tape or disc by applying a magnetic field during the coating operation, known as *longitudinal orientation* or direction.

10.6.2 Superparamagnetism and Thin Films

Domain walls are created because of the competition between long-range dipole–dipole interactions, which lead to an antiparallel alignment of magnetic dipoles, and short-range electrostatic interactions, which tend to produce a parallel alignment of magnetic dipoles. When particles of a magnetic solid are below the domain size, the electrostatic interactions dominate. The magnetic dipoles tend to align parallel to each other, and a *superparamagnetic* state results. This need not only occur in small particles. Isolated clusters of magnetic atoms or ions in a solid can also exhibit the feature. For example, nanoclusters of ferromagnetic particles embedded in polymers often show superparamagnetic behaviour, as can small magnetic precipitates in a non-magnetic matrix. The same seems to be true of the magnetic particles that occur in certain bacteria that are able to navigate along the Earth's magnetic field gradient.

The same interactions lead to the observation that the magnetisation in thin films changes with film thickness. In thick films, of ordinary dimensions, domains form and the magnetic flux is trapped in the film (Figure 10.17c). When the thickness of a film is reduced to below single domain size, the magnetic dipoles align in a parallel direction in a longitudinal manner. The film has a north and south pole at its extremities (Figure 10.25a). An external magnetic field is now detectable that runs more or less parallel to the film surface. Further reduction in film thickness to just a few atom layers results in the individual electron spins re-aligning normal to the film plane. North and south poles are now on the film surfaces, and the external magnetic field lies perpendicular to the film plane (Figure 10.25b). The implications of these scale effects are of importance in attempts to increase the density of magnetic recording media.

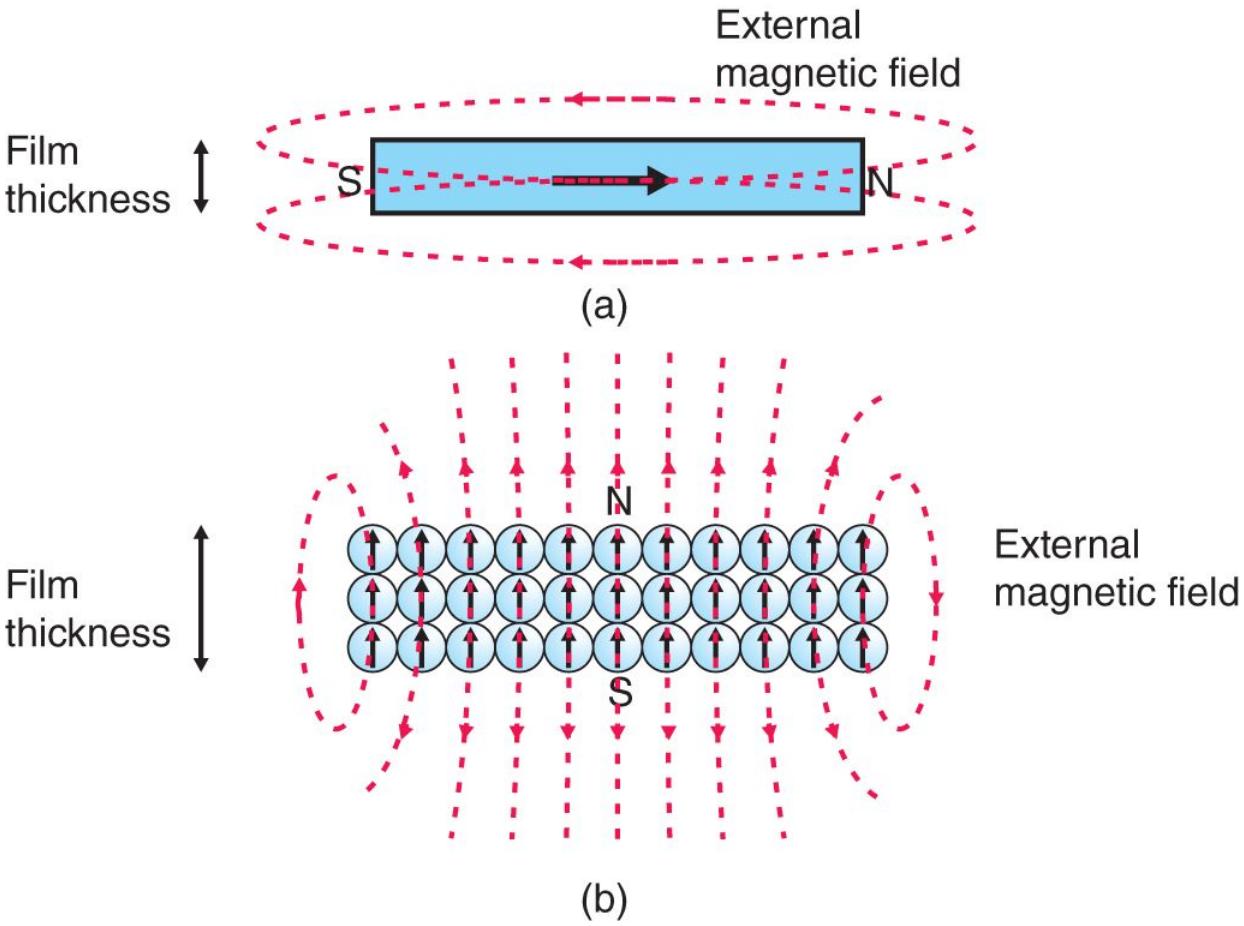


Figure 10.25 (a) Films less than a domain wide have magnetic dipoles aligned so that the magnetic flux escapes longitudinally. (b) An atomically thin film has magnetic dipoles aligned perpendicular to the film thickness so that the magnetic flux escapes normal to the film.

The magnetic properties of layers can also be investigated via layered structures such as MoS_2 (Figure 2.4c) and TiS_2 (Figure 3.36), so called van der Waals solids. These materials are composed of stacks of sheets, each of which consists of a pair of close packed (...AB...) anion layers containing a metal cation in each of the octahedral holes between them. The sheets are held together by weak van der Waals bonds. Because of this weak inter-sheet bonding, the solids can be cleaved to give single sheets. The magnetic (and electronic) properties of such single sheets are of great interest. For example, in the phase CrI_3, which has a structure similar to TiS_2, but with only 2/3 of the octahedral positions between the two close-packed anion layers filled

in an ordered fashion (Figure 10.26), a single CrI_3 sheet is ferromagnetic in a direction perpendicular to the sheet, while two sheets stacked as in a macroscopic crystal shows antiferromagnetic behaviour. In the structurally similar phase $CrBr_3$, a single $CrBr_3$ sheet is again ferromagnetic in a direction perpendicular to the sheet. However, when two stacked sheets are investigated, the magnetism depends upon the relative orientations of the two. One stacking, equivalent to a macroscopic rhombohedral crystal structure, remains ferroelectric, while an alternative stacking, equivalent to a hexagonal macroscopic crystal, is antiferromagnetic.

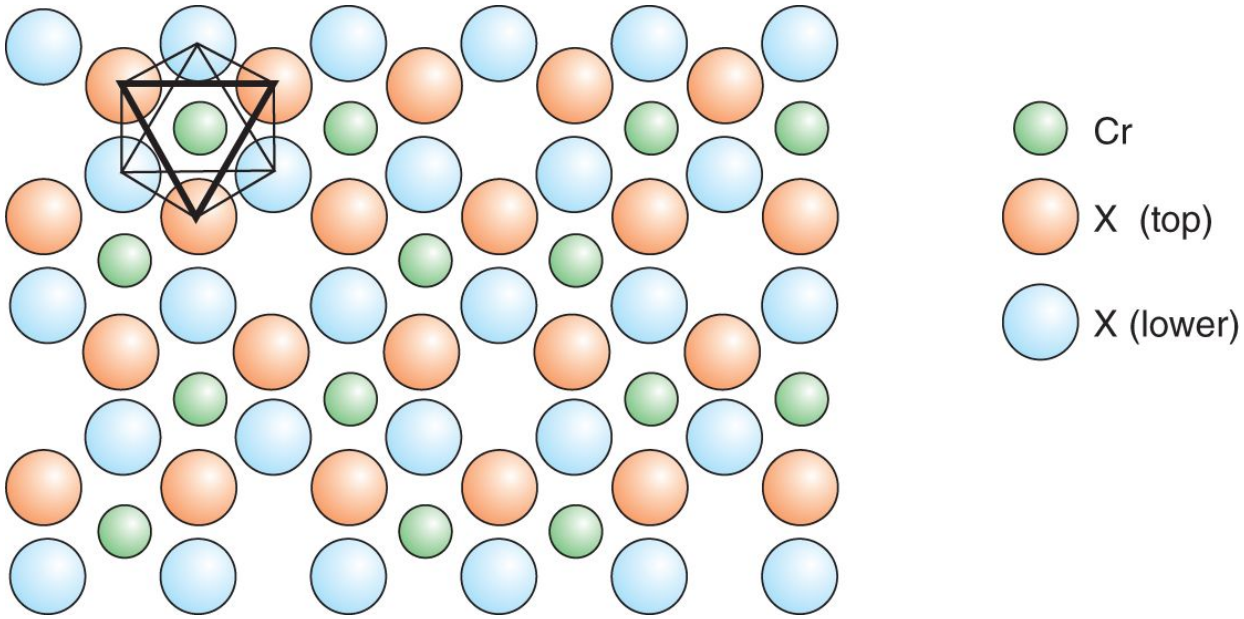


Figure 10.26 A single sheet of the CrX_3 structure (X = Cr, Br, I).

10.6.3 Perovskite Superlattices

Just as with ferroelectrics, superlattices consisting of alternating layers, a few unit cells thick, of two or three perovskite structure materials, grown sequentially so that a perfect crystal is produced, are eminently suitable for the investigation of magnetic properties and their modification via interface engineering. The properties of these superlattices can be tuned by varying the chemical constituents (ABO_3, $A'B'O_3$) of the layers and by changing the relative proportions of each layer, m unit cells of ABO_3 and n unit cells of $A'B'O_3$ (see, i.e. Section 9.3.9 and Figure 9.30).

An example of the way in which interfaces can change properties significantly is given by $LaNiO_3$, a paramagnetic metal and $LaMnO_3$, an antiferromagnetic insulator. These have been grown in a (111) orientation (Figure 10.27). $LaNiO_3$ contains the unusual valence state Ni^{3+}, a $3d^7$ ion. $LaMnO_3$ contains Mn^{3+}, a $3d^4$ ion with the electron distribution $t_{2g}^3 e_g^1$. In films with m and n both greater than 3, the superlattice becomes ferromagnetic although *neither* of the parent oxides is ferromagnetic. The underlying cause is that the less stable Ni^{3+} accepts the e_g electron from Mn^{3+} to give a more stable pairing of Ni^{2+} ($3d^8$) and M^{4+} ($3d^3$). These ions then order ferromagnetically in the interfacial regions. Provided that enough layers are present, the ferromagnetic contribution dominates the magnetic properties of the superlattice.

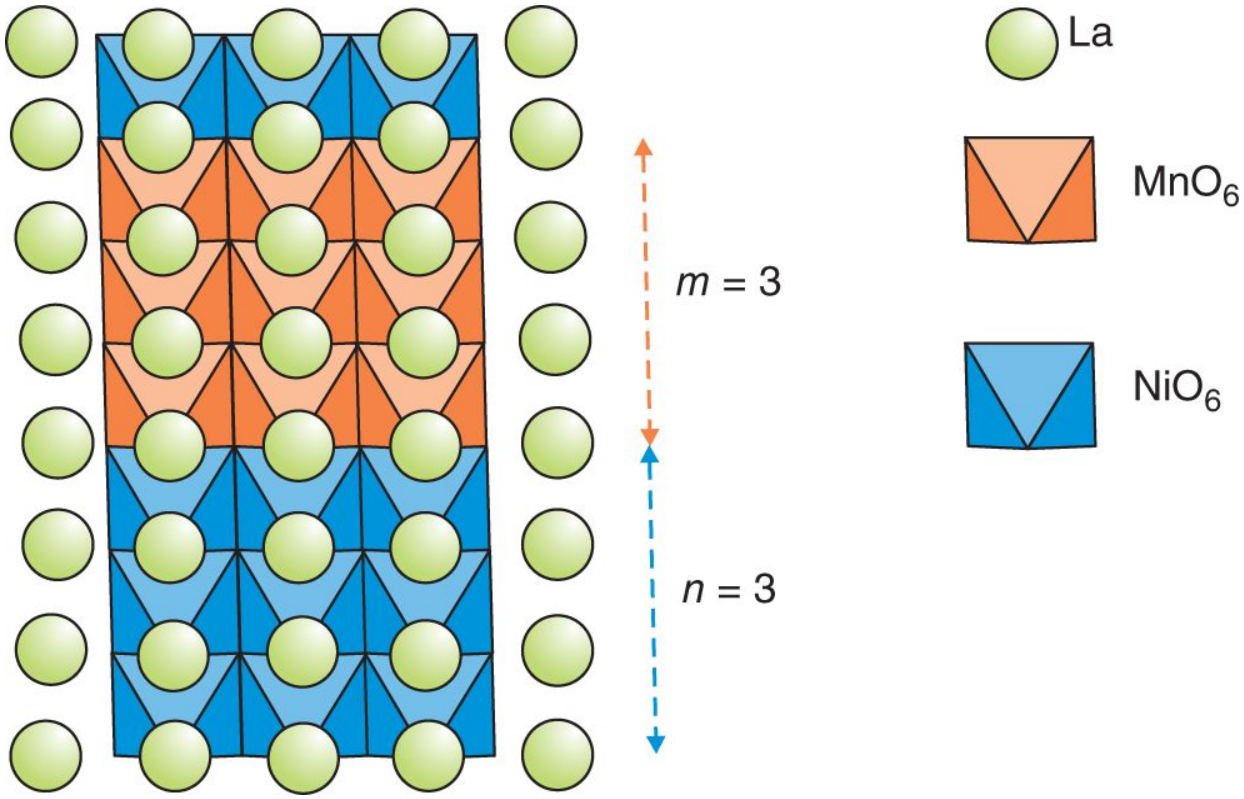


Figure 10.27 Part of an $m = n = 3$ magnetic $LaMnO_3/LaNiO_3$ superlattice in a (111) orientation.

10.6.4 Photoinduced Magnetism

A number of materials show *photoinduced magnetism* – modified magnetic behaviour when illuminated by laser light. The precise mechanism for the changes observed varies from one material to another, but it is frequently associated with the crystal field splitting of the d orbital energy levels on transition metal ions within the structure. The energy gap between the t_{2g} and e_g set of orbitals, responsible for the LS vs. HS magnetic properties of these ions is also similar to the energy of visible light. This correspondence means that the magnetic properties of an ion can be changed by irradiation of light with a suitable wavelength, which will promote electrons from the lower to the upper energy state, a phenomenon called *spin crossover* (Figure 10.28a). Molecules in which this state of affairs can occur are known as *spin-crossover* (SCO) complexes or compounds. (Spin crossover can also be induced by heat or pressure, but light induced transformations are most important.)

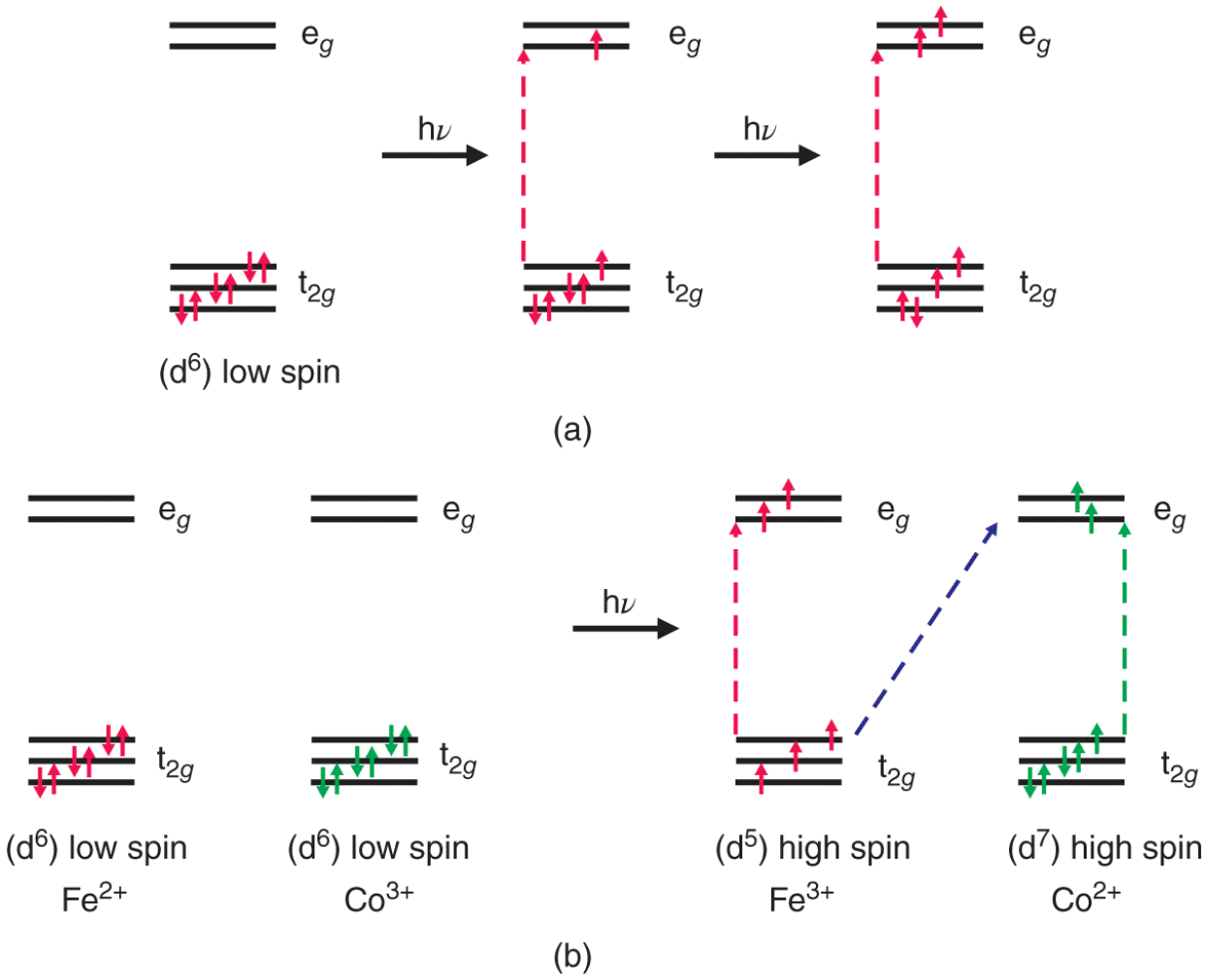


Figure 10.28 Photoinduced spin crossover in transition metal cations: (a) promotion of electrons from the t_{2g} ground state to the e_g state; (b) electron excitation and transfer in $KFeCo(CN)_6$.

In general, a light-induced transition is transitory, with photo-excited states reverting to the ground state rapidly, thus creating short-lived defects in the magnetic and optical structure of the solid. However, a stable change of magnetic structure has been achieved in a number of systems. A necessary precursor is a molecule in which the ground state is associated strongly with neither the HS nor LS state of the cation, so that the transformation is energetically reasonable. One set of compounds widely explored for this purpose is related to Prussian blue, $KFe^{2+}Fe^{3+}(CN)_6$, in which one of the Fe atoms is replaced by another transition metal ion such as Co or Cr. In Prussian blue, the Fe^{2+} and Fe^{3+} ions alternate with (CN^-) groups and the intense

colour is due to electron transfer from Fe^{2+} to Fe^{3+} via a cyanide intermediary. In molecules containing other transition metal ions, the aim is to induce this transfer optically and then to stabilise the photoinduced state. For example, in $KCoFe(CN)_6$ the non-magnetic to magnetic state transformation is:

$$Fe^{2+}(t_{2g}^6\text{LS}) - CN^- - Co^{3+}(t_{2g}^6\text{LS})$$
$$\rightarrow Fe^{3+}(t_{2g}^3 e_g^2\text{HS}) - CN^- - Co^{2+}(t_{2g}^5 e_g^2\text{HS})$$

In the initial state, all transition metal ions are in the LS state, giving an effectively non-magnetic solid. Irradiation transfers an electron from Fe to Co and also promotes electrons to create all HS ions with a considerable magnetic moment (Figure 10.28b). The magnetic complexes consist of the triple group (metal ion–cyanide–metal ion).

Many similar systems have been explored, including Mn–CN–Fe, Fe–CN–Nb, and Co–CN–W, all of which involve LS to HS electron transfer mediated by light. Although the energy gap between the HS and LS configurations is similar, they are sufficiently separated for the transition to be driven forward by one wavelength and reversed by another, making these systems switchable in either direction.

10.7 Magnetic Defects

Magnetic defects can be considered to form when magnetic ions are introduced into a non-magnetic structure, either as substituents or as interstitials, or when an low-spin state cation is transformed into an high-spin state. In this section point defects that considerably alter the magnetic properties of the host material are described.

10.7.1 Magnetic Defects in Semiconductors

Traditionally, semiconductor devices utilise the charge on electrons and ignore the spin component. However, the addition of spin to the observable properties of electrons in semiconductors has excited considerable interest for data storage and computing. The area in which both conductivity and spin are exploited simultaneously is called *spintronics*. Semiconductors that have a measurable spin or magnetic component added to them are called *diluted magnetic semiconductors*. The first materials to attempt to exploit the combination were semiconductors such as ZnSe or GaAs, in which some of the non-magnetic ions are replaced by magnetic ions such as Mn^{2+}. For example, (Ga, Mn)As has a Curie temperature of 190 K. The Mn^{2+} ion, with a $3d^5$ electron configuration, has a potential magnetic moment due to five unpaired electron spins. The magnetic ions form substitution impurity defects, in the case of Mn^{2+} doping, nominally Mn on Zn sites in ZnSe or Mn on Ga sites in GaAs. The magnetic ions can interact in the same way as they would in any other solid. Thus, at low concentrations, the materials are often paramagnetic, but higher concentrations can produce to ferromagnetic or antiferromagnetic structures.

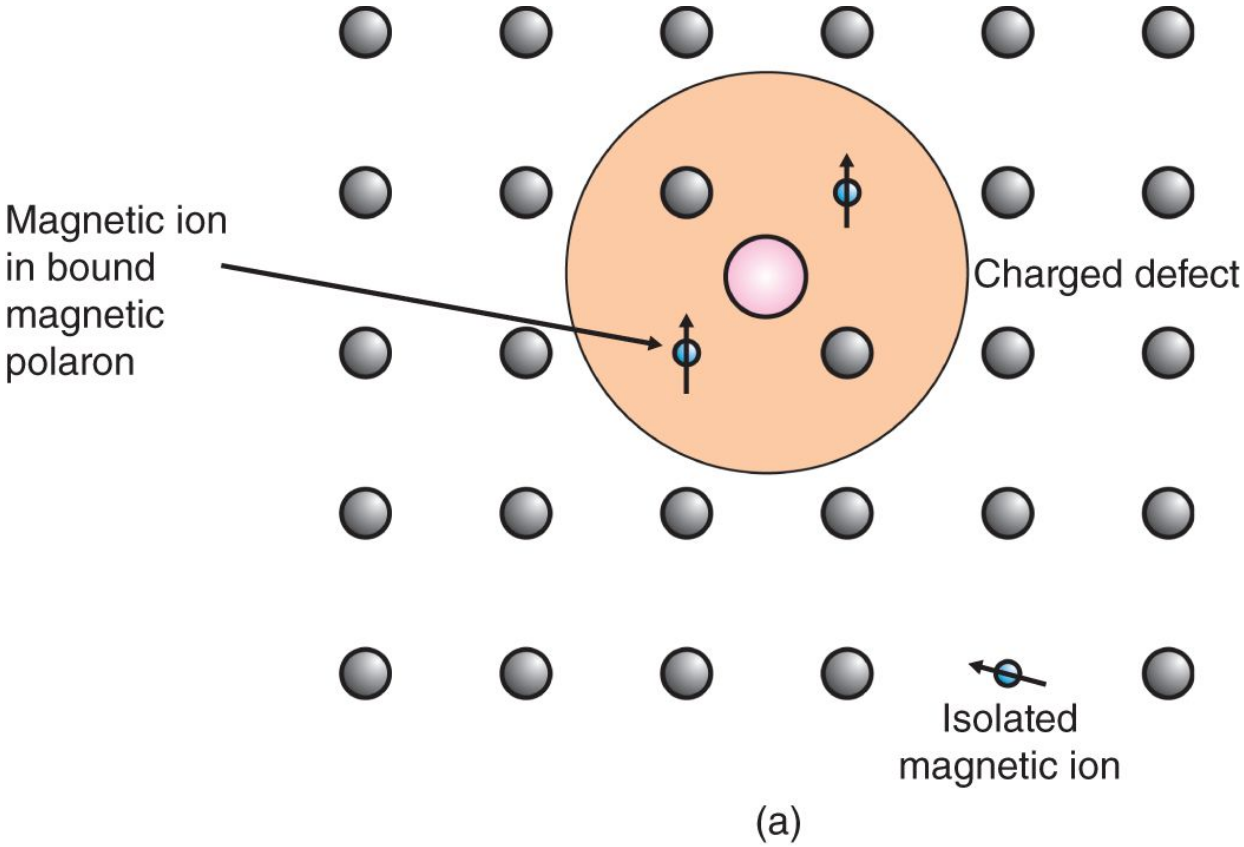
The magnetic properties of these doped semiconductors are, however, complicated and do not reside only in the magnetic impurities and there is considerable evidence to show that even non-magnetic defects may give rise to observable magnetic properties. For instance, doping of vanadium into bulk Sb_2Te_3 gives rise to a low temperature ferromagnetic phase with a Curie temperature of about 20 K, while the Curie temperature of similarly doped thin films is close to 200 K. The difference is related to the overall defect structure of the thin films compared with that of bulk samples.

Oxide films are being intensively studied as diluted magnetic semiconductors because doping with magnetic cations can produce materials that combine transparency, electronic conductivity, and room temperature ferromagnetism. The first oxide to show ferromagnetic behaviour above room temperature was ZnO doped with Mn^{4+} to form $Zn_{1-x}Mn_xO$ with x taking values below 0.02. The dopant nominally substitutes for Zn^{2+} in a bulk sample but there is evidence to suggest that many of the impurity cations favour surface rather than bulk sites, considerably altering the observed magnetic behaviour. The properties of these solids also depend upon preparation conditions. When prepared at moderate temperatures, the Mn^{4+} ions are distributed at random, giving a magnetic solid, but if the materials are heated at 700 °C or above, the Mn^{4+} ions cluster into antiferromagnetic units, and ferromagnetism is lost.

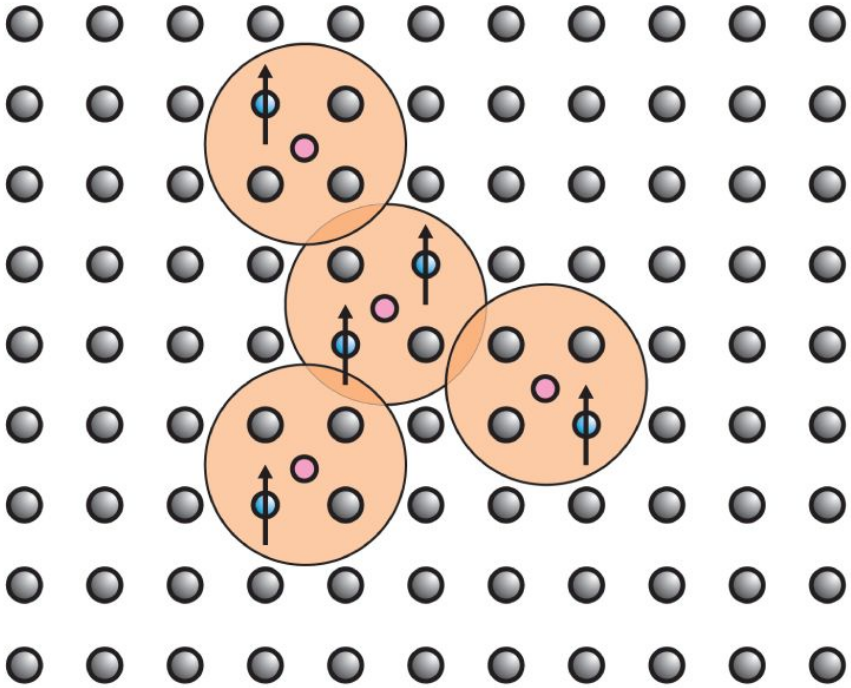
The interaction between a charged point defect and neighbouring magnetic ions in magnetically doped thin films has been described in terms of a defect cluster called a *bound magnetic polaron* (Figure 10.29a). The radius of a bound magnetic polaron due to an electron located on the defect, r, in units of Bohr radius (0.0529 nm) is given by:

$$r = \varepsilon_r \left(m_e / m_e^*\right)$$

where ε_r is the relative permittivity of the solid, m_e is the electron mass, and m_e^* the effective mass of the electron. The bound magnetic polaron has a volume of r^3. The charged defect acts to align the spins on magnetic ions that lie within the polaron volume. If there is an overlap of bound magnetic polarons ferromagnetic order can occur throughout the sample (Figure 10.29b).


Magnetic ion
in bound
magnetic
polaron
Charged defect
Isolated
magnetic ion

(a)



(b)

Figure 10.29 Bound magnetic polarons, schematic: (a) one bound magnetic polaron located on a charged defect; (b) overlapping bound magnetic polarons leading to ferromagnetic alignment of spins.

The complex role of defects in these materials is underlined by noting that many undoped oxide films, including TiO_2, HfO_2, ZnO, CeO_2, SnO_2, In_2O_3, and Al_2O_3, all show room temperature ferromagnetism *without* transition metal ion impurities. In ZnO films ferromagnetism is believed to arise with Zn interstitials. In films such as SnO_2, oxygen vacancies are implicated in the ferromagnetic properties. In particular it seems that exchange interactions between the defects and the surface give rise to the ferromagnetic ordering observed.

10.7.2 Charge and Spin States in Cobaltites and Manganites

Cobaltites and manganites, especially those related to $LaCoO_3$ and $LaMnO_3$, have been studied in detail because of their electrical and magnetic properties, all of which can be attributed to varying populations of point defects. Both adopt the perovskite ABO_3 structure, with nominal formulae $La^{3+}Co^{3+}O_3$ and $La^{3+}Mn^{3+}O_3$. The transition metal cations are located in octahedral sites in the oxides. Under high oxygen pressures both can take on more oxygen to give, $LaBO_{3+\delta}$, while under low oxygen pressures they can lose oxygen to form $LaBO_{3-\delta}$. These changes have profound effects on the magnetic properties. Doping of the compounds with different metal cations gives further modifications to the magnetic properties, so that overall, the phases show considerable magnetic complexity.

Take $LaCoO_3$ first, and suppose that the compound is ionic. In this case the material would be expected to be a non-magnetic semiconductor at ordinary temperatures. At lowest temperatures the Co^{3+} is in the LS state, with an electron configuration t_{2g}^6 (Figure 10.30a). However, the energy difference between the energy levels is small and as the temperature increases, a proportion of electrons from the ground t_{2g} state are promoted into the upper e_g

state. The configuration of these ions is $t_{2g}^5 e_g^1$, the *intermediate spin* (IS) state (Figure 10.30b), or $t_{2g}^4 e_g^2$, the HS state (Figure 10.30c). There is evidence that these spin states are ordered, especially when there are equal numbers of the two present. In the temperature range from 350 to 650 K, the materials smoothly transform from a semiconductor to a metal as the Co^{3+} ions change spin state and at high temperatures in the metallic phase the configuration is 50% HS and 50% IS.

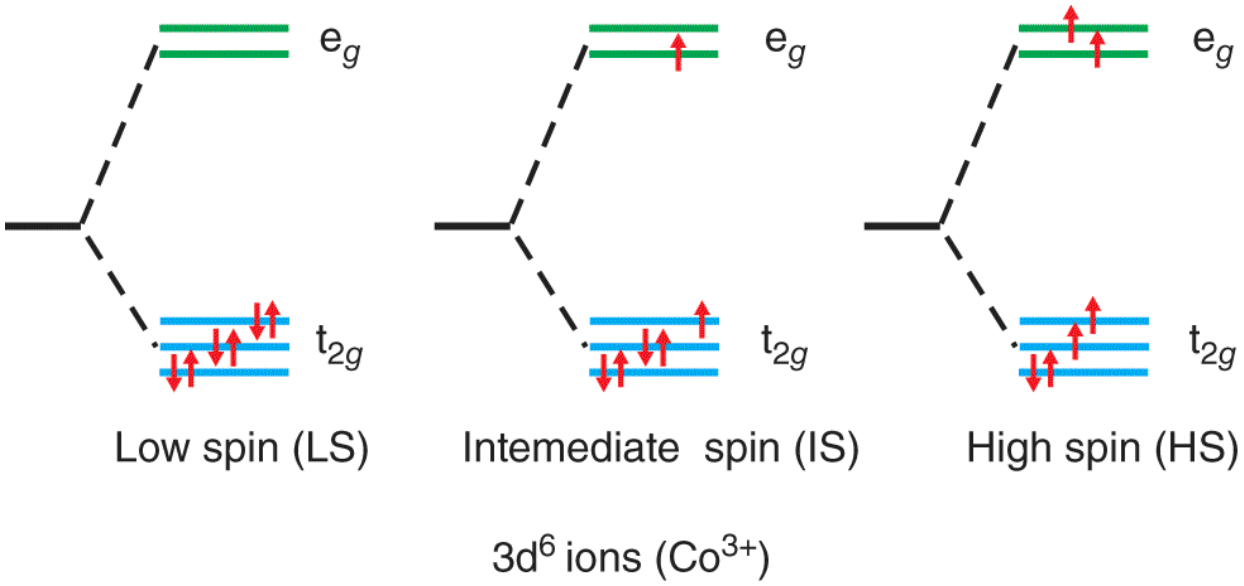


Figure 10.30 Possible electron spin configurations for Co^{3+} ($3d^6$) cations in an octahedral crystal field.

What will be the consequences of an increase in oxygen content? If this rises above 3.0 by an amount δ, then the compound will become negatively charged due to an excess of δO^{2-} anions. This can be balanced by converting some of the Co^{3+} cations into Co^{4+} cations to give an overall formula $La^{3+}Co^{3+}_{1-2\delta}Co^{4+}_{2\delta}O^{3+}_{3+\delta}$. Each additional oxygen ion is balanced by two Co^{4+} cations. In the situation when the oxygen content is reduced below 3.0 by an amount δ, the phase has an overall excess positive charge due to the absence of δO^{2-} anions. This can be balanced by converting some of the Co^{3+} cations into Co^{2+} cations to give an overall formula $La^{3+}Co^{3+}_{1-2\delta}Co^{2+}_{2\delta}O^{3+}_{3+\delta}$. Each

missing oxygen ion is balanced by two Co^{2+} cations. In oxygen-rich or oxygen-deficient materials, the various spin states give rise to a rich variety of magnetic behaviour. (In semiconductor terminology, Chapter 11, this is called acceptor doping or hole doping).

The variation in spin components induced by changing the oxygen stoichiometry can be mimicked by doping. The addition of extra oxygen to $LaCoO_3$ can be replaced by the incorporation of an alkaline earth cation, Ca^{2+}, Sr^{2+}, Ba^{2+}, in place of La, to form $La_{1-x}A_xCoO_3$. The result is that each A^{2+} cation must be balanced by changing one Co^{3+} cation to Co^{4+}. In this state the Co^{3+} ions are thought to be mainly in the IS state and Co^{4+} in the HS state, with an electron configuration for this $3d^5$ ion of $t_{2g}^3 e_g^2$. In these and other doped cobaltites, small changes in oxygen content dramatically alter the magnetic structure and properties of the phase.

Stoichiometric $LaMnO_3$ contains only octahedrally coordinated Mn^{3+} in an HS state. It is an antiferromagnetic insulator; due to superexchange via the O^{2-} ions that separate the cations. The similar phase $CaMnO_3$ contains only octahedrally coordinated Mn^{4+} in an HS state. It is also an antiferromagnetic insulator, again, believed to be due to superexchange via the O^{2-} ions that separate the cations. The system $La_{1-x}Ca_xMnO_3$ contains both Mn^{3+} and Mn^{4+}, and encompasses a complex collection of magnetic phases. In barest outline: with x between 0 and ~0.21 antiferromagnetic insulating phases occur; with x between ~0.21 and ~0.5 a metallic ferromagnetic metal occurs, and with x between ~0.5 and 1.0 another succession of antiferromagnetic phases. The double exchange mechanism was first invoked by Zener in 1951 to explain the occurrence of the metallic ferroelectric phase. In this case the participating ions, Mn^{3+} and Mn^{4+}, would give rise to a ferromagnetic phase by double exchange provided that the cations were well separated and mobile electrons were present (Figure 10.31).

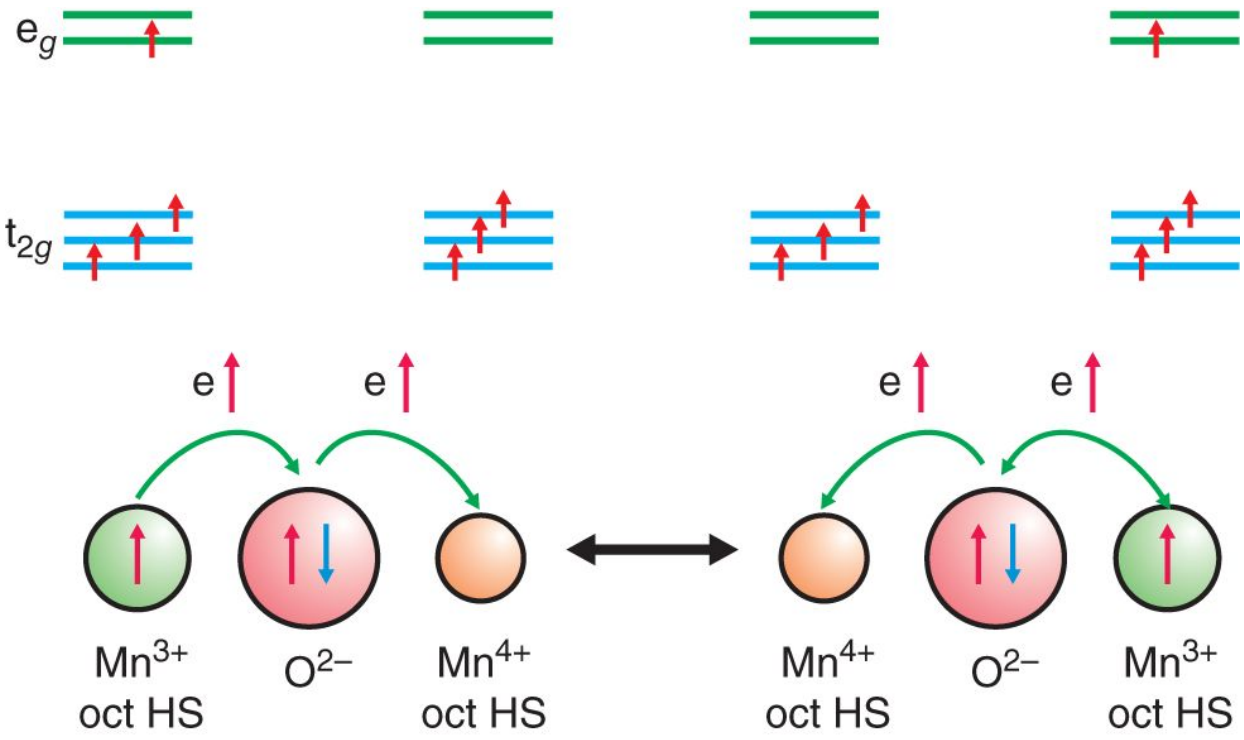


Figure 10.31 Double exchange between Mn^{3+} and Mn^{4+} high spin (HS) cations in ferromagnetic $La_{1-x}Ca_xMnO_3$, schematic.

Interest in $LaMnO_3$ intensified when it was discovered that doped forms show *colossal magnetoresistance* (CMR). The magnetoresistance of a solid is the change of resistance when a magnetic field is applied. The magnetoresistive (MR) ratio is defined as the change in resistance when a magnetic field is applied to that in zero field:

$$\text{MR ratio} = (R_H - R_0)/R_0 = \Delta R/R_0$$

where R_H is the resistance in a magnetic field and R_0 is the resistance in the absence of the field. For the doped manganites the MR ratio is close to 100% and is negative, meaning that the resistance falls to almost nothing in a magnetic field. In such materials, a better measure of the change is the ratio of R_0/R_H, where values of up to 10^{11} have been recorded.

The CMR effect is found in a variety of doped systems, including $La_{1-x}A_xMnO_3$, with A being an alkaline earth cation (Ca^{2+}, Sr^{2+},

Ba^{2+}) and ordered $Ln_{0.5}A_{0.5}MnO_3$, with Ln being a lanthanoid, typically Pr^{3+}. In all of these compounds, each substituted A^{2+} cation transforms one Mn^{3+} cation into Mn^{4+} cation. In addition, disproportionation of $2Mn^{3+}$ into Mn^{2+} and Mn^{4+} may also occur, so that several spin states are found in the matrix. It is believed that the Mn^{3+} ions are in the HS $t_{2g}^3e_g^1$ state as are the Mn^{4+} ions, with an electron configuration $t_{2g}^3e_g^0$. The Mn^{2+} ions can adopt either an LS $t_{2g}^5e_g^0$ configuration, an IS $t_{2g}^4e_g^1$ configuration, or an HS $t_{2g}^3e_g^2$ configuration (Figure 10.32). The distribution between these states is temperature sensitive, and at room temperature the IS state appears to dominate. Substitution of some of the Mn with both magnetic or non-magnetic ions, including Ti^{4+}, Sn^{4+}, Fe^{3+}, Cr^{3+}, Al^{3+}, Ga^{3+}, In^{3+}, Mg^{2+}, and Ni^{2+}, increases the CMR effect considerably. For $Pr_{0.7}Ca_{0.2}Sr_{0.1}MnO_3$ the ratio R_0/R_H is approximately 230. This jumps to 4×10^5 for the modestly doped $Pr_{0.7}Ca_{0.2}Sr_{0.1}Mn_{0.98}Mg_{0.02}O_3$.

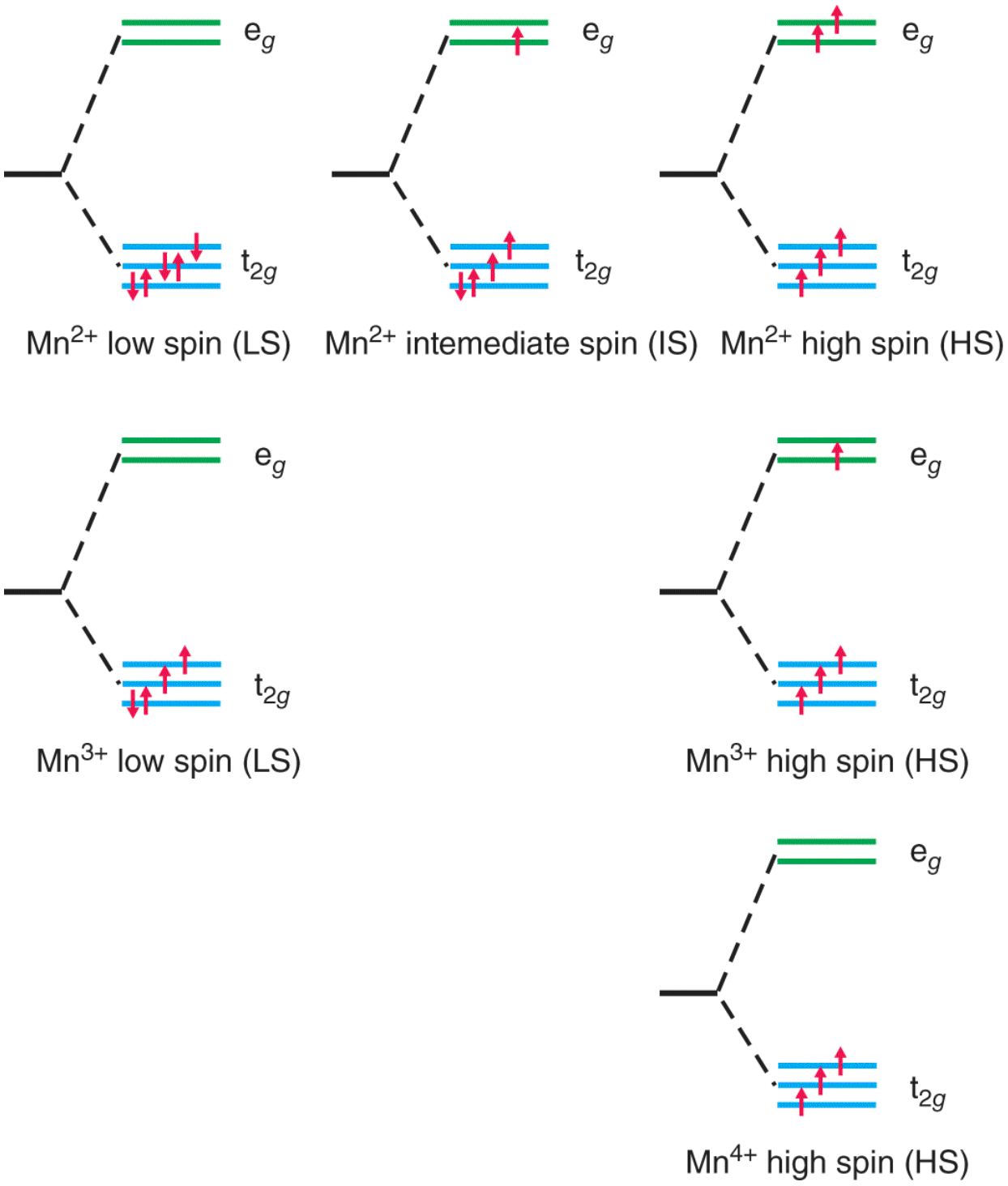


Figure 10.32 Electron configurations possible for Mn cations in an octahedral crystal field.

Neither the magnetic defect structures of these phases nor the mechanism of the CMR effect are yet completely clear, although

Mn^{3+}–Mn^{4+} interactions are believed to be important in inducing CMR and double exchange between these ions is believed to occur in the metallic ferromagnetic manganites.

Further Reading

General:

Epstein, A.J. (2003). Organic-based magnets: opportunities in photoinduced magnetism, spintronics, fractal magnetism and beyond. *Mater. Res. Soc. Bull.***28**: 492.

Spaldin, N. (2003). *Magnetic Materials*. Cambridge: Cambridge University Press.

Magnetic States:

Chen, W., Sun, Z., Wang, Z. et al. (2019). Direct observation of van der Waals stacking-dependent interlayer magnetism. *Science***366**: 983.

Drokina, T.V., Petrakovskii, G.A., Molokeev, M.S. et al. (2015). Spin-glass magnetism in $RFeTi_2O_7$ (R □ Lu, Tb) compounds. *Phys. Proc.***75**: 580.

Judy, J.H. (1990). Thin film recording media. *Mater. Res. Soc. Bull.***XV**: p63.

Kamber, U., Bergman, A., Eich, A. et al. (2020). Self-induced spin glass state in elemental and crystalline neodymium. *Science***368**: 966.

Khan, G.G., Ghosh, S., and Mendel, K. (2012). Origin of d^0 ferromagnetism and characteristic photoluminescence in pristine SnO_2 nanowires: a correlation. *J. Solid State Chem.***186**: 278–282.

Sharrock, M.P. (1990). Particulate recording media. *Mater. Res. Soc. Bull.***XV**: p53.

Yan, H., Feng, Z., Qin, P. et al. (2020). Electric-field-controlled antiferromagnetic spintronic devices. *Adv. Mater.***32**: 1905603.

A Starting Point for the Detection of Magnetic Fields by Animals:

Edera, S.H.K., Cadioub, H., Muhamadb, A. et al. (2012). Magnetic characterisation of isolated candidate vertebrate magnetoreceptor cell. *Proc. Natl. Acad. Sci. U.S.A.***109**: 12022–12027. This paper details magnetic receptor cells in trout.

Winklhofer, M. (2012). An avian magnetometer. *Science***336**: 991–992.

Density Functional Theory Calculations of Magnetic Properties is Outlined by:

Matar, S.F. (2003). *Ab initio* investigations in magnetic oxides. *Prog. Solid State Chem.***31**: 239–299.

Magnetic Superlattices:

Gilbert, M. et al. (2012). Exchange-bias in $LaNiO_3$-$LaMnO_3$ superlattices. *Nat. Mater.***11**: 195–198.

Hwang, H.Y., Iwasa, Y., Kawasaki, M. et al. (2012). Emergent phenomena at oxide interfaces. *Nat. Mater.***11**: 103–113.

A Starting Point for Studies on Photomagnetism:

Dong, D.-P., Liu, T., Kanegawa, S. et al. (2012). Photoswitchable dynamic magnetic relaxation in well-isolated $\{Fe_2Co\}$ double-zigzag chains. *Angew. Chem. Int. Ed.***51**: 5119–5123.

Ohkoshi, S.-i., Imoto, K., Tsunobuchi, Y. et al. (2011). Light induced spin-crossover magnet. *Nat. Chem.***3**: 564–569.

Ozaki, N. et al. (2011). Photoinduced magnetization with a high Curie temperature and large coercive field in a Co-W bimetallic assembly. *Adv. Funct. Mater.***22**: 2089–2093.

Problems and Exercises

Quick Quiz

10.1 A material which is slightly attracted to a magnetic field is:

a. Diamagnetic.

b. Paramagnetic.

c. Ferromagnetic.

10.2 A material that does not contain any magnetic dipoles on atoms or ions is:

a. A diamagnetic material.

b. A paramagnetic material.

c. An antiferromagnetic material.

10.3 The Curie law describes the magnetic behaviour of:

a. Paramagnetic solids.

b. Diamagnetic solids.

c. Ferrimagnetic solids.

10.4 Above the Curie temperature, ferromagnetic materials become:

a. Diamagnetic.

b. Paramagnetic.

c. Antiferromagnetic.

10.5 The magnetic behaviour of ferromagnetic solids is described by the Curie–Weiss law:

a. Above the Curie temperature.

b. Below the Curie temperature.

c. At low temperatures.

10.6 The Néel temperature is the temperature at which an antiferromagnetic material becomes:

a. Ferromagnetic.

b. Ferrimagnetic.

c. Paramagnetic.

10.7 A ferrimagnetic compound contains:

a. All magnetic dipoles arranged parallel to one another.

b. Two sets of magnetic dipoles arranged antiparallel to each other.

c. One set of magnetic dipoles arranged in an antiparallel arrangement.

10.8 The magnetic properties of electrons are due to:

a. The electron spin only.

b. The orbital motion only.

c. Both the orbital motion and spin.

10.9 The orbital contribution to the magnetic moment is quenched for:

a. Lanthanoid atoms and ions.

b. 3d transition metal atoms and ions.

c. Both lanthanoids and transition metals.

10.10 The interaction of d orbitals and oxygen atoms that leads to antiferromagnetic ordering is called:

a. Exchange interaction.

b. Superexchange.

c. Double exchange.

10.11 The interaction between magnetic atoms in a ferrimagnetic material is called:

a. Double exchange.

b. Superexchange.

c. Exchange interaction.

10.12 Ferrites with the spinel structure are:

a. Paramagnetic materials.
b. Ferromagnetic materials.
c. Ferrimagnetic materials.

10.13 The general formula of hexagonal ferrites is:

a. $A^{2+}Fe_{12}O_{19}$.
b. $A^{4+}Fe_{12}O_{19}$.
c. $A^{3+}Fe_{12}O_{19}$.

10.14 Hexagonal ferrites are:

a. Paramagnetic materials.
b. Ferromagnetic materials.
c. Ferrimagnetic materials.

10.15 A Bloch wall in a ferromagnetic solid is:

a. The external surface of the ferromagnet.
b. The region between two domains.
c. A grain boundary.

10.16 The walls between magnetic domains are the result of:

a. Competition between long-range interactions and thermal energy.
b. Competition between short-range interactions and chemical bonding.
c. Competition between long-range and short-range interactions.

10.17 Pauli paramagnetism refers to:

a. The paramagnetic behaviour of ions.
b. The paramagnetic behaviour of ferromagnetic metals above the Curie temperature.
c. The paramagnetic behaviour of non-ferromagnetic metals.

10.18 The ferromagnetic properties of the 3d transition metals are explained by:

a. The overlap of 3s and 3d orbitals.

b. The overlap of 3p and 3d orbitals.

c. The overlap of 3d and 4s orbitals.

Calculations and Questions

10.1 Determine the ground state values of *S*, *J*, *L*, and **m** for the f^3 ion Nd^{3+}.

10.2 Determine the ground state values of *S*, *J*, *L*, and **m** for the f^7 ion Gd^{3+}.

10.3 Determine the ground state values of *S*, *J*, *L*, and **m** for the d^4 ion Mn^{3+} in both high and low spin states.

10.4 Determine the ground state values of *S*, *J*, *L*, and **m** for the d^7 ion Co^{2+} in both high and low spin states.

10.5 Why don't the lanthanoid ions possess high spin and low spin states?

10.6 Estimate the saturation magnetisation, $\mathbf{M}_S$, for a sample of ferromagnetic nickel metal,

a. assuming only the unpaired d electrons contribute to the magnetism and the spins can be added as if the material were paramagnetic,

b. using the measured magnetic moment per nickel atom of 0.58 μ_B.

The metal has a face-centred cubic structure with a cubic unit cell parameter of 0.3524 nm.

10.7 Iron has a saturation magnetisation of 1.72×10^6 A m^{-1}. What is the measured magnetic moment, in Bohr magnetons, of an iron atom? Iron has the body-centred cubic structure, with a cubic unit cell parameter of 0.2867 nm.

10.8 The saturation magnetisation of cobalt is 1.446×10^6 A m^{-1}.

a. Calculate the number of magnetic dipoles per unit volume in cobalt knowing the effective magnetic moment per atom is 1.72 μ_B.

b. If it is assumed that there is one atom per primitive cubic unit cell, determine the length of the unit cell edge.

10.9 The magnetic moment of Fe^{3+} ions in the species $[Fe(H_2O)_6]^{3+}$ is 5.3 μ_B. What is the likely electron configuration of the Fe^{3+} ion?

10.10 The magnetic moment of Fe^{3+} ions in the species $[Fe(CN)_6]^{3-}$ is 2.3 μ_B. What is the likely electron configuration of the Fe^{3+} ion? What can you conclude about the orbital contribution to the magnetic moment for this species?

10.11 The species $[Co(NH_3)_6]^{3+}$, containing Co^{3+} ions, gives rise to diamagnetic solids while the species $[CoF_6]^{3-}$, also containing Co^{3+} ions, has a strong magnetic moment and gives rise to paramagnetic solids.

a. What is the likely electron configuration of the Co^{3+} ions in these two species?

b. Calculate the expected magnetic moment of Co^{3+} in $[CoF_6]^{3-}$.

10.12 The species $[Fe(CN)_6]^{4-}$, containing Fe^{2+} ions, gives rise to diamagnetic solids while the species $[Fe(NH_3)_6]^{2+}$, also containing Fe^{2+} ions, has a strong magnetic moment and gives rise to paramagnetic solids.

a. What is the likely electron configuration of the Fe^{2+} ions in these two species?

b. Calculate the expected magnetic moment of Fe^{2+} in $[Fe(NH_3)_6]^{2+}$.

10.13 The crystal field splitting of diamagnetic $[Co(NH_3)_6]^{3+}$, containing Co^{3+} ions is 4.57×10^{-19} J. What wavelength light would produce photoinduced paramagnetism in this molecule?

10.14 The crystal field splitting of diamagnetic $[Fe(CN)_6]^{4-}$, containing Fe^{2+} ions, is 6.56×10^{-19} J. What wavelength light would produce photoinduced paramagnetism in this molecule?

10.15 The light absorbed by the complex ion $[FeF_6]^{3-}$ peaks at 719 nm. This absorption is due to the promotion of an electron from the lower (t_{2g}), to the upper (e_g), state in the Fe^{3+} ion.

a. Calculate the magnitude of the crystal field splitting of the Fe^{3+} d-orbitals due to F^-.

b. What is the relative population of the two levels at 300 K?

10.16 Calculate the paramagnetic energy level splitting for a Pr^{3+} ion in a magnetic flux density of 0.5 T, and the corresponding wavelength of radiation for a transition between these energy levels.

10.17 Calculate the paramagnetic energy level splitting for a Ho^{3+} ion in a magnetic flux density of 0.75 T, and the corresponding wavelength of radiation for a transition between these energy levels.

10.18 Calculate the paramagnetic energy level splitting for a V^{4+} ion in a magnetic flux density of 0.25 T if the orbital contribution is quenched, and the corresponding wavelength of radiation for a transition between these energy levels.

10.19 Calculate:

a. the paramagnetic energy level splitting for a Ni^{2+} ion in a magnetic flux density of 0.6 T if the orbital contribution is quenched;

b. and the corresponding wavelength of radiation for a transition between these energy levels.

10.20 Calculate the mass susceptibility of $NiSO_4 \cdot 7H_2O$ at 20 °C (units kg^{-1}), which contains isolated Ni^{2+} ions. The density of the compound is 1980 kg m^{-3}. Assume that the spin only formula is adequate.

10.21 Calculate the mass susceptibility of $CuSO_4 \cdot 5H_2O$ at 20 °C (units kg^{-1}), which contains isolated Cu^{2+} ions. The density of the compound is 2284 kg m^{-3}. Assume that the spin only formula is adequate.

10.22 Calculate the volume susceptibility of $MnSO_4 \cdot 4H_2O$ at 20 °C (units m^{-3}), which contains isolated Mn^{2+} ions. The density of the compound is 1980 kg m^{-3}. Assume that the spin only formula is adequate.

10.23

a. Calculate the value of x in the Brillouin function $B_J(x)$, where $x = g_J \mu_B JB/k_B T$ for an Mn^{2+} ion with five unpaired electrons in an inductance of 0.5 T. Assume that the orbital angular momentum is quenched, so that $L = 0$. Take S as 5/2 (Table 10.2).

a. The Curie law requires that $x \ll 1$. Estimate the temperature at which x is 0.01.

10.24

a. Calculate the value of x in the Brillouin function $B_J(x)$, where $x = g_J \mu_B JB/k_B T$ for a Ti^{3+} ion with one unpaired electron in an inductance of 0.45 T. Assume that the orbital angular momentum is quenched, so that $L = 0$, and take S as ½ (Table 10.2).

a. Repeat the calculation assuming that the orbital angular momentum is not quenched and $L = 2$, $J = 3/2$.

a. The Curie law requires that $x \ll 1$. Estimate the temperature at which x is 0.005.

10.25 Estimate the saturation magnetisation and the magnetic inductance for the cubic ferrite $CoFe_2O_4$ with the inverse spinel structure. Co^{2+} is a d^7 ion. The cubic unit cell has a lattice parameter of 0.8443 nm and contains eight formula units. Assume that the orbital angular momentum is quenched.

10.26 Estimate the saturation magnetisation and the magnetic inductance for the cubic ferrite $NiFe_2O_4$ with the inverse spinel structure. Ni^{2+} is a d^8 ion. The cubic unit cell has a lattice parameter of 0.8337 nm and contains eight formula units. Assume that the orbital angular momentum is quenched.

10.27 Estimate the saturation magnetisation and the magnetic inductance for the hexagonal ferrite ferroxdure, $BaFe_{12}O_{19}$. The hexagonal unit cell has parameters a = 0.587 78 nm, c = 2.1236. There are two formula units in the unit cell. Assume that the orbital angular momentum is quenched.

10.28 Estimate the saturation magnetisation and the magnetic inductance for the hexagonal ferrite $SrFe_{12}O_{19}$. The hexagonal unit cell has parameters a = 0.588 36 nm, c = 2.303 76 nm. There are two formula units in the unit cell. Assume that the orbital angular momentum is quenched.

10.29 Derive a formula for the saturation magnetisation of a cubic ferrite $A^{2+}Fe_2O_4$, which is partly inverse. The fraction of Fe^{3+} ions on tetrahedral sites is given by λ, where $0 < \lambda < 0.5$ (Section 3.2.2.8).

Note

1 Ferromagnetic and ferrimagnetic materials (Section 10.5) behave in the same way with respect to the subject matter of this section. To avoid repetition only the term ferromagnetic will be used, but it should be understood that ferrimagnetic materials are identical.

11
Electronic Conductivity in Solids

- What is n-type and p-type silicon?
- How are conducting polymers produced?
- What is a cuprate superconductor?

Solids that allow an electric current to flow when a small voltage is applied are called *conductors* or *semiconductors*. Conductivity requires the presence of mobile charge carriers. In this chapter, solids that have reasonable numbers of mobile charge carriers present, either because of their native electronic properties, or because they have been deliberately introduced by doping, are considered. In addition, superconductors, a group of materials that appear not to use 'normal' conductivity mechanisms, are described.

11.1 Metals

11.1.1 Metals, Semiconductors, and Insulators

One of the defining physical properties of a metal is its electrical conductivity, defined via Ohm's Law,

$$V = I\,R$$

where V is the voltage applied across the material, I is the resultant current, and R is the *resistance*. The resistance is proportional to the length of the material, L, and the cross-sectional area, A:

$$R = \rho(L/A)$$

where ρ is the *resistivity*. The resistivity of a solid is an intrinsic property, whereas resistance depends upon the dimensions of the sample. The *conductivity* of a solid is the inverse of the resistivity:

$$\sigma = 1/\rho$$

Electrical conductivity in a metal is due to electrons that are free to move, i.e. gain energy, under the influence of an applied voltage. Metallic bonding allows conductivity to be understood most easily. In this model, the electrons on the atoms making up the solid are allocated to energy bands that run throughout the whole of the solid. A simple one-dimensional band-structure diagram, called a *flat band* diagram, allows the broad distinction between conductors, semiconductors, and insulators to be understood.

If the number of electrons available fills an energy band completely and the energy gap between the top of the filled band and the bottom of the next higher (empty) energy band is large, the material is an *insulator* (Figure 11.1a). This is because the electrons have no means of taking up the additional energy needed to allow them to move under a low voltage, because all of the energy levels are filled or inaccessible. Only a considerable voltage will cause electrons to jump from the completely filled band to the next highest empty band, and when such a transfer does occur, the insulator has *broken down*.

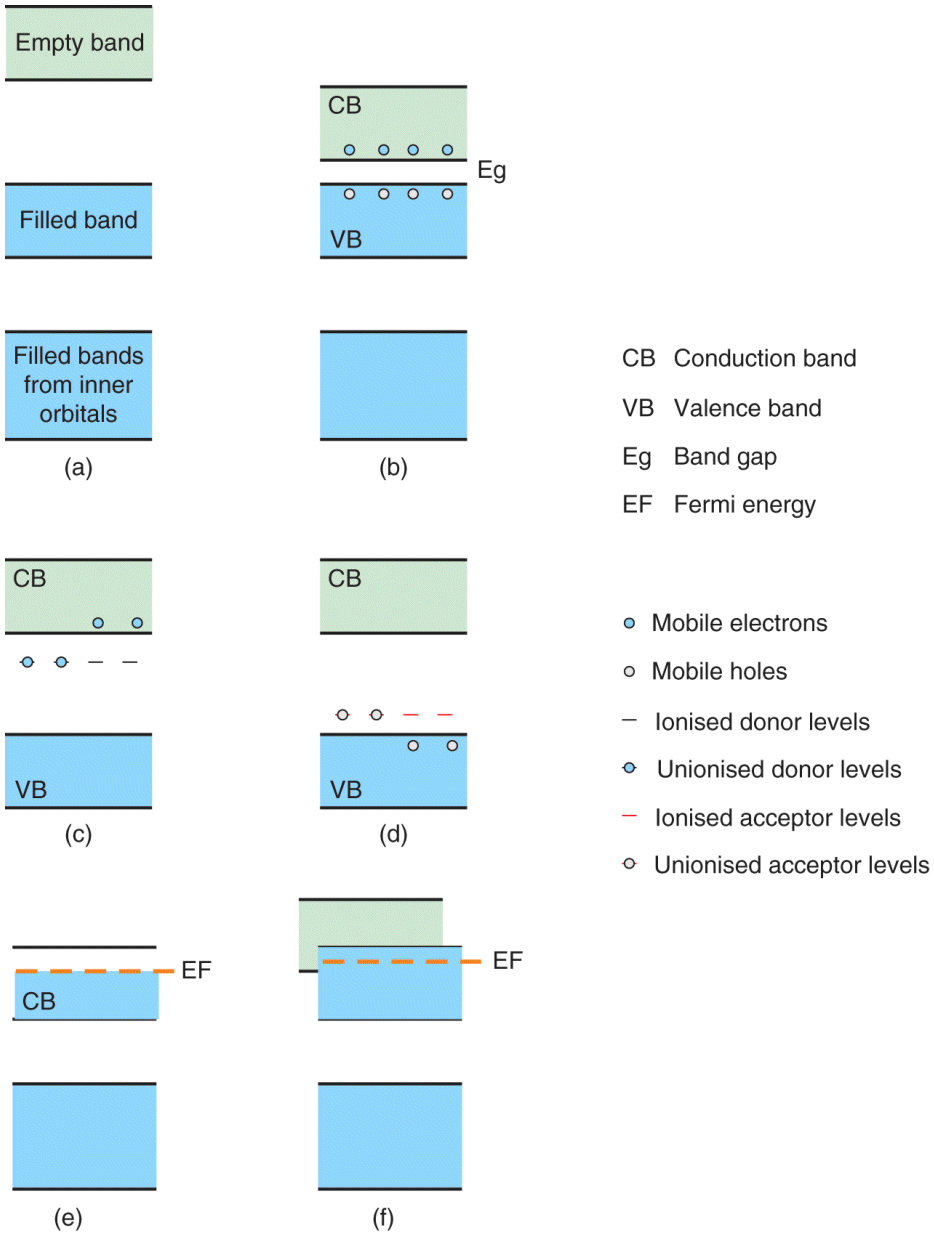


Figure 11.1 Schematic energy band representations of materials: (a) insulators; (b) intrinsic semiconductors; (c) n-type extrinsic semiconductors; (d) p-type extrinsic semiconductors; (e) metals; (f) semimetals.

If the energy gap between the filled and empty band is small enough that thermal energy is sufficient to cause some electrons to jump from the lower filled band to the upper empty band, a degree of the electronic conduction becomes possible. Such materials are called *intrinsic semiconductors* (Figure 11.1b). Once electrons arrive in the empty band, they can contribute to electrical conductivity, as there are empty energy levels around them and the solid is transformed from an insulator into a poor electronic conductor. The now almost filled lower energy band is called the *valence band* (VB), and the higher energy almost empty energy band is called the *conduction band* (CB). The energy gap is called the *band gap*, E_g.

Although this picture is simple, it reveals an important feature. It is found that the 'vacancies' left in the valence band when electrons are promoted to the conduction band also contribute

to the conduction process. To a good approximation these vacancies can be equated to positive electrons, and move in the opposite direction to the electrons in an applied field. They are called *positive holes* or just *holes*. Semiconductors are characterised by an increase in the conductivity with temperature because the number of mobile charge carriers, electrons and holes, will increase as the temperature increases. If the band gap is so small that thermal energy at normal temperatures is sufficient to generate a very high number of charge carriers in each band the material is classed as a *degenerate semiconductor*. At 0 K intrinsic semiconductors become insulators.

Semiconductivity can arise in an insulator if the material contains an appreciable number of impurities (added intentionally or not). Similarly, the conductivity of an intrinsic semiconductor can be manipulated by adding impurities. These can act as *donors*, liberating electrons to the conduction band, to form n-type semiconductors (Figure 11.1c), or as *acceptors*, receiving electrons from the valence band, and thus contributing holes to the valence band, to form p-type semiconductors (Figure 11.1d). Such materials are called *extrinsic semiconductors*. When all of the impurities are fully ionised (i.e. all the donor levels have lost electrons or all the acceptor levels have gained electrons), the *exhaustion range* has been reached. If the donors and acceptors are present in equal numbers, the material is said to be a *compensated semiconductor*. At 0 K these materials are also insulators. It is difficult in practice to distinguish between compensated extrinsic semiconductors and intrinsic semiconductors.

When there are insufficient electrons to fill the highest (conduction) band, even small amounts of energy will be able to move the topmost electrons into higher energy levels and small voltages will produce significant conductivity and the solid is a *metal* (Figure 11.1e). The uppermost filled energy levels form the *Fermi surface* at the *Fermi energy*. Should the bottom of the $(n + 1)$th band lie energetically lower than the top of a full nth band, electrons will spill over into the bottom of the empty band until the Fermi level intersects both sets of bands. Holes and electrons now exist, and coexist at 0 K. This type of material is called a *semimetal* (Figure 11.1f).

From the point of view of metallic conductivity, the nature of the atoms composing the structure is not important. The primary point is that there should be a conduction band that is party filled with electrons. In these terms a large number of compounds can be classed as metals or semimetals, including oxides, nitrides, phosphides, sulphides, selenides, and tellurides.

11.1.2 Electronic Conductivity

Although the detailed conductivity of a metal is dependent upon the shape of the Fermi surface, a good idea of conduction can be gained by considering the properties of a simple one-dimensional metal. The energy of a free electron moving through a solid is related to its wave vector **k** by a parabolic curve (Figure 11.2a). The wave vector can also be expressed in terms of the momentum vector or the velocity vector of the electron:

$$\mathbf{k} = 2\pi/\lambda = (2\pi/h)\mathbf{p} = (2\pi m/h)\mathbf{v}$$

where λ is the wavelength of the electron wave, **p** is the momentum of the electron, m is the mass of the electron, **v** is the velocity of the electron, and h is Planck's constant. Thus, **k** is proportional to both momentum and velocity. Each energy level is then associated with an electron velocity, which increases as the Fermi surface is approached. When the metal is at normal equilibrium, the velocities sum to zero because all velocity states (equivalent to energy levels) are filled up to the Fermi surface. Thus, although the electrons are in motion, no current flows (Figure 11.2b). When an electric field, **E** (also a vector), is applied to the metal,

each electron experiences a force $-e\mathbf{E}$ and a change in momentum, $\Delta\mathbf{p}$, which is equivalent to a change in the value of the wave vector, $\Delta\mathbf{k}$, and velocity, $\Delta\mathbf{v}$, of the electrons. An electric field applied along the **x**-axis will translate the velocity distribution parallel to the **x**-axis by a small amount opposite to the direction of the applied field. The velocities no longer sum to zero along **x**, and electrons drift in a direction opposite to that of the applied field, causing a current to flow (Figure 11.2c). The effect is limited by collisions with atoms, impurities, and defects and ultimately a steady current is reached.

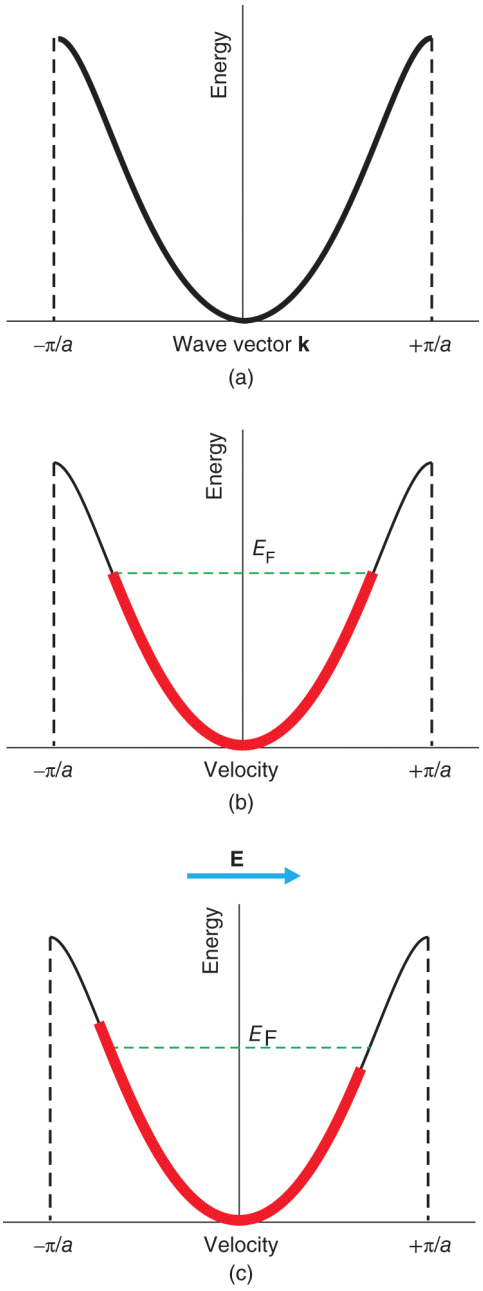


Figure 11.2 Free electrons in a one-dimensional metal: (a) energy vs. wave vector at equilibrium; (b) energy vs. electron velocity at equilibrium; (c) energy vs. electron velocity in an electric field **E**.

In the case of a two-dimensional metal, the Fermi surface can be represented as a circle, with velocity components along the **x**- and **y**-axes (Figure 11.3a). The circular periphery marks the boundary up to which the velocity states are occupied. When the metal is at normal equilibrium, the velocities sum to zero because all velocity states are filled up to the Fermi surface, and although the electrons are in motion, no current flows. As in the one-dimensional case, an electric field applied along the **x**-axis will translate the velocity distribution parallel to the **x**-axis by a small amount opposite to the direction of the applied field (Figure 11.3b). The velocities no longer sum to zero along *x*, and electrons drift in a direction opposite to that of the applied field giving rise to a current along –**x**.

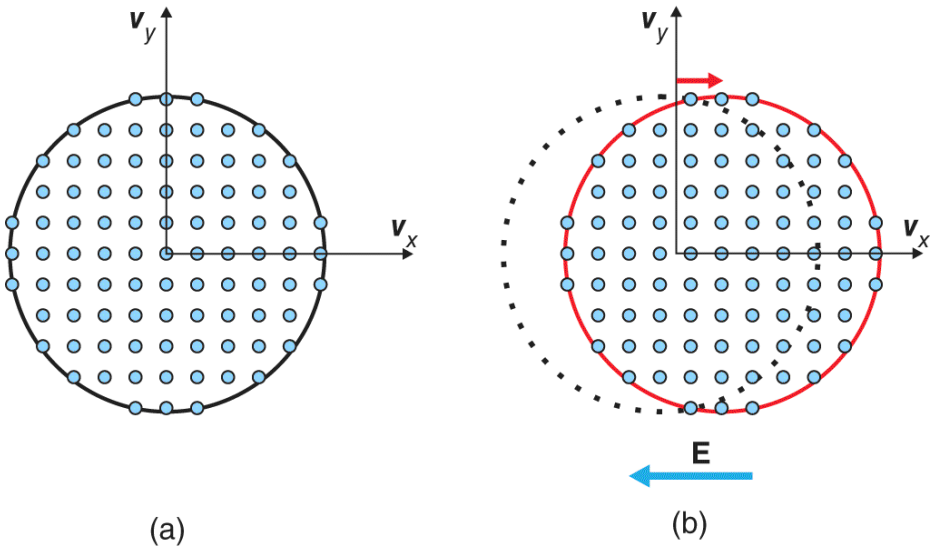


Figure 11.3 The circular Fermi surface of a two-dimensional metal, dots represent velocity states: (a) in the absence of an electric field; (b) in an electric field **E**.

The relationship between the current and applied voltage in a metallic conductor can be most easily described in terms of a one-dimensional model, in which the variables such as the electric field, wave vector, and velocity are treated as scalar quantities, E, k, v, rather than vectors. The magnitude of the current, I, that results from electron drift is then defined as the amount of charge that passes through unit cross-sectional area of the conductor per second:

$$I = n e A v$$

where n is the number of mobile electrons per unit volume, A is the cross-sectional area of the conductor, and v is the (scalar) *drift velocity*. The force exerted on an electron by an electric field, E, is given by $-eE$ and the acceleration, a, imposed on an electron is then given by:

$$a = e\ E/m_e^* \qquad (11.1)$$

where m^*_e is the effective mass of the electron. (This latter term is used to account for the fact that the dynamics of electrons in solids is quite different from that in a vacuum, and measurements show that the mass that applies to electrons in a vacuum, the rest mass, needs to be replaced by an effective mass. The effective mass is not a constant, but depends upon temperature and the direction in the crystal that the electron is travelling. It is invariably expressed as a fraction of the electron rest mass m_e, i.e. 0.067 m_e.)

The current in a conductor is steady, and not ever increasing, as would be expected if the electrons were accelerating. To account for this, it is assumed that the electrons constantly

collide with the atoms and defects in the material and that each collision resets the drift velocity to zero. In this case the current will decay to zero in a time τ, the *relaxation time*, after the voltage is turned off, where τ is equivalent to the time between successive collisions. The *mean free path* of the electron, which is the length of the path between successive collisions, is given by:

$$\Lambda = \tau \, v_F$$

where v_F is the electron velocity at the Fermi surface. In addition, the drift velocity of the electrons is:

$$v = a \, \tau$$

Substituting for the acceleration from Eq. (11.1)

$$v = e E \tau / m_e^* \tag{11.2}$$

The total current flowing is therefore:

$$I = n \, e^2 A \, E \tau / m_e^*$$

If the length of the conductor is L, the electric field can be replaced by V/L, where V is the voltage applied to the conductor, to give:

$$I = (n \, e^2 \, \tau / m_e^*)(A \, V / L)$$

Ohm's law can be written:

$$I = (A \, V / L \, \rho)$$

so that:

$$1/\rho = \sigma = n \, e^2 \, \tau / m_e^* \tag{11.3}$$

where σ is the conductivity and ρ the resistivity of the solid.

The conductivity is often written in terms of another variable, the *mobility* of the electrons, μ_e, defined as the drift velocity gained per unit electric field, that is:

$$v = \mu \, E$$

Comparing this with Eq. (11.2) makes it apparent that:

$$\mu = e \, \tau / m_e^*$$

This is sometimes called the *drift mobility*, to distinguish it from mobility measured via the Hall effect (Section 11.2.5). The conductivity and the mobility are then related by substituting Eq. (11.3) into (11.2) to give:

$$\sigma = n \, e \, \mu_e$$

11.1.3 Resistivity

Scattering is the main cause of resistivity. The electron wave can be scattered by interaction with *phonons* (lattice vibrations), called *thermal scattering*. As the temperature increases so do the lattice vibrations, and the resistivity rises. At low temperatures the resistivity drops gradually to a finite value, maintained at absolute zero (Figure 11.4), except for the superconductors (Section 11.5). This is an *intrinsic* property of a metal and cannot be altered. *Structural imperfections* present in the solid also contribute to resistivity. These are mainly defects such as dislocations, grain boundaries, and impurities. As with lattice vibrations, they scatter the electron waves and so increase resistivity. Because of this, impure solids, including alloys, have a higher resistivity than pure metals at all temperatures. Defects and impurities are *extrinsic* features that can be removed by careful processing.

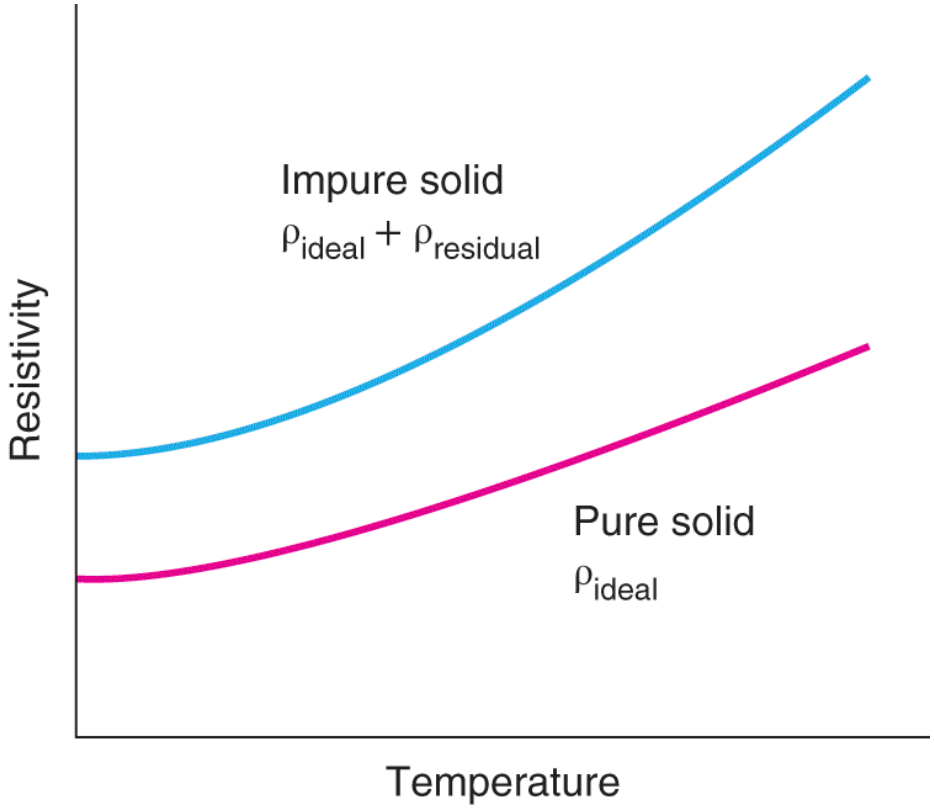


Figure 11.4 The variation of the resistivity of a metal with temperature, schematic.

The different scattering processes can be allocated relaxation times. Suppose that the distance between thermal scattering events is Λ_{th}. The relaxation time for this process will be given by:

$$\tau_{th} = \Lambda_{th}/v_F$$

where v_F is the electron velocity at the Fermi surface, and τ_{th} is the thermal relaxation time. Exactly analogous equations can be written in respect of the distance between defect scattering, Λ_d, and between impurity scattering, Λ_i. Thus, the resistivity can be written:

$$\rho = m_e^*/e^2 n\,\tau_{th} + m_e^*/e^2 n\,\tau_{def} + m_e^*/e^2 n\,\tau_{imp}$$

Taking into account these features, the total resistivity can be written as:

$$\rho = \rho_{th} + \rho_d + \rho_i$$

This is known as *Matthiessen's rule*, sometimes written as:

$$\rho = \rho_{\text{ideal}} + \rho_{\text{residual}}$$

where ρ_{ideal} is the intrinsic component due to phonon interactions and ρ_{residual} is the extrinsic contribution. The resistivity of a substitutional solid solution alloy will generally be greater than that of a pure metal because the elements added to form the alloy have the same effect as impurities (Figure 11.5). If the alloying atoms order to form a new crystal structure, the disruptive scattering of the electron wave is suppressed and the resistivity will drop.

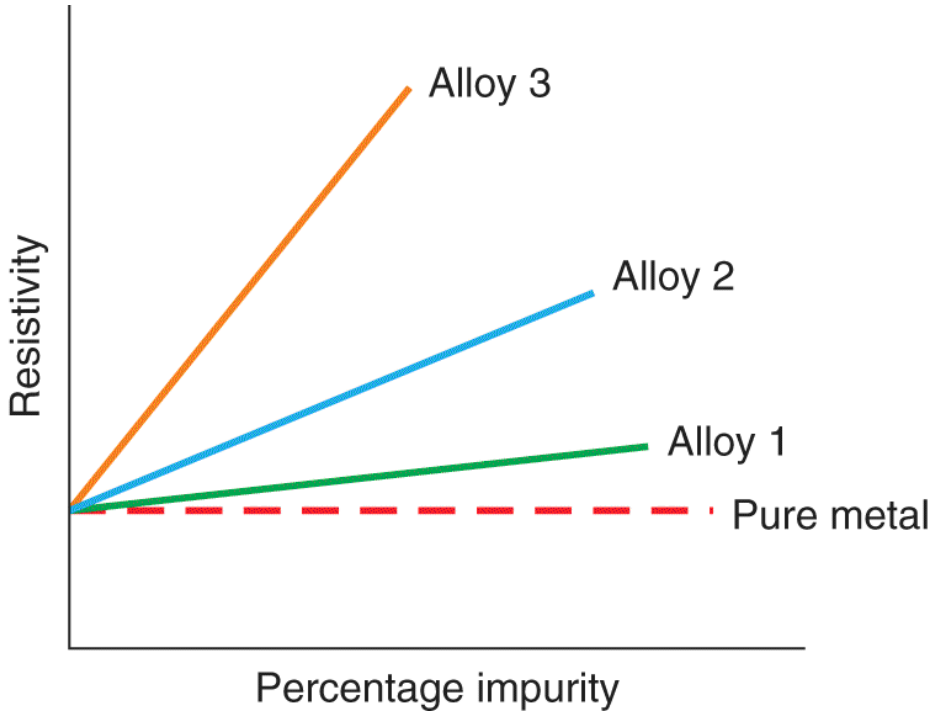


Figure 11.5 The variation of the resistivity of alloys with concentration of alloying elements, schematic.

A consequence of the electron collisions is a transfer of energy from the mobile electrons to the structure. This is revealed as heat energy, and accounts for the fact that an electric current generates heat. This effect, which takes place uniformly along the length of the conductor, is called *Joule heating*. The amount of heat generated is:

$$P = I^2 R$$

where P is the power output, I is the current, and R is the resistance. This heating poses problems for closely packed electronic circuits, which have to be cooled to function correctly.

11.2 Semiconductors

11.2.1 Intrinsic Semiconductors

The simplest flat band picture of a semiconductor describes the energy gap E_g between the top of the valence band, E_v, and the bottom of the conduction band, E_c, as constant (Figure 11.1b). Electron energy *increases* (and is defined to be *positive*) when measured *upwards* from the top of the valence band, which is usually taken as the energy zero. Hole energy *increases* (and is defined to be *positive*) when measured *downwards* from the top of the valence band. At absolute zero, the valence band will be full and the conduction band empty. As the temperature increases, some electrons will be promoted across the narrow band gap and the material will show a small degree of conductivity. Because the number of electrons promoted will increase with temperature, the conductivity will rise with temperature. This increase is characteristic of a semiconductor, rather than the magnitude of the conductivity. (Remember that for metals, conductivity falls with increasing temperature.)

The conductivity, σ, of a semiconductor is made up of two components, one due to electrons:

$$\sigma_e = n\, e\, \mu_e$$

and one due to holes:

$$\sigma_h = p\, e\, \mu_h$$

so that the overall conductivity will be given by:

$$\sigma = n\, e\, \mu_e + p\, e\, \mu_h$$

where the subscripts e and h refer to electrons and holes, respectively, the number of electrons is given by n, the number of holes by p, and the mobility by μ.

It is possible to determine the number of electrons excited into the conduction band by thermal energy in an intrinsic semiconductor using Fermi–Dirac statistics. If, as is usual, the energy at the top of the valence band, E_v, is set at zero, and the energy difference between the top of the valence band and the bottom of the conduction band is written as E_g, for ordinary semiconductors at normal temperature:

$$n = N_c \exp[-(E_g - E_F)/k_B T]$$

where E_F is the Fermi energy and N_c is the effective density of states function in the conduction band, given by:

$$N_c = 2[(2\pi\, m_e^* k_B T)/h^2]^{3/2}$$

where m_e^* is the effective electron mass for electrons in the bottom of the conduction band.

Similarly, the number of holes at the top of the valence band is found to be:

$$p = N_v \exp[-E_F/k_B T]$$

where N_v is the effective density of states of states function in the valence band, given by:

$$N_v = 2[(2\pi\, m_h^* k_B T)/h^2]^{3/2}$$

where m_h^* is the effective mass of the holes at the top of the valence band, and is generally different from that of electrons in a semiconductor.

In an intrinsic semiconductor, the number of holes is equal to the number of electrons (i.e. $n = p$) so:

$$N_c \exp[-(E_g - E_F)/k_B T] = N_v \exp[-E_F/k_B T]$$

Taking logarithms and rearranging allows the Fermi energy to be written as:

$$E_F = \tfrac{1}{2}E_g + \tfrac{1}{2}[k_B T \ln (N_v/N_c)] = \tfrac{1}{2}E_g + \tfrac{3}{4}k_B T \ln (m_h^*/m_e^*)$$

In the case when the effective masses are identical:

$$E_F = \tfrac{1}{2} E_g$$

In general, the Fermi level is always close to the centre of the band gap in an intrinsic semiconductor.

The approximate intrinsic carrier densities n_i and p_i can be found by substituting $\frac{1}{2} E_g$ for E_F, thus:

$$n_i = N_c \exp [-E_g/2k_B T]$$

$$p_i = N_v \exp [-E_g/2k_B T]$$

$$n_i p_i = N_c N_v \exp (-E_g/k_B T)$$

$$\text{As } n_i = p_i : n_i = p_i = (N_c N_v)^{1/2} \times \exp(-E_g/2k_B T) = 2(2\pi k_B T/h^2)^{3/2}(m_e^* m_h^*)^{3/4} \times \exp (-E_g/2k_B T)$$

Inserting values for the constants, and separating the effective mass of electrons and holes, gives:

$$n_i = p_i = 4.826 \times 10^{21}(m_e^* m_h^*/m_e^*)^{3/4} T^{3/2} \times \exp(-E_g/2k_B T)$$

(units m^{-3}).

Writing the total conductivity, σ, as

$$\begin{aligned}\sigma &= n\,e\,\mu_e + p\,e\,\mu_h \\ &= 4.826 \times 10^{21}(m_e^*\,m_h^*/m_e^*)^{3/4}T^{3/2} \\ &\quad \times e(\mu_e + \mu_h)\exp(-E_g/2k_BT) \\ &= 773.1\,(m_e^*\,m_h^*/m_e^*)^{3/4}T^{3/2}(\mu_e + \mu_h) \\ &\quad \times \exp(-E_g/2k_BT)\end{aligned}$$

The mobility of holes and electrons falls with increasing temperature due to interactions with the crystal structure, but at all but the lowest temperatures the exponential term is dominant and the overall conductivity increases with temperature.

11.2.2 Band Gap Measurement

The total conductivity, σ, can be expressed in the form:

$$\sigma = \sigma_0 \exp(-E_g/2k_BT)$$

where σ_0 is a constant, E_g is the band gap, k_B is Boltzmann's constant, and T is the absolute temperature. Taking logarithms:

$$\ln\sigma = \ln\sigma_0 - E_g/2k_BT$$

The gradient of a plot of conductivity vs. $1/T$ will therefore yield a value for the *thermal* band gap (Figure 11.6).

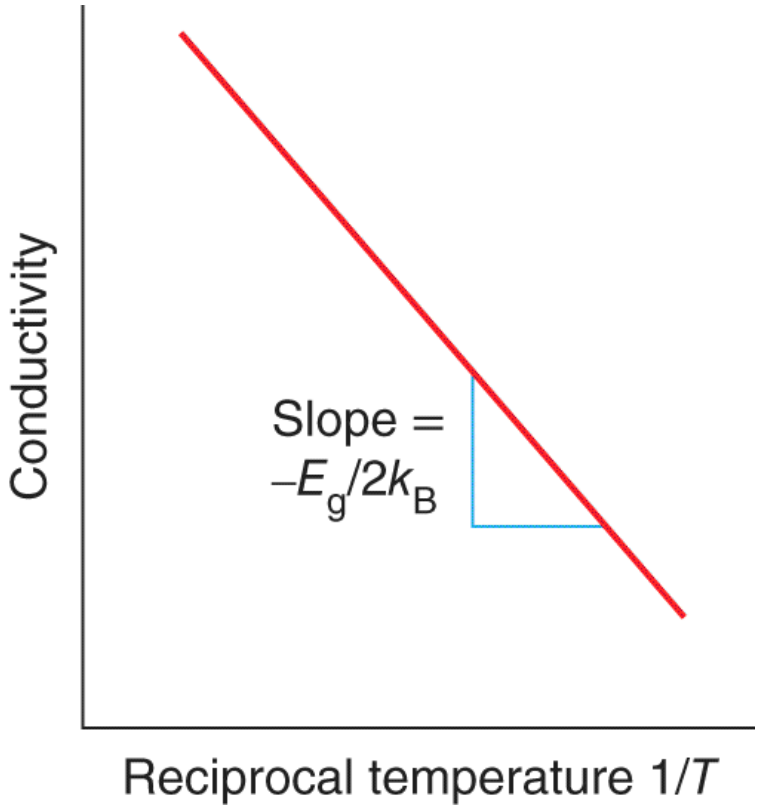


Figure 11.6 The variation of resistivity vs. temperature for an intrinsic semiconductor.

An alternative method of obtaining the magnitude of the band gap is via the absorption of radiation. Radiation falling onto a semiconductor crystal will only be absorbed by the electrons at the top of the valance band if they can then jump to higher energy levels. As there are no energy levels in the band gap, low energy radiation directed at a crystal will not interact with the electrons and the crystal will be transparent. As the energy of the radiation gradually increases, eventually it will just be sufficient to promote an electron from the top of the valence band to the bottom of the conduction band. The radiation will now be absorbed and the crystal will become opaque. The *optical* band gap can be equated to the energy at which this change occurs. Thus,

$$E_g = h\nu_g$$

where $h\nu_g$ is the energy of the photon required to promote an electron from the valence band and create a hole in its place. The measured value of the optical band gap is usually slightly different than the thermal band gap, reflecting the fact that the bands in a semiconductor are not flat, but have a more complex shape.

The band gap in a semiconductor decreases as the size of the component atoms increases (Table 11.1). The decrease is simply a consequence of the fact that the outer orbitals of larger atoms overlap more and give rise to wider bands. Thus, within the group C, Si, Ge, Sn, and Pb, diamond is best regarded as an insulator; silicon and germanium are the classical semiconductors, while grey tin and lead are regarded as metals. Note that the III–V semiconductors are so called because they are compounds of elements in groups III and V (now groups 13 and 15) of the periodic table and the II–VI semiconductors because they are compounds of elements in groups II and VI (now groups 12 and 16) of the periodic table.

Table 11.1 Approximate values for the band gap of some semiconductors.

Compound	**Formula**	**Band gap (eV)**[a]
Elements		
Diamond	C	5.47
Silicon	Si	1.12
Germanium	Ge	0.66
Grey tin	Sn	0.08
III–V semiconductors		2.26
Gallium nitride	GaN	3.34
Gallium phosphide	GaP	2.24
Gallium arsenide	GaAs	1.35
II–VI semiconductors		
Cadmium sulphide	CdS	2.42
Cadmium selenide	CdSe	1.70
Cadmium telluride	CdTe	1.56

[a] The band gap is normally given in electron volts in most compilations. 1 eV is equal to 1.60219×10^{-19} J.

11.2.3 Extrinsic Semiconductors

The deliberate addition of carefully chosen impurities to silicon, germanium, and other semiconductors is called *doping*. It is carried out so as to modify the electronic conductivity, and the dopants are chosen so as to add either electrons or holes to the material. It is possible

to gain a good idea of how this is achieved very simply. Suppose that an atom such as phosphorus, P, ends up in a silicon crystal. This can occur, for example, if a small amount of phosphorus impurity is added to molten silicon before the solid is crystallised. Experimentally the impurity atom is found to occupy a position in the crystal that would normally host a silicon atom, and so forms a *substitutional defect*.

Silicon adopts the diamond structure, in which each atom is linked to four tetrahedrally disposed neighbours by four sp^3 hybrid bonds (Figure 3.8). These use all of the four ($3s^2$ $3p^2$) valence electrons available. Phosphorus is found one place to the right of silicon in the periodic table, which indicates that the atom has one more electron in its complement. The outer electron structure of phosphorus is $5s^2$ $5p^3$, and after forming four sp^3 hybrid bonds, one electron is spare, and still associated with the phosphorus atom (Figure 11.7a). This electron is available to enhance the electrical conductivity if it can enter the conduction band. Atoms such as phosphorus are called *donors* when they are added to silicon as they can donate the unused valence electron to the conduction band.

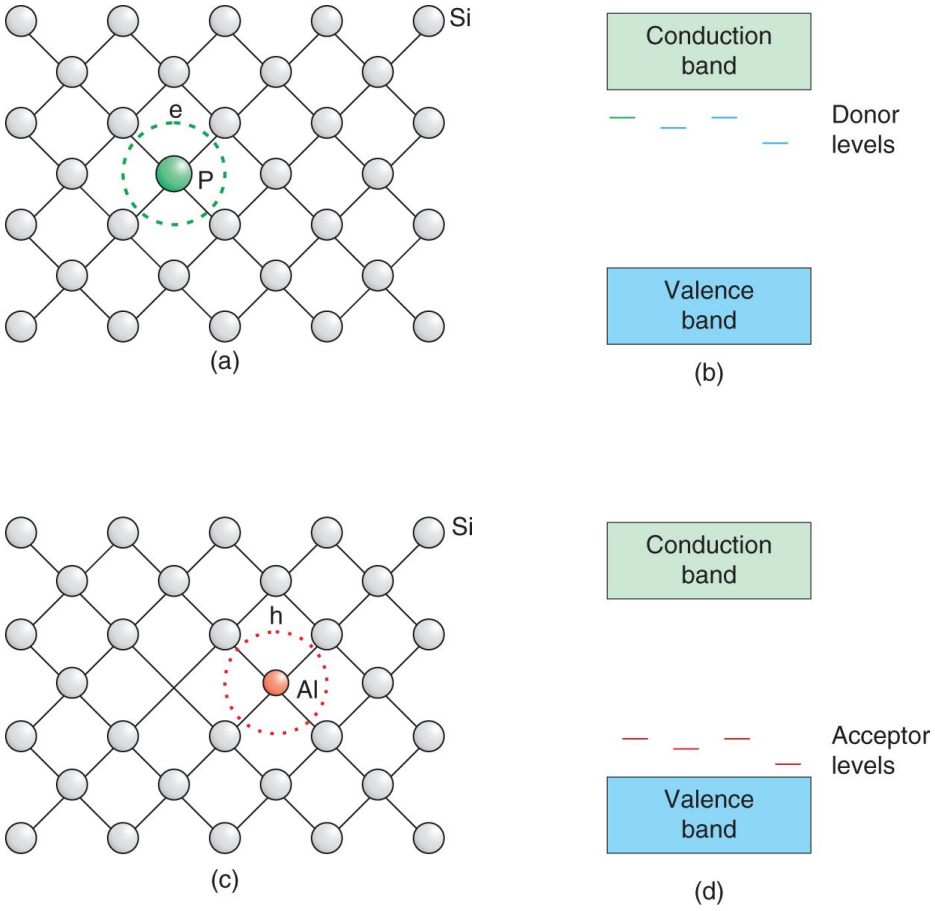


Figure 11.7 Impurities in a crystal of silicon: (a) donor (P) atoms; (b) donor energy levels; (c) acceptor (Al) atoms; (d) acceptor energy levels.

The simplest model to employ for the estimation of the energy needed to free the electron uses the Bohr theory of the hydrogen atom. In this model, a single electron is attracted to a positive nucleus consisting of a single proton. The energy needed to free this electron is given by:

$$E = -m_e e^4 / 8 \varepsilon_0^2 h^2$$

where m_e is the mass of the electron, e is the electron charge, ε_0 is the permittivity of free space, and h is Planck's constant. The value of E is -2.18×10^{-18} J (−13.6 eV). The negative value reflects the fact that zero is taken as the energy of a completely free electron. To apply this to an electron located at a phosphorus atom, suppose that the attraction of the phosphorus nucleus is 'diluted' by the relative permittivity of silicon, ε_r, and the mass of the electron is replaced by the effective mass m_e^*. The energy to free the electron is now:

$$E(\mathrm{P}) = -m_e^* e^4 / 8\,\varepsilon_0^2\,\varepsilon_r^2\,h^2$$
$$= E\,m_e^*/m_e\,\varepsilon_r^2$$

As the effective mass of an electron in silicon is approximately one tenth of the electron mass and the relative permittivity of silicon is about 10, the energy needed to free the electron, approximately 0.0136 eV, is about one hundredth of the band gap, which suggests that the electron should be very easily liberated. Donor energies are often represented by an energy level, the donor level, drawn under the conduction band (Figure 11.7b).

An analogous situation arises if silicon is doped with an element such as aluminium. Aluminium also forms substitutional defects. However, aluminium is found one place to the left of silicon in the Periodic Table, with a valence electron configuration $3s^2\ 3p^1$, and has one valence electron less than silicon. One of the four resulting sp^3 hybrid bonds will be an electron short. This is equivalent to the introduction of a positive hole, which is localised on the aluminium impurity (Figure 11.7c). Providing that the energy needed is not too great, an electron from the full valence band can be promoted to fill the bond, leaving a hole behind. The energy needed to free the hole, calculated using the Bohr model, is similar to that of an electron. Once in the valence band, the hole is free to move and to contribute to the conductivity. Dopant atoms from the left of silicon in the periodic table, such as aluminium, are called *acceptors*, because they can be imagined to accept an electron from the filled valence band and so create a hole that takes part in conductivity. These impurities can also be represented by energy levels, drawn just above the top of the valence band (Figure 11.7d).

The energy to liberate both donor electrons and acceptor holes is about 8×10^{-21} J (0.05 eV). These values are comparable with room temperature thermal energy, and most extrinsic electrons and holes should be free at room temperature. In this state the donors and acceptors are said to be *ionised*, and the semiconductor crystal will be a reasonable conductor. If donor atoms are present in great numbers, they will govern the conductivity, which will be by electrons. The material is said to be *n-type*. If acceptors are present in greatest quantities, then holes will control the conduction, and the material is said to be *p-type*. When both electrons and holes are present and both contribute to the conductivity we talk of *majority* and *minority carriers*.

11.2.4 Carrier Concentrations in Extrinsic Semiconductors

The creation of mobile electrons and holes in an *intrinsic* semiconductor is a dynamic process. There is a continuous excitation of electrons and these are continuously falling back to the valence band and recombining with holes. At ordinary temperatures dynamic equilibrium holds and we can write:

$$K = n\,p = n_i^2$$

where K is the equilibrium constant of the process, n is the concentration of electrons in the conduction band, p is the concentration of holes in the valence band, and the number of holes

equals the number of electrons. The equilibrium constant will depend upon temperature but not upon pressure or how much semiconductor is in the sample.

This equilibrium equation also applies to *extrinsic* semiconductors as the *origin* of the electrons and holes is not relevant. Therefore, the equilibrium constant derived for a pure intrinsic semiconductor is valid for a doped sample of the same semiconductor. This is an extremely useful finding, because it means that as the concentration of electrons in a semiconductor is increased by doping, so that n increases, the number of holes decreases proportionately. Similarly, if large numbers of acceptors are added, so that p increases, the number of electrons decreases. During doping the position of the Fermi level changes in order to maintain a balance between the charges in compliance with the relationship (n p) is constant. In the case of an intrinsic semiconductor, the number of holes and electrons is equal and the Fermi energy, E_F, is at the mid-point of the band gap, E_g (Figure 11.8a). In an n-type semiconductor the number of electrons is much higher than the number of holes and the Fermi energy moves towards the n-type side of the band gap to maintain the balance (Figure 11.8b). In a p-type semiconductor the number of holes is much higher than the number of holes and the Fermi energy moves towards the p-type side of the band gap to maintain the balance (Figure 11.8c).

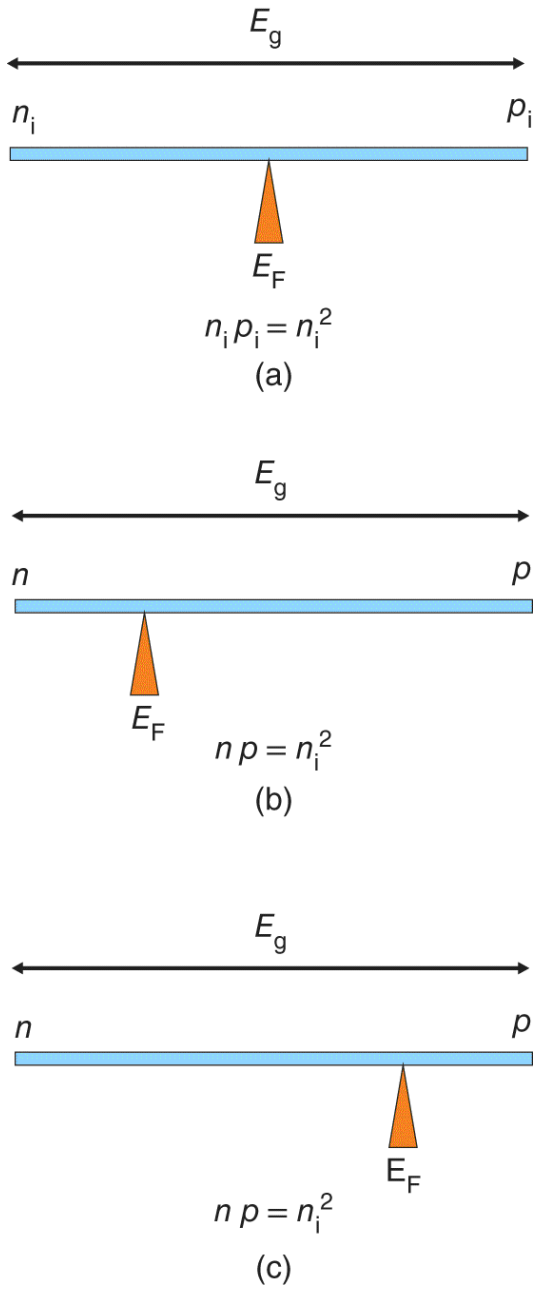


Figure 11.8 The position of the Fermi energy, E_F, in semiconductors: (a) intrinsic; (b) n-type; (c) p-type.

The number of holes created by the addition of acceptors is equal to N_a^-, where the number of acceptor atoms per unit volume is N_a, and the number of acceptors that have gained an electron, ionised acceptors, is N_a^-. In this case, at ordinary temperatures, the Fermi energy, E_F, moves so as to lie approximately half way between the top of the valence band and the acceptor energy levels (Figure 11.9a) Similarly, the number of electrons created by the addition of the donors is equal to N_d^+, where the number of donor atoms per unit volume is N_d, and the number of donors that have lost an electron, ionised donors, is N_d^+. In this case,

at ordinary temperatures, the Fermi energy, E_F, now moves so as to lay approximately half way between the bottom of the conduction band and the donor energy levels (Figure 11.9b).

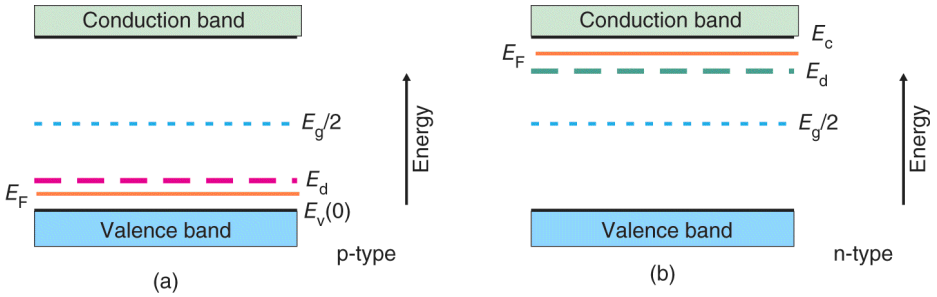


Figure 11.9 The position of the Fermi energy, E_F, in a semiconductor at ordinary temperatures; (a) p-type; (b) n-type.

As a doped crystal must remain electrically neutral:

$$n + N_a^- = p + N_d^+$$

At elevated temperatures, most of the donors and acceptors will be ionised so:

$$n + N_a = p + N_d$$

At high temperatures when an n-type semiconductor contains only completely ionised donors, the number of electrons is approximately equal to the number of donors and the number of holes is given by the equilibrium equation:

$$n\,(\mathrm{n-type}) \sim N_d$$

$$p\,(\mathrm{n-type}) \sim n_i^2/N_d$$

Similarly, for a p-type semiconductor crystal that contains only acceptors at high temperatures, the number of holes is approximately equal to the number of acceptors and the number of electrons is given by the equilibrium equation:

$$p\,(\mathrm{p-type}) \sim N_a$$

$$n\,(\mathrm{p-type}) \sim n_i^2/N_a$$

The position of the Fermi level is also a function of the temperature. As stated, at low temperatures, the position of the Fermi level approaches the conduction band in n-type materials and the valence band in p-type materials. As the temperature increases, the contribution from the intrinsic electrons and holes increases. In essence, the semiconductor gradually changes from an extrinsic towards an intrinsic material. Because of this, the Fermi level approaches the position found in intrinsic semiconductors, near to the middle of the band gap (Figure 11.10). The rapidity of this change will depend upon the concentrations of donors and acceptors in the semiconductor. A semiconductor will maintain its extrinsic character to higher temperatures when doped with higher concentrations of impurities.

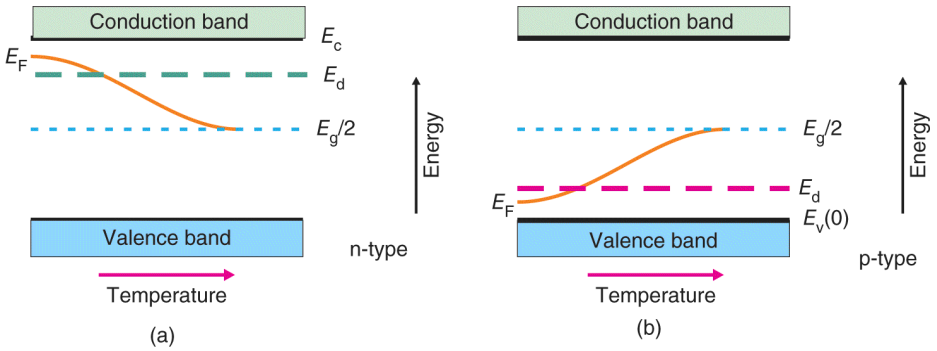


Figure 11.10 The temperature variation of the position of the Fermi energy, E_F, in a semiconductor; (a) n-type; (b) p-type.

11.2.5 Characterisation

Although the resistance of a metal can be measured easily by attaching two contacts, this gives unreliable results for semiconductors. For these materials, the resistivity is most often determined using a *four-point probe*, on rectangular or disc shaped samples. The probes are sharply pointed, equally spaced needles, and press down with a known force on the surface of the sample. The current passes between two outer probes, while the voltage drop is measured between the two inner probes. The relationship between the measured voltage and current, and the resistivity depends upon the geometry of the experimental set up. For a bulk specimen, in which the thickness of the sample is much greater than the spacing between the probes:

$$\rho = 2\pi s\,(V/I)$$

where ρ is the resistivity of the material, s the distance between the probe needles, V is the voltage drop between the middle two needles, and I is the current between the outer two needles. In the case of a thin film, in which the probe spacing, s, is much greater than the film thickness, t, and much smaller than the distance to the edge of the film, that is $(l, w) \gg s \gg t$ (Figure 11.11a), or $d \gg s \gg t$ (Figure 11.11b), the expression for the resistivity is:

$$\rho = (\pi\, t/\ln 2)\,(V/I) = 4.54\, t\,(V/I)$$

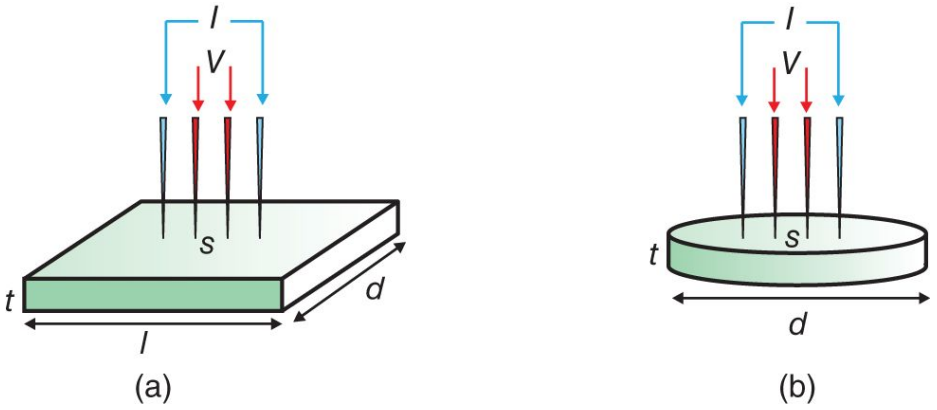


Figure 11.11 Arrangement of the four-point probe method of measurement of resistance of semiconductors, schematic: (a) rectangular specimen; (b) disc.

This formula is independent of the probe spacing, s.

The resistance of thin films is sometimes reported as the *sheet resistance*, R_s. This is defined in terms of the bulk resistance of a material, R, of resistivity ρ, and the dimensions of the sample. For a rectangular specimen:

$$R_s = \rho/t\ \rho = (R\,w\,t)/I$$

hence

$$R_s = R\,w/I\ \rho = R_s\,t$$

The sheet resistance is quoted as ohms per square, Ω/□.

The carrier type can be found in two ways, via the *Hall effect* or the *Seebeck effect* (Section 13.4). The Hall effect, discovered in 1879, relies upon the displacement of moving charges in a magnetic field to determine the nature of the mobile carriers. A slab of material is arranged so that the current flow is normal to a fairly strong (0.2 T), magnetic induction (Figure 11.12). The moving charges will experience a force due to the magnetic field generated in the solid that is normal to both current direction and magnetic field. That is, if the current is along the **x**-axis and the magnetic induction is along the **z**-axis, the charge carriers will be deflected in the **y**-axis direction. This deflection will build up until the electric field is just strong enough to oppose further charge displacement. The result is a voltage, the *Hall voltage*, along the **y**-axis. Perhaps counter intuitively, the charge carriers will be deflected in the same direction, irrespective of the charge that they carry, and the sign of the voltage will give the sign of the charge carriers. The value of the Hall voltage, V, can be appreciable.

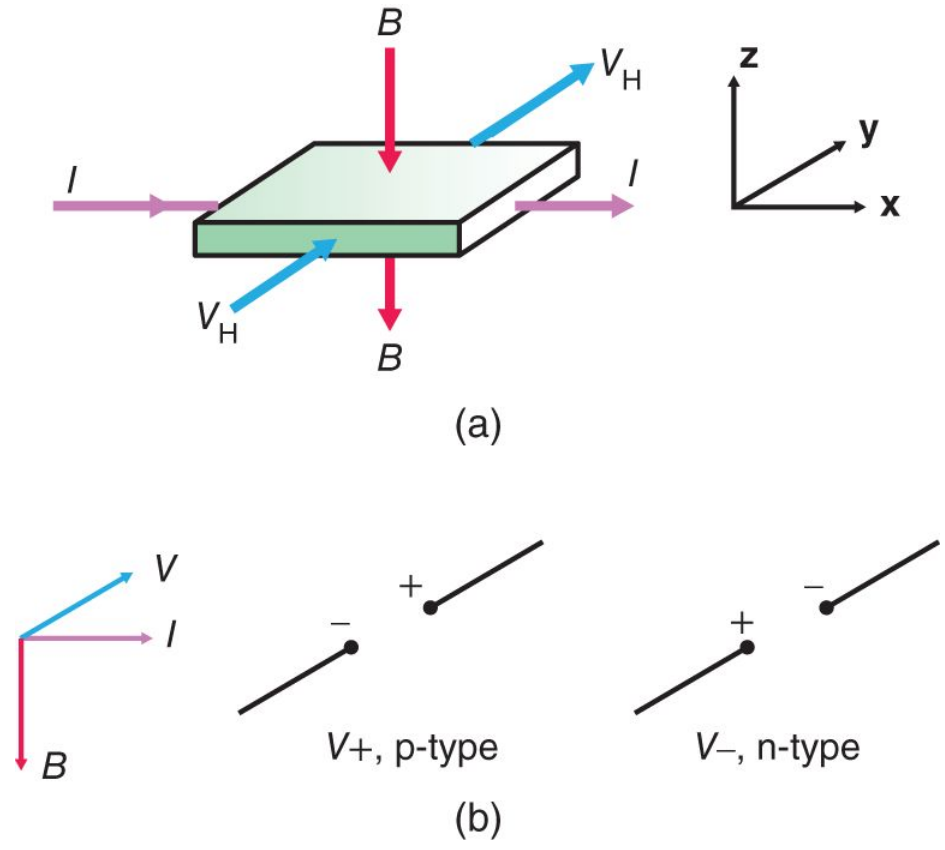


Figure 11.12 Measurement of the Hall effect, schematic: (a) reference axes; (b) Hall voltage sign.

The relationship between the current, magnetic induction and electric field is:

$$\mathbf{E}_y = \pm R_{\mathrm{H}} \mathbf{J}_x \mathbf{B}_z \tag{11.4}$$

where $\mathbf{E}_y$ is the electric field along the y-axis, $\mathbf{J}_x$ is the current density along the **x**-axis, and $\mathbf{B}_z$ is the magnetic induction along the **z**-axis. The sign of the Hall coefficient differentiates between n-type and p-type semiconductors. For example, if the magnetic induction is aligned along **–z** (Figure 11.12a), a positive Hall voltage compared with the orthogonal **V**, **B**, and **I** axes, yields a positive value for R_{H}, and the material is p-type, while a negative voltage and negative R_{H} indicates that the material is n-type (Figure 11.12b).

The Hall coefficient is related to the number of mobile charge carriers in a simple way. Suppose that the current is made up of electrons flowing parallel to **x** with a drift velocity $\mathbf{v}_x$. The magnetic field present will exert a force, **F**, on an electron:

$$\mathbf{F} = \mathrm{e}\, \mathbf{v}_x \mathbf{B}_z$$

The force exerted by an electric field, $\mathbf{E}_y$, on an electron is

$$\mathbf{F} = \mathbf{E}_y \mathrm{e}$$

When the two forces are equal and equilibrium is achieved:

$$\mathbf{E}_y = \mathbf{v}_x \mathbf{B}_z$$

The current density is given by:

$$\mathbf{J}_x = -n\, e\, \mathbf{v}_x$$

where n is the number of mobile electrons per unit volume. Substituting for $\mathbf{E}_y$ and $\mathbf{J}_{\mathrm{x}}$ into Eq. (11.4):

$$R_{\mathrm{H}} = -1/n\, e$$

where n is the number of electrons per unit volume, each with a charge of $-e$. For a flow of positive holes:

$$R_{\mathrm{H}} = 1/p\, e$$

where p is the number of mobile holes per unit volume, each with a charge of $+e$.

In general, only one mobile charge carrier, either electrons or holes, is present. In this case, by measuring both the conductivity, σ, and the Hall coefficient, R_{H}, it is possible to determine the number of charge carriers and their mobility, as:

$$\sigma_{\mathrm{e}} = n\, e\, \mu_{\mathrm{e}}$$

$$-R_{\mathrm{H}}\, \sigma_{\mathrm{e}} = \mu_{\mathrm{e}}$$

or

$$\sigma_{\mathrm{h}} = p\, e\, \mu_{\mathrm{h}}$$

$$+R_H \sigma_h = \mu_h$$

If both electrons and holes are present:

$$\sigma\,(\text{total}) = n\,e\,\mu_e + p\,e\,\mu_h$$
$$= |e|(n\,\mu_e + p\,\mu_h)$$

and

$$R_H = [(n\,\mu_e{}^2 + p\,\mu_h{}^2)/(n\,\mu_e + p\,\mu_h)^2]/|e|$$

The mobility derived from a Hall measurement is often called the *Hall mobility* to distinguish it from the drift mobility (Section 11.1.2).

For non-cubic single crystals, the Hall coefficient varies with direction, although the effect is averaged when measurements are made on polycrystalline samples. A number of ordinary metals have positive Hall coefficients, including Be, Al, Cd, In, As, and W. These results could not be explained in terms of the classical electron gas model of electronic conductivity. In some metals, such as Er and Ho, the Hall coefficient varies from negative to positive as a function of crystallographic direction. In addition, the Hall effect becomes more complex when thin films subjected to high magnetic fields at low temperatures are investigated. In such cases the resistivity of the sample will jump from one level to another as the magnetic field increases in abrupt steps. These sudden jumps are a manifestation of the fact that the electron dynamics are under the control of quantum mechanics, not, as is supposed earlier, classical (continuum) mechanics. This phenomenon is called the *quantum Hall effect*. The explanation of all these findings requires a detailed knowledge of the three-dimensional shape of the Fermi surface in these metals and goes far beyond the flat-band picture and the mathematics used here. (For further information on this see Further Reading, especially S. Michalakis.)

11.2.6 The p–n Junction Diode

The electrical behaviour that emerges when a region of p-type semiconductor is adjacent to a region of n-type semiconductor, a *p–n junction diode*, often just called a *diode*, is different than the behaviour of the separate components. The electronic properties of this important device can be explained in terms of the simple flat-band model. In separated materials, the Fermi energies are unequal and this can be imagined to be the situation occurring immediately on joining the two components (Figure 11.13a,b). (Note that p–n junction diodes are fabricated by selective doping of different regions of a semiconductor crystal and not by joining separate crystals together.) When a p-type region abuts an n-type region, electrons move into the p-type region from the n-type side and holes move into the n-type region from the p-type region, by diffusion. (In a simplistic way it is possible to say that if the Fermi level slopes it is energetically favourable for electrons to ‘roll downhill’ and holes to ‘roll uphill’.) Most of these displaced charge carriers will *recombine*; holes with electrons in the n-region and electrons with holes in the p-region. As electrons leave the n-type material, positively charged donor atoms are left behind, while negatively charged acceptor atoms are left in the p-type material as holes leave. These charges will create an electric potential, the *contact potential*, of about 0.3 V. At equilibrium, the Fermi level must be the same on each side of the junction (Figure 11.13c). To achieve this, the energy levels have been shifted vertically with respect to each other to give a distorted band structure in the junction region (Figure 11.13d). This means that the populations of electrons and holes change dramatically as the junction is traversed (Figure 11.13e).

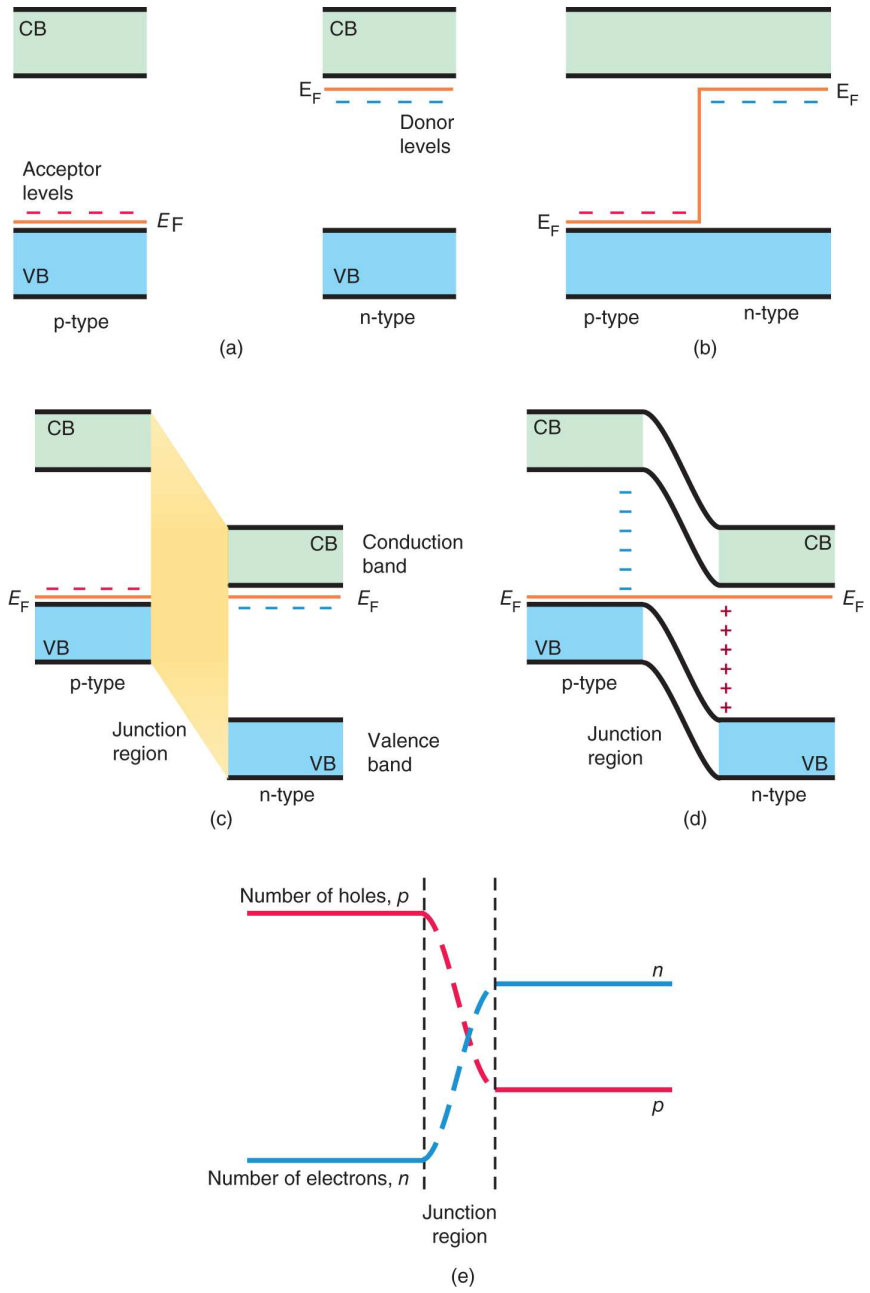


Figure 11.13 The p-n junction schematic: (a) energy bands of separate p-type and n-type solids; (b) energy bands of p-type and n-type solids juxtaposed; (c) energy bands displaced to make the Fermi energy continuous; (d) distorted energy bands in the junction region at equilibrium; (e) electron and hole populations across the junction at equilibrium.

The transition region has a width of about 1 μm. The density of mobile charge carriers in the transition region is low, and for this reason the transition region is also called the *depletion region*. It is important to note that at equilibrium (thermal and electrical) there will still be an *exchange of carriers* at the junction, but the current flows will cancel and the region is in a state of dynamic equilibrium.

The conductivity of the p–n junction in one direction is totally different to the other. An applied voltage, which will drop across the transition region, because of the absence of mobile

charge carriers, can be applied with the positive side connected either to the p-type region, called *forward bias*, or to the n-type region, called *reverse bias*. Under a forward bias the potential barrier is significantly reduced. In effect the Fermi level changes so that it is energetically favourable for the electrons to move 'downhill' and the holes to move 'uphill', causing a rapid increase in the current flowing across the junction (Figure 11.14a). The effect of a reverse bias is to raise the potential barrier, so that it is energetically unfavourable for both the electrons and the holes cross it. Current flow now virtually ceases (Figure 11.14b).

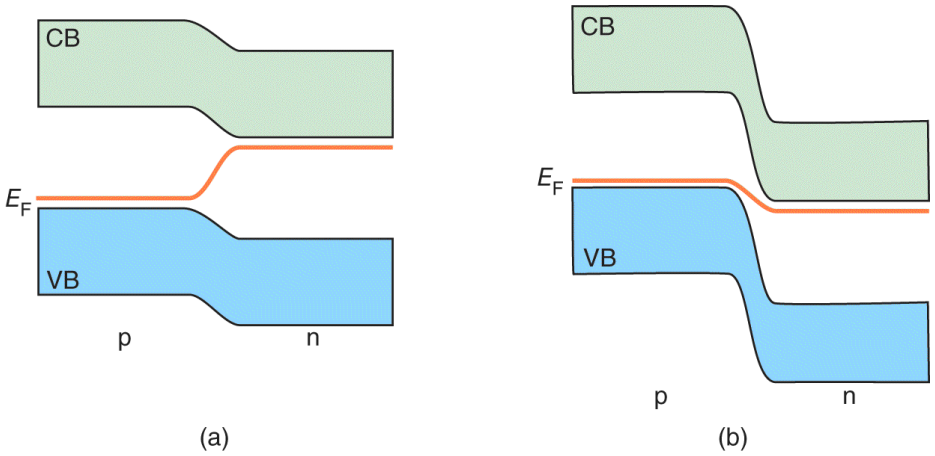


Figure 11.14 Band structure across a p–n junction: (a) under forward bias; (b) under reverse bias.

The change of current with applied voltage is given by the *Shockley equation*. In a simplified form this is:

$$I = I_0 \left[\exp\left(e\,V/k_B\,T\right) - 1\right]$$

where I_0 is a constant term, the saturation current, determined by the junction geometry and the doping levels, e is the charge on the electrons and holes, V is the applied voltage, k_B is Boltzmann's constant, and T the temperature (Figure 11.15).

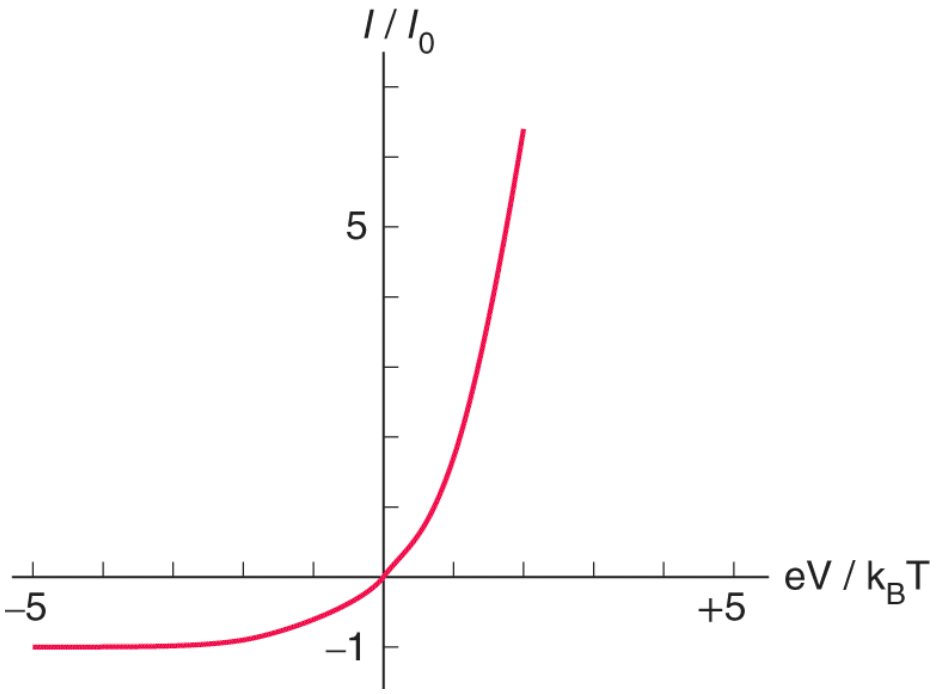


Figure 11.15 Ideal current–voltage curve for a p–n junction.

The total current across the device, which is constant for any applied voltage, is made up of hole and electron flows in opposite directions. When a forward bias is applied the number of electrons moving to the left will increase rapidly, by a factor of $\exp(V/k_BT)$. If V is 0.1 V, this is a factor of about 55× at room temperature. Thus, the number of electrons appearing at the p-type boundary is about 55 times higher than the equilibrium concentration there. A similar situation describes the holes appearing at the n-type boundary. In general, the hole current will be greater than the electron current under forward bias, but the actual currents on each side will depend upon the doping levels.

When the electrons reach the p-type region, they are annihilated by combination with holes. The penetration depth depends upon a number of factors, but can be taken to be about 1 mm. In order to maintain charge neutrality, the hole population must be replenished and holes must be move into the p-type region from the left to balance the electron density. Similarly, the holes that arrive in the n-type region are gradually annihilated by recombination with electrons. In order to maintain charge balance, an extra electron current must flow into the n-type region from the right.

The total current flowing will be made up of six components (Figure 11.16):

$$I_a = I_b + I_e + I_g = I_c + I_d + I_f$$

where I_a is the constant total current; I_b is the electron current flowing in the n-type region (constant); I_c is the injected (introduced) electron current in p-type region (decaying); I_d is the hole current in p-type region (constant); I_e is the injected hole current in n-type region (decaying); I_f is the declining hole current in p-type region to balance and annihilate I_c; and I_g is the declining electron current in n-type region to balance and annihilate I_e. This is quite different than in a metal, in which an applied voltage allows the mobile electrons to acquire a drift voltage. In a p–n junction, minority carriers (that is, electrons in the p-type region and

holes in the n-type region) are inserted into both regions. They were not there originally and arise as a consequence of the applied voltage.

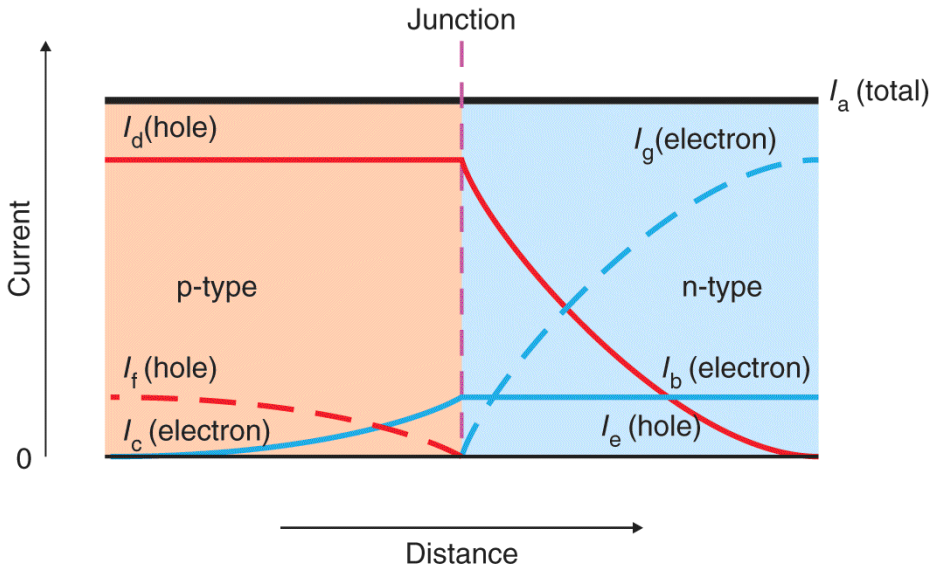


Figure 11.16 The currents flowing across a p–n junction under forward bias.

From what has been said, it can be seen that a p–n junction diode acts as a *rectifier*. That is, the device exhibits a very low resistance to current flow in one direction and a very high resistance in the other.

11.3 Metal–Insulator Transitions

11.3.1 Metals and Insulators

The simple theory of metals indicates that metallic conduction arises when the highest energy band is only part filled with electrons. This idea works well considering its simplicity. For example, WO_3, with no outer electrons, is an insulator while ReO_3, which adopts essentially the same structure as WO_3, has a $5d^1$ outer electron and a partly filled d-band, making it metallic. Similarly, WO_2 has a $5d^1$ outer electron band and is also metallic. However, a large group of solids are found to be insulators when the theory suggests that they should be metals. These materials are often called *Mott insulators*. Furthermore, many also show *metal-insulator transitions* where a change from an insulating to a conducting state occurs.

The most significant of the problems of the metallic vs. insulating state are exhibited by compounds of the 3d transition metals and the lanthanoids. A classic case is that of the 3d transition-metal monoxides, MO, consisting of TiO, VO, MnO, FeO, CoO, and NiO. These all adopt the halite (NaCl) structure and in all of them the d orbitals, and hence the d bands that arise are only partly filled, which should endow all with metallic conductivity. In reality only TiO and VO are metals at room temperature; all of the others are antiferromagnetic insulators.

It is possible to make a number of immediate suggestions to try to remedy this situation. It could be argued that the intervention of oxygen atoms between the metal atoms disrupts the formation of a d band. This is a satisfactory first answer, and the idea that the d electrons remain shielded from external influences and remain localised on their respective metal atom cores works well in practice. However, it does not account for metallic TiO and VO compared to insulating MnO and FeO.

A second suggestion might be that the antiferromagnetic ordering in, for example, NiO, is sufficiently strong to prevent any electron movement. This also fails because NiO remains an insulator above the Néel temperature at which the antiferromagnetism is lost. Another suggestion is that the d orbitals in these compounds are split into the t_{2_g} and e_g sets, thus giving rise to two bands, which may inhibit metallic conduction. However, this cannot be so as the d-electron population of these orbitals is such that one or the other band would be only part filled. In NiO, for instance, each Ni^{2+} (d^8) would contribute six electrons to the t_{2_g} band, filling it, but only two electrons to the e_g band that is capable of holding four, resulting in a half-filled band.

These difficulties were tackled during the middle years of the twentieth century and gave rise to a number of modifications to simple band theory. The mathematics of these modifications is rather complex and just one example – the effects of electron–electron repulsion will be described.

11.3.2 Electron–Electron Repulsion

The simple theory of metals suggests that when atoms are close atomic orbitals can overlap to form energy bands and electrons in partly filled bands can move freely throughout the solid. As the atoms get further and further apart, the bands become narrower and the electron movement is impeded more. The electrons take on an increased effective mass, but they are still free to move. Now this does not correspond with reality. At some stage the atoms must be considered to be essentially isolated, and in this situation, electrons are localised on each atomic core. Suppose that the atoms each have a single valence electron outside of a closed

core in the ground state (Figure 11.17a). To move this electron from one atom now requires an energy expenditure equal to the ionisation energy of the atom, *I*. Placing the electron onto the receiving atom recoups energy equal to the electron affinity of the atom, *A* (Figure 11.17b). The difference between these two energy terms, *U*, is given by

$$U = I - A$$

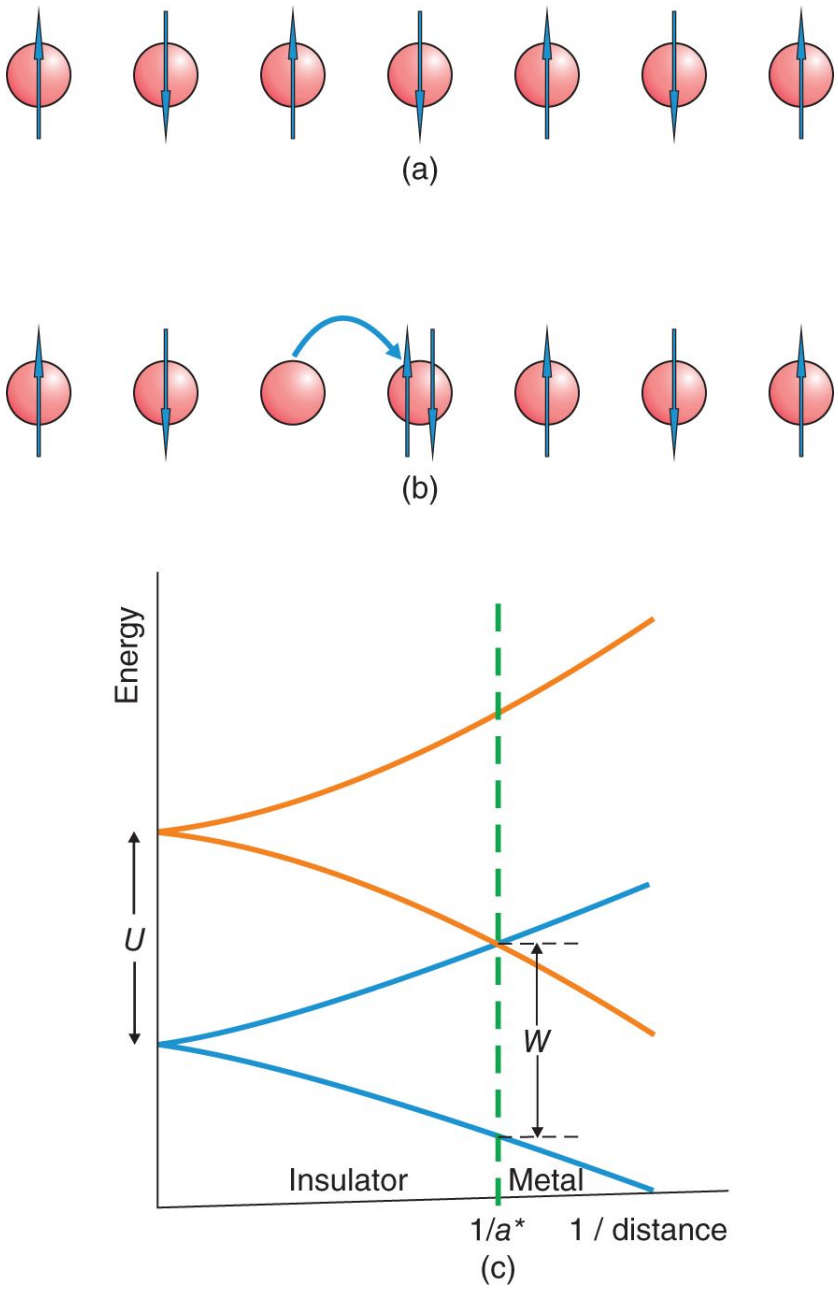


Figure 11.17 Weakly interacting atoms: (a) ground state with one localised electron per atom; (b) excited state after electron transfer; (c) Hubbard subbands as a function of inverse interatomic distance.

The value of *U* can be considered essentially as the extra energy due to electron–electron repulsion between electrons now on the same atomic core.

The energy level depiction will show a ground state, when the electrons are each on a separate atom core and an excited state when electron transfer has been achieved. The two bands, termed *Hubbard subbands*, are separated by an energy gap U, called the *Mott–Hubbard splitting* (Figure 11.17c). The lower subband, the ground state, is only half-full, because each atom contributes just one electron. However, this is an *insulating band*, because energy U is needed to move an electron to another atom and, on the energy level diagram, jump from the ground state to the excited state. The insulating nature is primarily due to the electron–electron repulsion.

The situation changes as the spacing, a, between atoms decreases. This results in increased interaction between the atoms, leading to broader bands. Continued decrease of interatomic spacing will ultimately make the bands wide enough to overlap when the band width, W, is approximately equal to the Mott–Hubbard splitting, U. At this point, when the reciprocal of the interatomic spacing is $1/a$, the energy to transfer an electron from one atom to its neighbour becomes zero (Figure 11.17c). The now composite band allows free electron transport and the material undergoes an insulator to metal transition at the specific interatomic spacing that generates band overlap.

The development of Hubbard subbands is believed to occur in the transition metals and lanthanoids. In the case of the 3d transition metals the electron transfer is from a cation with an electron configuration $3d^n$ to a similar neighbour, giving rise to a $3d^{n+1}$ ion and leaving behind a $3d^{n-1}$ ion:

$$M(3d^n) + M(3d^n) \rightarrow M(3d^{n-1}) + M(3d^{n+1})$$

The simple d band is split into two subbands the lower, ground state band corresponding to ($3d^n$) cations, although only partly filled, is n insulating band, because it requires an energy U to move an electron. The changes in the electronic behaviour of the 3d transition metal monoxides can be understood in these terms. Nickel oxide, with eight electrons in the subband, has a low degree of d-d interaction and the subband is insulating. As we move along the series from NiO towards TiO, the ionic radii of the M^{2+} cations increase from 0.083 nm for Ni^{2+} to 0.10 nm for Ti^{2+}. This increase in size is a reflection of the increased interaction between the d orbitals on adjacent cations, which causes the bands widen. Eventually when V and Ti are reached the overlap of the upper and lower subbands has occurred and the oxides are metallic.

11.3.3 Modification of Insulators

Metal–insulator changes can also be caused by a change of crystal structure, temperature, pressure, or composition. An example of a metal–insulator transition due to a change in crystal structure is given by VO_2. Above 67 °C VO_2 adopts the rutile structure, in which the V^{4+} ($3d^1$) atoms are found centrally within chains of edge-sharing VO_6 octahedra running parallel to the **c**-axis. The d band is partially filled and as expected from simple theory, the material is metallic. At temperatures below 67 °C, however, the structure distorts into a monoclinic symmetry, and V^{4+} atoms in adjacent octahedra pair. The resistivity increases by a factor of about 10^5 and the metallic properties are lost. The reason is that the d electrons are now localised in this pair bonding and are no longer able to move freely through the bulk. The partly filled d band is now an insulating band. The transition is reversible, and raising the temperature above 67 °C restores the rutile structure and metallic properties are regained. (For further consequences of this metal-insulator transition, see Section 13.6.2.)

An increase in pressure will generally cause interatomic spacing to decrease. In terms of the Hubbard model, this results in an increase in band width W, with little change in repulsion energy, U. A metal-insulator transition at high pressures is thus to be expected for many solids. In this respect it has long been suggested that solid hydrogen could become metallic at sufficiently high pressures. (This was verified experimentally in 2020.) There is also speculation that metallic hydrogen might be present on the planet Jupiter.

A typical example of such a pressure induced metal–insulator transformation is shown by the perovskite-related Mott insulator phase Ca_2RuO_4, which adopts the K_2NiF_4 structure (Figure 11.18). In this phase, Ru^{4+} ($4d^4$) ions occupy oxygen octahedra, but the width of the d band is small because the d orbital interactions are weak and electron–electron repulsion dominates. Under normal pressure the partly filled d band is insulating and this material is an antiferromagnetic insulator. It transforms into a metal when subjected to pressures greater than 0.5 GPa due to the increase in band width as the d orbitals are forced into greater overlap.

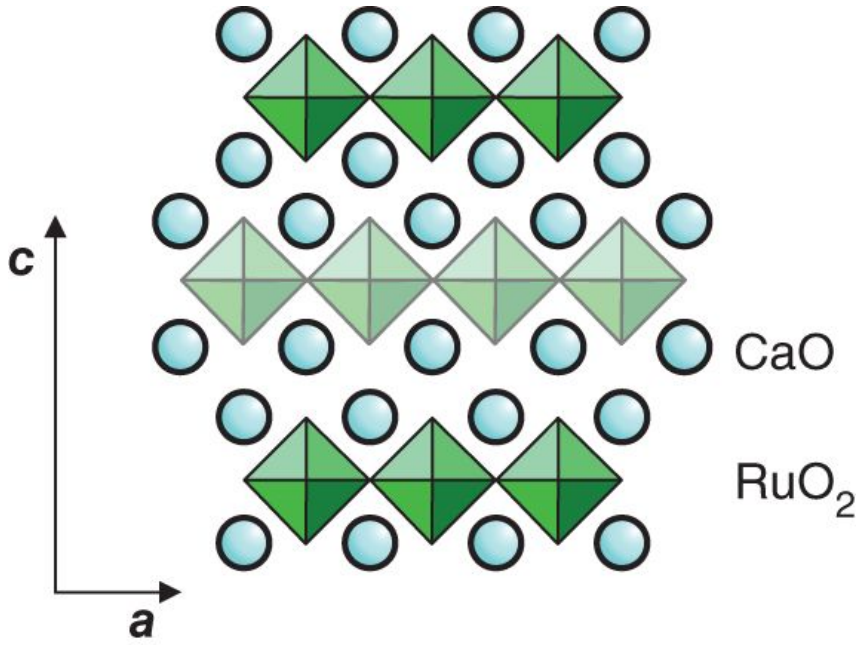


Figure 11.18 The (slightly idealised) structure of Ca_2RuO_4 projected down [010]. The central layer is displaced relative to its neighbours by $(a + b)/2$.

Pressure changes can often be mimicked by controlled doping. The insulating parent phase Ca_2RuO_4 can be manipulated in this way by doping with Sr in the system $Sr_xCa_{2-x}RuO_4$. It is found that the doped phase becomes metallic at ordinary pressures at a composition of approximately $Sr_{0.2}Ca_{0.8}RuO_4$.

Doping in this way generally forces the metal–insulator transition in the parent phase to lower temperatures as the amount of dopant increases (also see Section 13.6.2). As well as modifying transition temperatures, doping can also change the electronic properties of insulators by acting as donors or acceptors. Nickel oxide is an antiferromagnetic insulator, nominally containing equal numbers of Ni^{2+} and O^{2-} ions. Green nickel oxide can be reacted with colourless lithium oxide, Li_2O, to give a black solid solution $Li_xNi_{1-x}O$. The Li^+ ions occupy Ni^{2+} sites in the structure to form substitutional defects. In order to maintain charge neutrality, every Li^+ ion in the crystal must be balanced by a Ni^{3+} ion. This can be regarded as

a Ni^{2+} ion together with a trapped hole. The situation is thus analogous to that of Al^{3+} doped into silicon, and the defect can be regarded as an acceptor. The process of creating electronic defects in a crystal in this way is called *valence induction*. As expected, black $Li_xNi_{1-x}O$ is a p-type semiconductor, and as the holes are only weakly bound to the cations, the material possesses a high conductivity.

It is equally possible to impart *n*-type conductivity to an insulator by suitable doping. An example is provided by the reaction of small amounts of gallium oxide, Ga_2O_3, with zinc oxide, ZnO. In this case, Ga^{3+} ions substitute for Zn^{2+} ions in the zinc oxide structure to form $Ga_xZn_{1-x}O$. To maintain charge neutrality, one electron must be added to balance each Ga^{3+} ion in the structure. It is generally believed that these rest on Zn^{2+} ions to generate Zn^+ ions in the crystal. The electrons are not strongly attached to the Zn^{2+} ions, and each Zn^+ ion can be regarded as a donor.

The electronic properties of complex oxides can be changed in the same way, provided that one of the cations present can take part in the valence change. The lanthanoid containing manganites $RMnO_3$ provide many examples (Section 10.7.2). The parent phases adopt the (sometimes distorted) perovskite structure with the lanthanoid R^{3+} cations in the large sites between the Mn^{3+} ($3d^4$) centred metal oxygen octahedra. These materials are usually transparent paramagnetic or antiferromagnetic insulators. When the lanthanoid cations are partly replaced with lower valence cations, typically the alkaline earths, A^{2+} (Mg^{2+}, Ca^{2+}, Sr^{2+}, and Ba^{2+}) or alkali metals A^+ (Na^+, K^+) to give $R_{1-x}A_xMnO_3$, the resulting phases exhibit metal–insulator transitions. Doping with A cations causes the formation of Mn^{4+}: one A^{2+} gives rise to one Mn^{4+} ($La^{3+}{}_{1-x}A^{2+}{}_xMn^{3+}{}_{1-x}Mn^{4+}{}_xO_3$) while one A^+ gives rise to two Mn^{4+} ($La^{3+}{}_{1-x}A^+{}_xMn^{3+}{}_{1-2x}Mn^{4+}{}_{2x}O_3$). Electron transfer can now take place between Mn^{3+} and Mn^{4+} via double exchange (Section 10.5.4) but only if the spins on the two Mn ions are parallel; that is, provided that the metallic state is ferromagnetic.

11.3.4 Transparent Conducting Oxides

A semiconductor can only absorb light if the band gap is smaller than the energy of the incident photons. If high levels of doping can be achieved in a semiconductor with a fairly large band gap, the conductivity may be appreciable while the material remains transparent. This is the situation in *transparent conducting oxides* (TCOs), sometimes referred to as *transparent metals*, which are widely used in computer screens and related applications. The best known of these materials is indium oxide, In_2O_3, doped with between 5 and 15 mol% tin oxide, SnO_2, known as *indium tin oxide* or ITO.

Indium oxide is a lemon-yellow colour in the bulk but when prepared as a thin film appears transparent. Surprisingly, in view of the importance of this material, there are still ongoing attempts to explain its electronic properties. (Indeed, there continues to be disagreement about the true band gap of pure In_2O_3, which is reported to vary from about 2.8 to 3.75 eV.) Incorporation of SnO_2 into In_2O_3 leads to the formation of defects, which in turn leads to the increase in conductivity while retaining the large band gap that makes the oxide transparent. While there is still uncertainty about the nature of these defects and how the doping influences the band structure of the host In_2O_3, the following broad-brush picture describes the state of affairs that is believed to occur. The Sn^{4+} ions are thought to mainly occupy In^{3+}

sites forming $Sn_{In}^{\bullet}$ point defects. The additional oxygen is accommodated as interstitial O^{2-} ions, $O_i^{2'}$.

$$2SnO_2(In_2O_3) \rightarrow 2Sn_{In}^{\bullet} + O_i^{2'}$$
$$\leftrightarrow (2Sn_{In} + O_i)^x$$

There is some evidence to suppose that the tin and oxygen interstitial defects may aggregate into a neutral defect complex $(2Sn_{In}\ O_i)^x$ rather than remain isolated. In either case, the number of interstitial oxygen defects will vary with the ambient oxygen pressure during thin film preparation according to the reversible equation:

$$O_i^{2'} \leftrightarrow 1/2 + O_2 + 2e'$$

This means that at higher ambient oxygen pressures, interstitial oxygen defects are preferred while at lower pressures electrons are produced. This accounts for the fact that highly conducting oxide films are prepared under reducing conditions, that is, at relatively low oxygen partial pressures incorporation of SnO_2 into In_2O_3 leads to the production of electrons. These electrons enter the conduction band to enhance the *n*-type conductivity of the oxide. As the dopant concentration rises, the number of electrons in the oxide increases. At a dopant concentration of about 2×10^{19} cm^{-3}, the material is so heavily doped that the behaviour is that of a degenerate semiconductor and the conductivity approaches that of a metal. The band gap, however, although varying with dopant concentration, remains wide enough for the material to appear transparent in thin film form. The result is a transparent film showing metallic conductivity.

A number of other *n*-type transparent oxide conductors have been found, including tin oxide doped with F and zinc oxide doped with Al_2O_3. Unfortunately, a matching transparent *p*-type oxide conductor has not yet been found, although delafossite structure oxides $CuM^{3+}O_2$ including $CuGaO_2$, $CuInO_2$, and $CuScO_2$, have potential in this respect. Such a material is considered to be important because it would allow for highly desirable transparent electrodes at each face of a light-emitting device instead of only one.

11.4 Conducting Polymers

Ordinary polymers are good insulators, and they are widely used in this capacity as insulating covering on cables and other electrical conductors. The molecular feature that allows polymers to become electronically conducting is the presence of *conjugated* double and single bonds (Figure 11.19). The framework of these molecules is composed of sp^2-hybrid σ-bonds, at 120° to each other. The double bonds are formed by overlap of π orbitals above and below the plane of the carbon chains. They are *not* located between specific pairs of carbon atoms, as drawn, but are spread over all of the molecule as *delocalised π orbitals*, similar to those above and below the planes of the sheets of carbon atoms in graphite A long molecule will have extensive delocalised π orbitals, and it could be anticipated that these would lead to high electronic conductivity along the backbone of the molecule, to give a 'one-dimensional' metal.

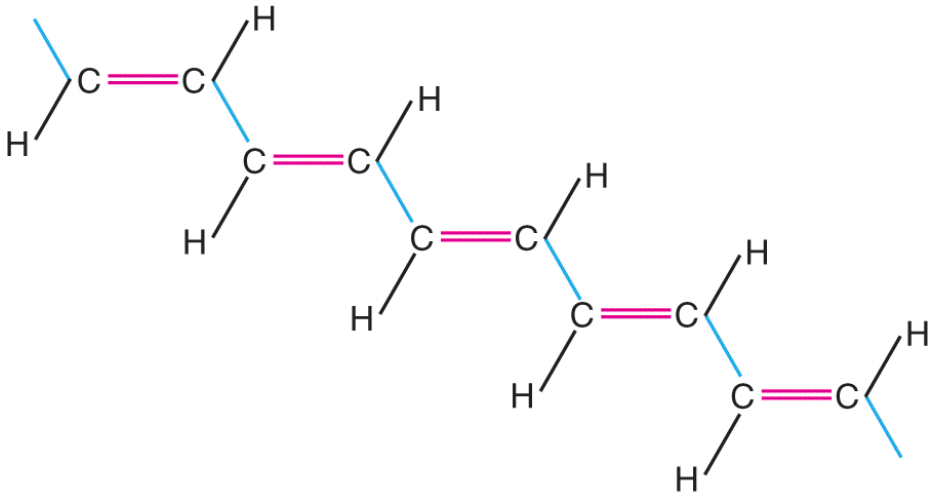


Figure 11.19 Part of a conjugated hydrocarbon molecule (schematic) in which C atoms are linked by alternate single and double bonds: C, carbon; H, hydrogen.

The first conducting polymer to be synthesised was polyacetylene. When polymerised, acetylene (ethyne) forms a silvery flexible film of polyacetylene. Acetylene has a formula C_2H_2. The carbon atoms are linked by a triple bond, consisting of 1 sp-hybrid σ bond and two π bonds (Figure 11.20). The molecules link by opening one of these bonds, to leave a molecular unit containing a double bond. Generally, polymerisation leads to the all-*cis* polymer. At room temperature, this changes to the thermodynamically stable all-*trans* form. Both are poor insulators, with the *trans* form having a conductivity similar to that of silicon (approximately 10^{-3} S m^{-1}), and the *cis* form with a conductivity similar to that of water (approximately 0.1 S cm^{-1}).

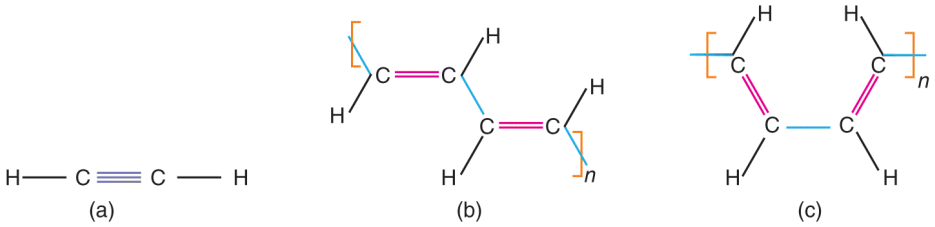


Figure 11.20 The structures of polyacetylenes, schematic: (a) acetylene (ethyne); (b) repeating unit of *trans*-polyacetylene; (c) repeating unit of *cis*-polyacetylene.

This is rather surprising if the band structure of these materials is considered. The sp^2-hybrid orbitals are filled and would give lower energy filled bands that do not contribute to the conductivity. The remaining p_z orbitals on each carbon contain one electron. A chain of CH units, each with one unpaired outer electron is analogous to a chain of alkali metal atoms such as lithium. If the separation of the carbon atoms is rather large, each electron would be completely localised on each carbon and the polymer would be an insulator, possibly ferromagnetic or antiferromagnetic, with an empty conduction band (Figure 11.21a). On the other hand, if the inter-carbon distance is small, each electron is able to delocalise over all of the carbon atoms in a π-band. Each atom would contribute just one electron, which would produce a half-full conduction band and the material should show metallic conductivity (Figure 11.21b).

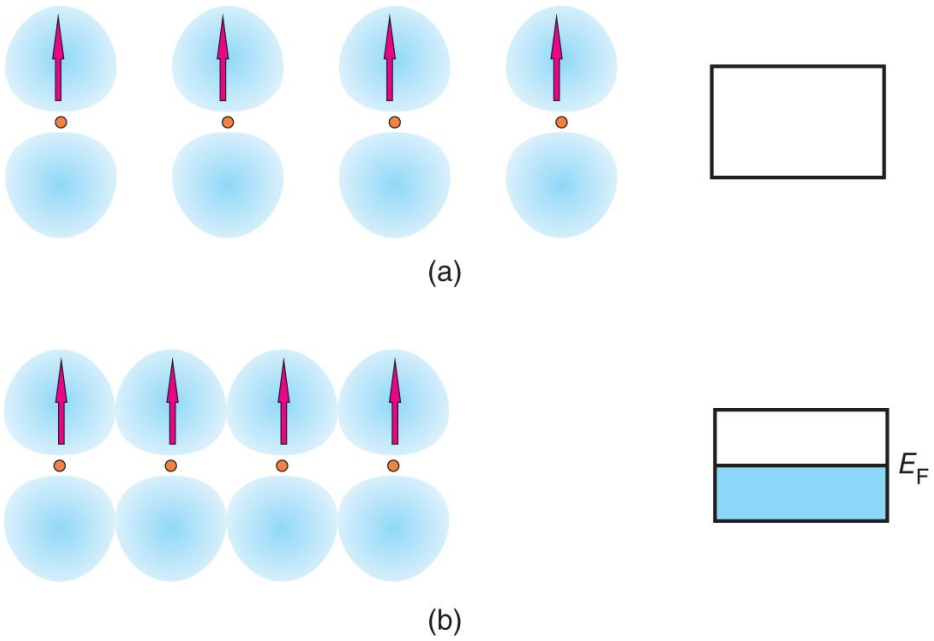


Figure 11.21 Conjugated polymers: (a) a chain of isolated half-occupied p orbitals leads to an insulating phase; (b) a chain of overlapping half-occupied p orbitals leads to a half-filled conduction band.

The real situation corresponds to neither of these alternatives. A linear chain of close equispaced atoms, which should give rise to a metallic state (Figure 11.22a), is found to be energetically unstable. Instead, the spacing between the atoms adjusts so that an energy gap opens at the Fermi level. The variation in spacing is called a *Peierls distortion*. As a consequence, the half-filled band is transformed into two, a filled band and an empty band, and as a result the solid becomes a semiconductor (Figure 11.22b,c). In agreement with this, C□C bonds along the chain in *trans*- and *cis*-polyacetylene alternate between two lengths. The band gap is approximately 1.8 eV (= 2.88×10^{-19} J) in the *trans*- form and about 2 eV (= 3.20×10^{-19} J) in the *cis*- form of the polymer, accounting for the observed electronic properties.

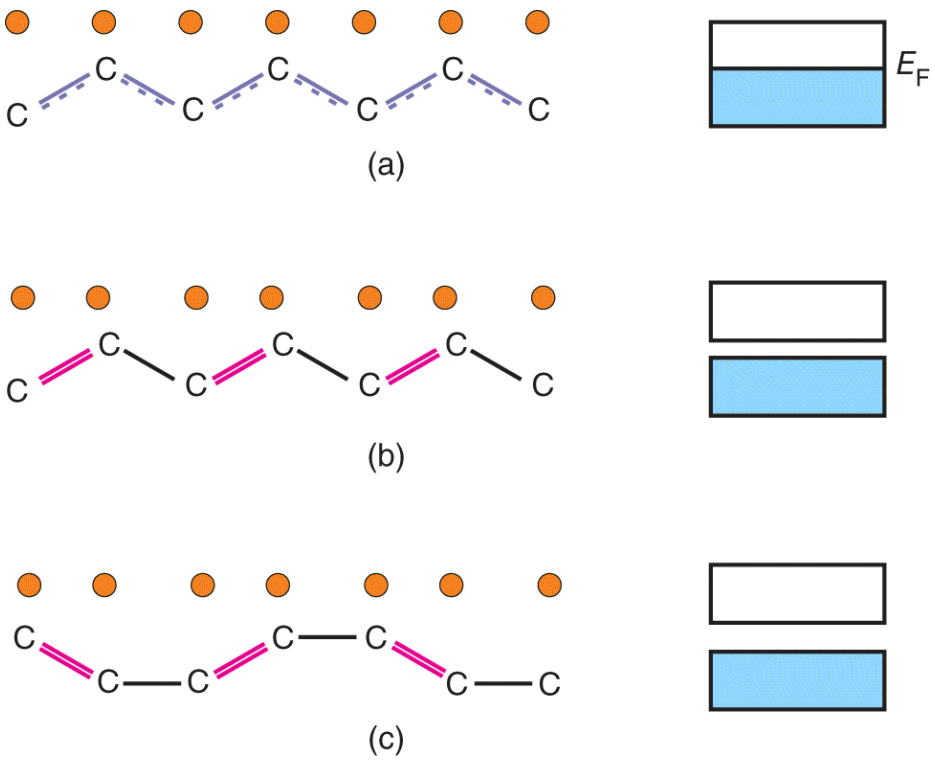


Figure 11.22 Distortion of a polymer chain: (a) delocalized orbitals lead to equally spaced atoms and a half-filled energy band; (b), Peierls distortion in *trans*-polyacetylene and (c) cis-polyacetylene, leads to energy band splitting.

Polyacetylene can, however, be transformed into a metallic conductor by doping. This involves oxidation or reduction of the polymer. Electron acceptors such as halogens (chlorine, iodine, etc.) oxidise the polymer. In this process, electrons are taken from the filled lower band and used to form halide ions, leaving holes, which result in a p-type material. A typical reaction is:

$$[CH]_n + (3x/2)\, I_2 \rightarrow [(CH)_n{}^{x+}\, x(I_3{}^-)]$$

The iodine enters the polymer between the molecular chains and the conductivity varies from that typical of a poor semiconductor to that of a metal. Doping with alkali metals (lithium, sodium, etc.) reduces the polymer and has the same effect. In this process, the alkali metal donates electrons to the empty band, forming an alkali metal ion and transforming the polymer into an n-type material. A typical reaction can be written as:

$$[CH]_n + x\, Li \rightarrow [(CH)_n{}^{x-}\, xLi^+]$$

The conductivities of the doped polymers, of the order of 10^8 S m^{-1}, are similar to that of copper or silver and vary with dopant concentration (Figure 11.23). Notice that much greater concentrations of dopant are needed (atomic %) than those used in the traditional semiconductors such as silicon (parts per million).

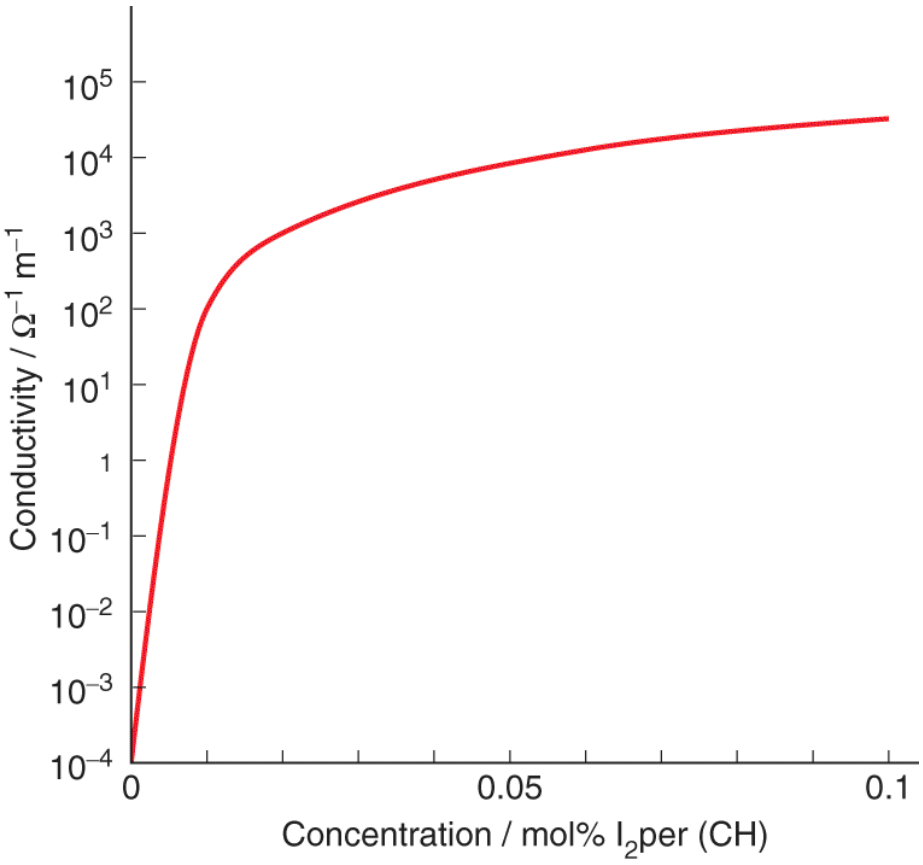


Figure 11.23 Variation of the conductivity of polyacetylene with iodine dopant concentration.

The conductivity and other electrical properties of both the semiconducting and the metallic region are not identical to the semiconductors and metals already discussed. For example, the conductivity in the metallic region of iodine doped polyacetylene decreases as the temperature decreases (Figure 11.24), whereas in an ordinary metal the reverse is true. Moreover, the resistivity seems to show Arrhenius-like temperature dependence:

$$\sigma = \sigma_0 \exp(-E_a/k_B T)$$

where E_a is the activation energy, k_B is Boltzmann's constant, T the temperature, and σ_0 is a constant.

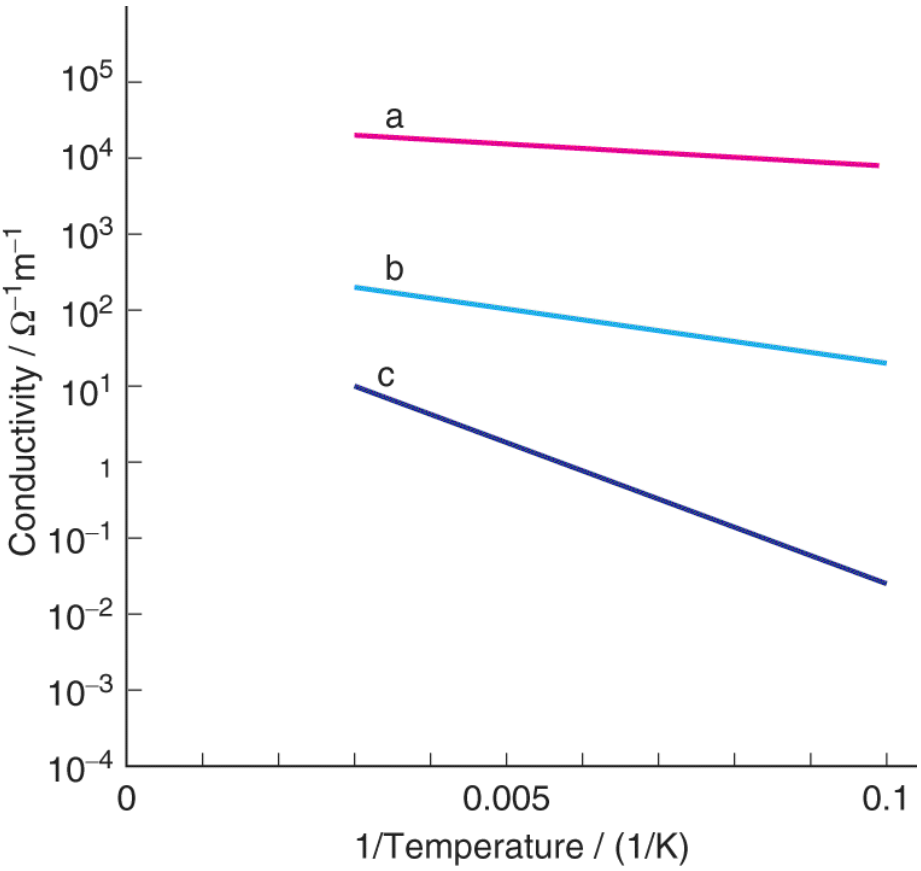


Figure 11.24 Variation of the conductivity of polyacetylene doped with iodine vs. reciprocal temperature. Dopant concentrations approximately (a) 0.15 mol%; (b) 0.01 mol%; (c) 0.005 mol% I_2 per CH unit.

This Arrhenius-like behaviour suggests that the dopant does not simply add mobile carriers to a conduction or valence band, as with silicon. The explanation of this feature is bound up with the structure of the polymer. Polyacetylene is composed of ordered, fairly crystalline regions linked by disordered regions in which the polymer chains are twisted and coiled. The dopants modify the structure of the polymer chains. Acetylene polymerises preferentially to form the *cis*-isomer, and this gradually transforms into the stable *trans*- form. The transformation process is initiated at random, rather like the nucleation of a crystal in a liquid. In the same way that crystallites grow together to form a polycrystalline mass, when two regions of polymer chain that have transformed from *cis*- to *trans*- meet, there is frequently a mismatch (Figure 11.25). This mismatch results in the formation of a *soliton*,[1] which can be thought of as an interruption in the orderly pattern of conjugated double bonds, extending over several adjacent carbon atoms. In some cases, there will be a π electron missing, which will give the soliton a positive charge. In other cases, the soliton can attract an unbound π electron to give a neutral soliton, or two π electrons to give a negative soliton. Doping increases the soliton density and charge along the polymer chains. This gives rise to two effects. Firstly, at the soliton, the alternating bond lengths required for the Peierls distortion are suppressed. Ultimately, when enough solitons are present, the Peierls distortion fails and a half-filled band is reinstated, recreating the metallic state. Secondly, the solitons give rise to impurity energy levels in the band gap. When enough of these are present, they merge and bridge the band gap, again making high conductivity possible.

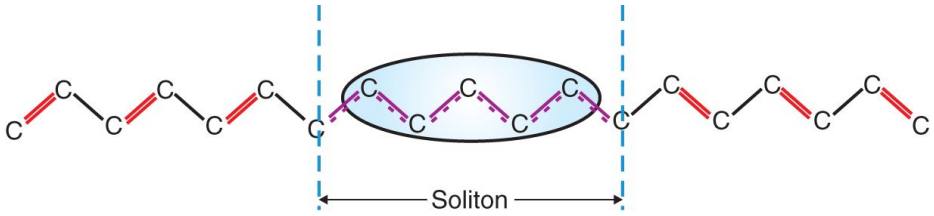


Figure 11.25 A soliton in *trans*-polyacetylene, schematic. The shaded ellipse may represent a positively charged soliton (containing a hole), a neutral soliton (containing a single electron) or a negative soliton (containing two electrons).

The conduction process is not simply the spread of electron waves throughout the solid, as in a crystalline metal. Instead, the soliton jumps from one location to a neighbouring one under the influence of an electric field, maintaining its shape all the time, similar to a soliton wave in water. This progress resembles ionic conductivity and the motion requires activation energy, resulting in the Arrhenius-like behaviour.

The conductivity of the polymer depends upon the microstructure of the solid. Stretching the polymer sheet at moderate temperatures increases the alignment of the chains and leads to significant improvement in properties along the chain direction. In this form polyacetylene is widely used as an electrode material in lightweight batteries.

11.5 Superconductivity

11.5.1 Superconductors

In 1911 H. Kamerlingh Onnes found that mercury lost all electrical resistance when cooled to the temperature of liquid helium (4.2 K), and was described as having reached the *superconducting* state. The feature is not restricted to mercury. Generally, the resistance of a metallic element gradually falls as the temperature falls, until there is a catastrophic drop in resistance to zero at the onset of a superconducting state. The temperature at which materials become superconducting is called the superconducting *transition temperature, T_c*. Most metallic elements have a transition temperature below 10 K. For many years the highest value of T_c recorded was close to 18 K, and this seemed to be a genuine limit, but new techniques of preparation pushed this up to 23.2 K in the alloy Nb_3Ge, by 1970.

Although other unusual superconductors were known in this period, metallic alloys were regarded as the likeliest contenders for showing superconductivity at temperatures appreciably higher than 23.2 K until the ceramic cuprate superconductors were discovered, in 1986. Although the transition temperature of the first ceramic compound recognised as a superconductor, $La_{1.85}Ba_{0.15}CuO_4$, was only about 30 K, the very existence of superconductivity in a ceramic lead to an explosion of research that produced large numbers of new oxide superconductors. The current record for T_c at ambient pressure is 135 K and is held by $Hg_{0.8}Tl_{0.2}Ba_2Ca_2Cu_3O_{8.33}$ (Figure 11.26).

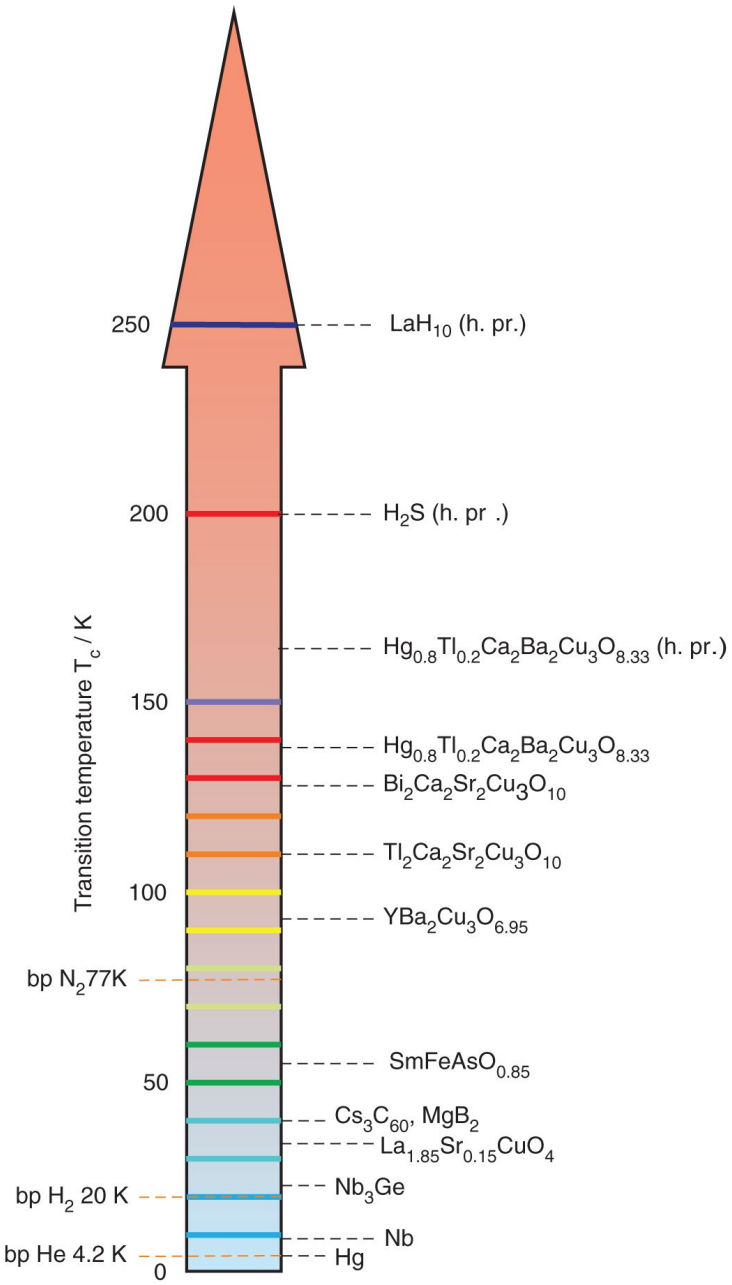


Figure 11.26 Superconducting transition temperatures.

Recently the variety of materials showing superconductivity have widened considerably. In 2001, the apparently simple ionic compound MgB_2 was found to become superconducting with a transition temperature of 39 K. In 2008, the superconducting iron pnictides (compounds of iron with the nitrogen group: N, P, As, Sb) and iron chalcogenides (compounds of iron with the oxygen group: O, S, Se, Te) were discovered. Since then, a dozen or more of these iron-containing superconducting compounds, typified by complex chemical formulae such as $NdFeAsO_{0.89}F_{0.11}$ ($T_c \approx 52$ K) and $Sr_{0.5}Sm_{0.5}FeAsF$ ($T_c \approx 56$ K) have been characterised. The situation became more complicated in 2015 by the announcement of superconductivity in H_2S subjected to high pressures, giving a T_c of 200 K under a pressure of 90 GPa, and the discovery of superconductivity in LaH_{10} in 2019, with a T_c of 250 K under

similarly high pressures. A superconducting transition at room temperature – a long sought-after goal – has been achieved, with a superconducting phase formed from C, H and S, with a structure yet to be determined (2020), at a pressure of 217 GPa.

The first 'classical' description of superconductivity, in which electrons are paired via the intervention of phonons (lattice vibrations), the *Bardeen–Cooper–Schrieffer* (BCS) theory (Section 11.5.3) explains the overall behaviour of superconductivity in elemental metals and alloys (Section 11.5.2). The cuprate and other newly discovered superconductors (Section 11.5.5) are called *unconventional superconductors*. These cannot be explained by the BCS theory and there is as yet no overall theoretical understanding of the way these materials behave.

11.5.2 The Effect of Magnetic Fields and Current

When a superconductor is cooled in a magnetic field, it *expels* the magnetic induction, B, in its interior. This is called the *Meissner effect*. Ideally this expulsion is complete so that a superconductor behaves as perfectly diamagnetic (Figure 11.27). In effect, in the superconducting state a surface current is produced that is sufficient to generate an internal magnetic field that exactly cancels the magnetic induction. This current also distorts the external magnetic field so that it does not penetrate the superconductor, which is thus screened. The surface current is confined to a thin layer, called the *penetration depth*, varying between 10 and 100 nm.

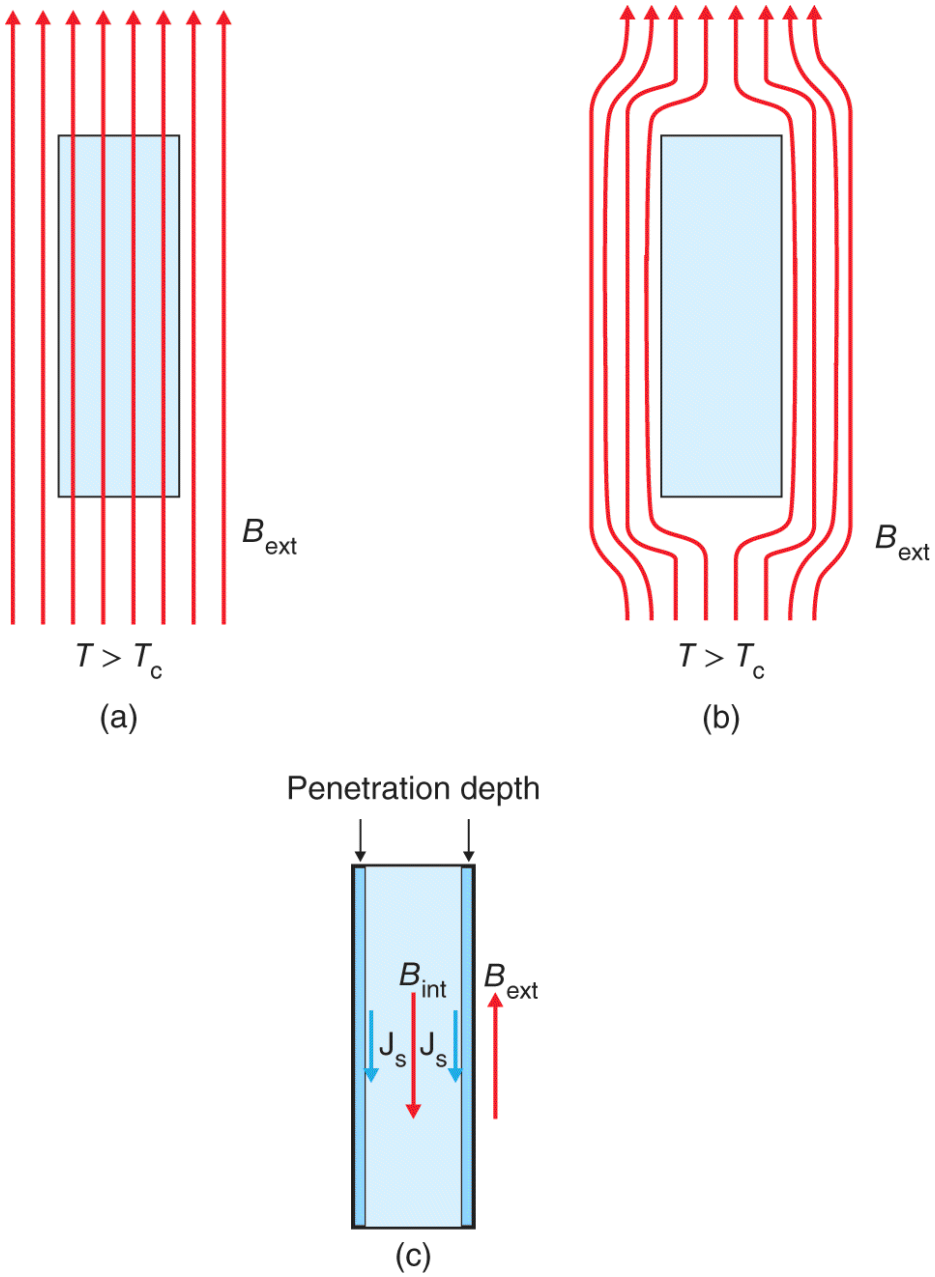


Figure 11.27 The Meissner effect: (a) normal metal or superconducting metal above T_c; (b) superconducting metal below T_c; (c) surface currents induced in a superconductor.

The exclusion of the magnetic induction costs energy, and when that cost outweighs the energy gained by the formation of the superconducting state, the material reverts to normal behaviour. This occurs at a *critical field*, H_c, in *Type I superconductors*. The value of the critical field, H_c, is temperature dependent so that a phase boundary can be mapped out in $H - T$ phase space (Figure 11.28a). The relationship between H_c and the critical temperature, T_c, is given by an approximate equation:

$$H_c \sim H_c(0)\,[1 - (T/T_c)^2] \tag{11.5}$$

where $H_c(0)$ is the critical magnetic field[2] at 0 K, and T is the absolute temperature.

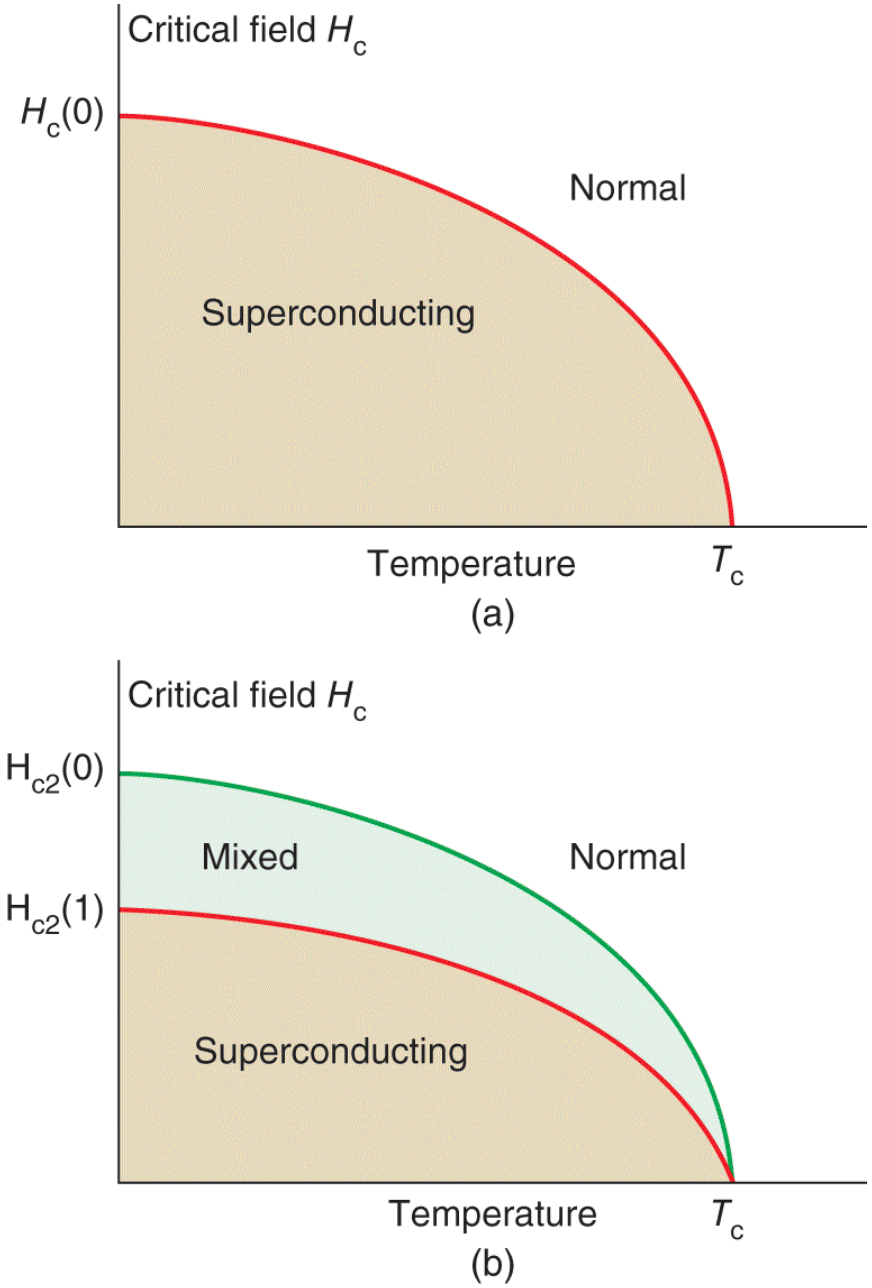


Figure 11.28 The variation of superconductivity with external magnetic field: (a) Type I superconductor; (b) Type II superconductor.

In many superconductors the transition between the superconducting and normal state is not sharp. When the external magnetic field reaches some *lower critical magnetic field* value, H_{c1}, filaments of magnetic flux start to penetrate the material to produce threads that are no longer superconducting (Figure 11.29). The small regions of normal material are isolated from the superconducting matrix by surface currents, called *vortices* or *fluxoids*, which act to repel each other, so that the flux lines form an ordered structure called a *fluxon lattice*, *flux lattice*, or *vortex lattice*. At the core of each flux vortex the material is effectively normal, but is surrounded by a region of superconductor. As the magnetic field increases, the amount of

normal phase also increases relative to the superconducting part. When the magnetic field reaches the *upper critical field,* H_{c2} the flux lines are so close together that no superconducting material exists between them and the solid becomes normal. This behaviour characterises *Type II superconductors* (Figure 11.28b). The temperature dependence given in Eq. (11.5) also holds for type II superconductors if H_c is replaced by $H_{c2}(0)$.

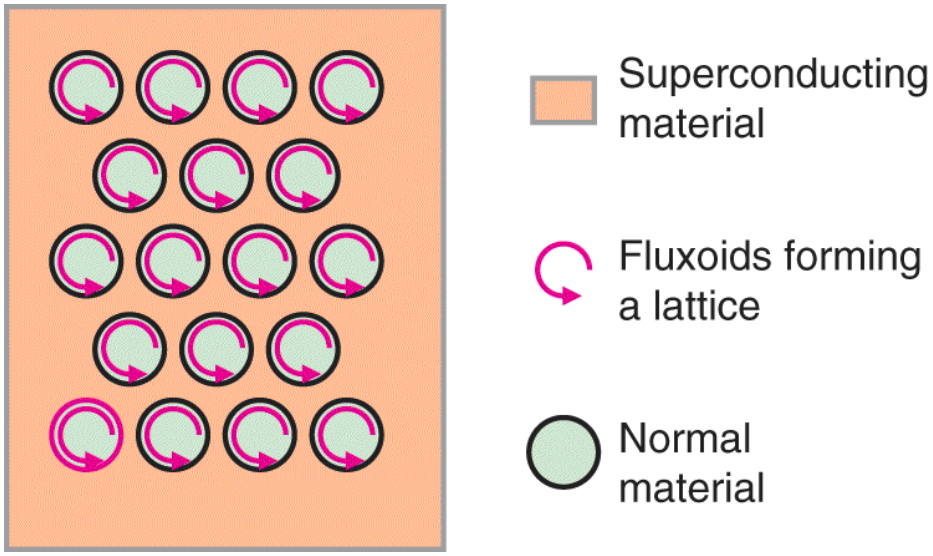


Figure 11.29 A Type II superconductor in the mixed state.

The magnetic flux enclosed by a loop of superconductor must be quantised in multiples of $h/2e$, where h is Planck's constant and e is the electronic charge. The unit of flux is called the *flux quantum*, or *fluxon*, Φ_0, with a value of 2.07×10^{-15} Wb. The flux enclosed in any circuit must then be $nh/2e$, with n taking integral values.

Although a superconducting solid exhibits no resistivity, at a certain current, the *critical current*, J_c, a superconductor reverts to a normal resistive state. The critical current is a function of the temperature and the external magnetic field, and hence the superconducting state of a solid can be mapped out as a volume with respect to current, field and temperature axes (Figure 11.30). As either temperature or magnetic field increases, the critical current decreases. The critical current is also greatly influenced by the microstructure of the solid, and careful processing is needed to obtain superconducting samples with high values of the critical current. For working devices, a value of J_c of about 10^6 A cm^{-2} or greater needs to be achieved.

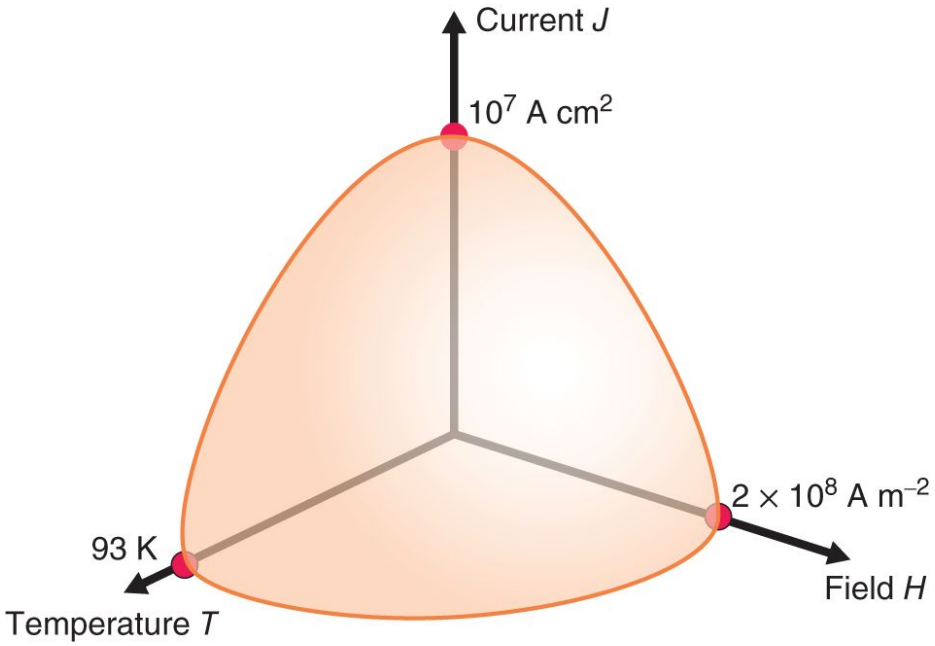


Figure 11.30 The temperature–current–magnetic field phase space delimiting superconductivity. The values of the critical field, temperature, and current are for $YBa_2Cu_3O_{7-x}$.

11.5.3 The BCS Theory of Superconductivity

In a normal metal, resistivity is the result of interactions of the current carrying electrons with the crystal structure. This clearly does not happen in the superconducting state and a completely different theory of electrical conduction is needed to account for the properties that superconductors possess. Type I superconductors are well explained by the BCS theory, set out in 1957, in which the superconducting state is characterised by having the mobile electrons coupled in pairs with opposite spins, called *Cooper pairs*. At normal temperatures, electrons strongly repel one another. As the temperature falls and the lattice vibrations diminish, a weak attractive force between pairs of electrons becomes significant. In Type I superconductors the 'glue' between the Cooper pairs are phonons (quantised lattice vibrations).

The coupling can be envisaged in the following way. As an electron passes through a crystal, it interacts with the surrounding positively charged atomic cores, weakly attracting them (Figure 11.31). This leads to a small lattice distortion and a slightly enhanced region of positive charge in the neighbourhood of the passing electron that has been likened to the wake from a moving boat. Naturally this weak attraction is swamped at high temperatures by thermal vibrations. However, at low temperatures another electron is able to feel the influence of the distortion and is weakly attracted to the slightly positive wake generated by the first electron, and is carried along with it. The two electrons are thus linked as a Cooper pair. Formally, when one electron in a Cooper pair passes through the crystal, the atom cores are attracted slightly and then spring back as the electron passes. This causes a phonon to be emitted that is picked up by the other electron in the Cooper pair. At the same time, this electron is also causing phonons to be emitted, which are picked up by the first electron. This phonon exchange acts as the weak glue between the electron pairs. This coupling is easily destroyed even at low temperatures, and Cooper pairs are constantly forming and breaking apart. The pairs of electrons behave quite differently than single electrons. For example, they share the same wave function, and are able to pass through the crystal unimpeded.

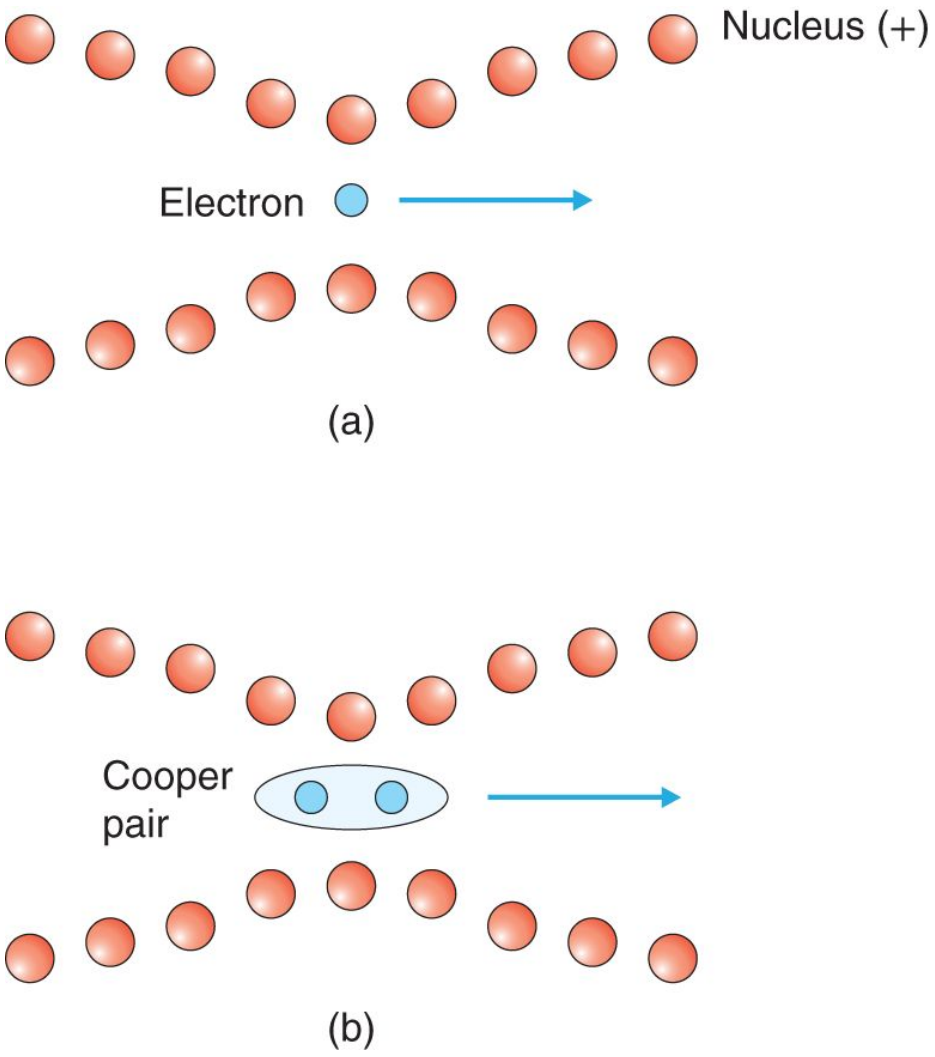


Figure 11.31 Cooper pairs: (a) a single electron passing through a superconductor attracts the positively charged nuclei slightly; (b) at low temperatures another electron can be attracted into this region to form a Cooper pair.

In the newly discovered 'unconventional' high temperature superconductors described in the following text, it appears that the electrons are also paired. The BCS theory is not able to account for the much stronger coupling that must occur in these solids, and no satisfactory overall theory has yet been suggested.

11.5.4 Josephson Junctions

Cooper pairs can tunnel through a thin layer of insulator, called a *weak link*, separating two superconducting regions, without destroying the coupling between them or the superconductivity of the adjoining phases. In such cases a direct current (dc) flows across the insulating layer without the application of an accompanying voltage. This is called the *dc Josephson effect*. In essence the insulator behaves as if it was superconductor. The Josephson effect only persists for a certain range of currents, and eventually, above a critical current i_c, a voltage develops across the junction.

When two Josephson junctions are connected in parallel (Figure 11.32a), the maximum current, i_c, that can flow across the device is function of the magnetic flux enclosed in the

loop. As the magnetic flux penetrating the loop varies, the value of i_{cmax} varies from a maximum at zero flux or an integral number of flux quanta to minimum for a half-integer number of flux quanta. The relationship is given by:

$$i_{cmax} = 2I_J[\cos(\pi \Phi/\Phi_0)]$$

where I_J is a constant depending upon the junction geometry, Φ is the enclosed magnetic flux, and Φ_0 is the flux quantum.

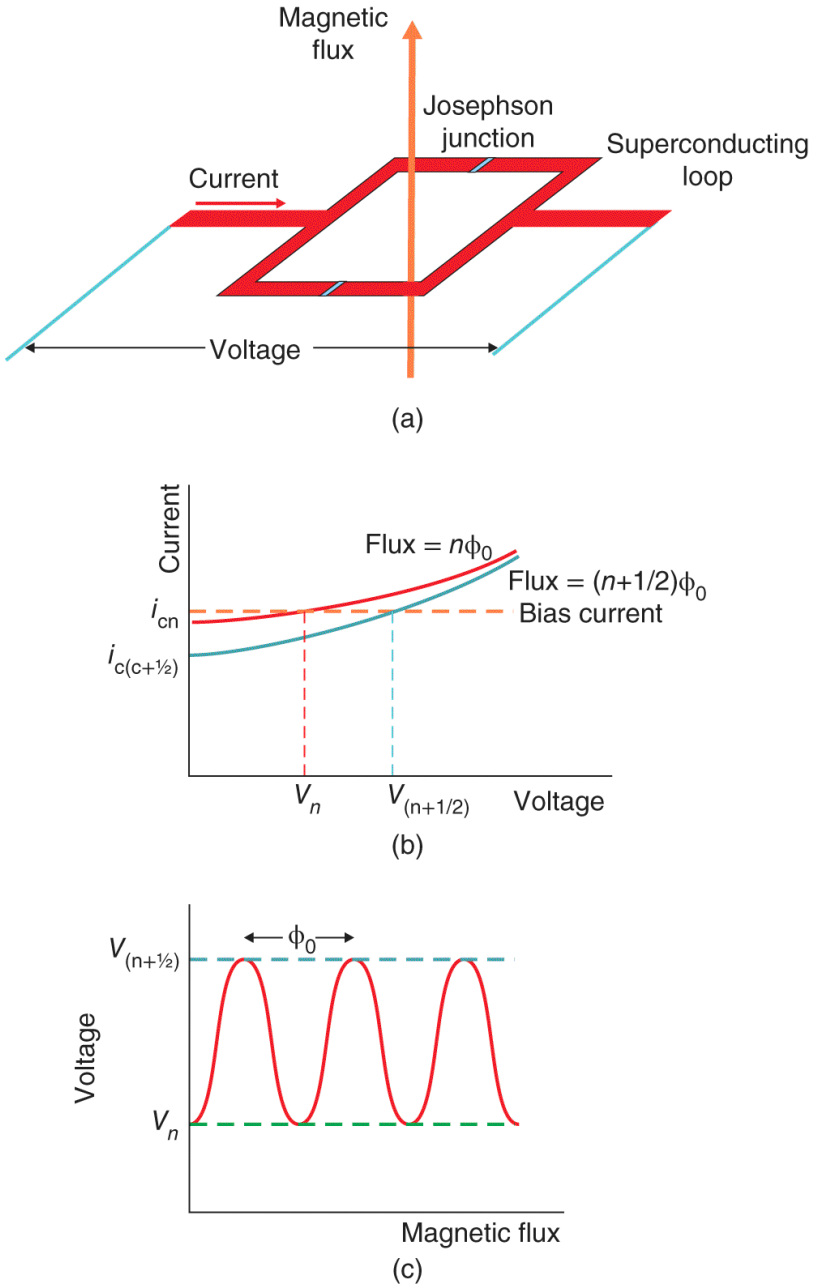


Figure 11.32 A dc SQUID: (a) the circuit consists of a loop of superconductor containing two Josephson junctions in parallel; (b) with a bias current above the critical current the voltage recorded depends upon the number of flux quanta that penetrate the loop; (c) in a varying magnetic field, the voltage cycles sinusoidally.

A *dc SQUID* (*superconducting quantum interference device*) is a device based upon this effect, used for the measurement of microscopic magnetic fields. In operation, a bias current is used, which is just above that needed to produce a voltage across the circuit. The critical current i_{cn} when a whole number of flux quanta penetrate the loop is higher than when an odd half number of flux quanta penetrate the loop, $i_{c(n+½)}$. Because of this, the voltage recorded, V_n, $V_{n+½}$, or an intermediate value, depends upon the magnetic flux penetrating the loop (Figure 11.32b), and as the magnetic flux changes, the voltage varies in a sinusoidal fashion with a period equal to the flux quantum (Figure 11.32c). An SQUID is thus a magnetic flux to voltage converter. Typically, the gain is about 1 volt per flux quantum. As fractions of a volt are easily measured, an SQUID has a sensitivity of approximately 10^{-4} to 10^{-6} of a flux quantum. This is sensitive enough to measure the magnetic fields produced by changes in the electrical activity of the brain.

11.5.5 Cuprate High Temperature Superconductors

Cuprate superconductors are non-stoichiometric copper oxides that maintain the superconducting state to relatively high temperatures, often above that of liquid nitrogen (Table 11.2). Crystallographically the phases are all related to the perovskite structure type, ABO_3 (Figure 3.11), where A is a large cation and B is Cu. The structures of the superconductors are built up of slices of perovskite structure linked by slabs with structures (in the main) of the halite (NaCl) or fluorite (CaF_2) type. The copper valence in most compounds lies between the formal values of Cu^{2+} and Cu^{3+}. The appearance of superconductivity, and the transition temperature, is closely connected with the composition of these solids, which all exhibit considerable degrees oxygen composition variation.

Table 11.2 Some cuprate superconducting oxides.

Compound[a]	T_c **(K)**
$La_{1.84}Sr_{0.16}CuO_4$	38
$Nd_{2-x}Ce_xCuO_4$	20
$YBa_2Cu_3O_{6.95}$	93
Bi_2O_2/Tl_2O_2 layers	
$Bi_2Sr_2CuO_6$	10
$Bi_2Sr_2CaCu_2O_8$	92
$Bi_2Sr_2\ Ca_2Cu_3O_{10}$	110
$Tl_2Ba_2CuO_6$	92
$Tl_2Ba_2CaCu_2O_8$	119
$Tl_2Ba_2\ Ca_2Cu_3O_{10}$	128
$Tl_2Ba_2\ Ca_3Cu_4O_{12}$	119
TlO/HgO layers	
$TlBa_2CuO_5$	10
$TlBa_2CaCu_2O_7$	103
$TlBa_2\ Ca_2Cu_3O_9$	110
$HgBa_2CuO_4$	94
$HgBa_2CaCu_2O_6$	127
$HgBa_2\ Ca_2Cu_3O_8$	133
$Hg_{0.8}Tl_{0.2}Ba_2\ Ca_2Cu_3O_{8.33}$	138
$HgBa_2Ca_3Cu_4O_{10}$	126

[a] The formulae are representative and do not always show the exact oxygen stoichiometry for the optimum T_c values given.

11.5.5.1 Lanthanum Cuprate, La_2CuO_4

The phase La_2CuO_4 contains trivalent La and divalent Cu, and adopts a slightly distorted version of the K_2NiF_4 structure, in which the CuO_6 octahedra are lengthened along the **c**-axis, compared to the regular octahedra in K_2NiF_4. The structure can also be thought of as LaO layers with the halite (NaCl) structure interleaved with CuO_2 layers and stacked in the sequence ...CuO_2, LaO, LaO, CuO_2 ... (Figures 11.33–11.35a). In the superconductor literature this structure is often labelled T or T/O. The dimensions of the room temperature orthorhombic unit cell are a = 0.535 nm, b = 0.540 nm, c = 1.314 nm.

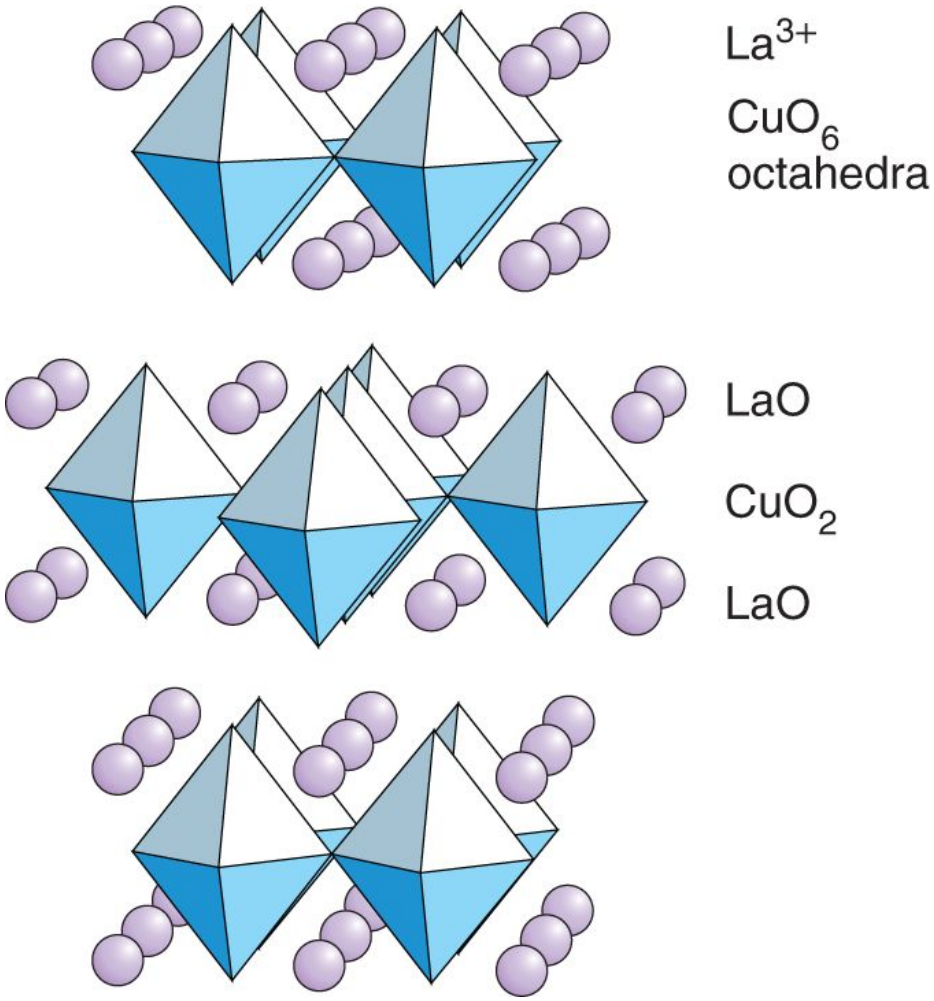


Figure 11.33 The crystal structure of La_2CuO_4. The **c**-axis is vertical.

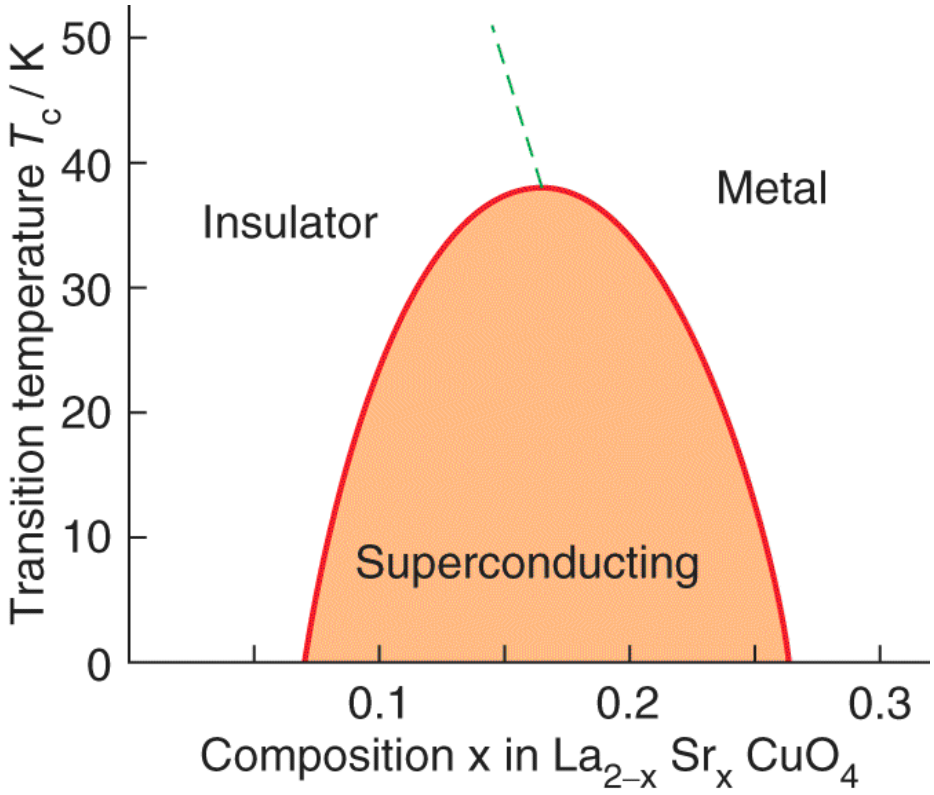


Figure 11.34 The phase region supporting superconductivity for $La_{2-x}Sr_xCuO_4$.

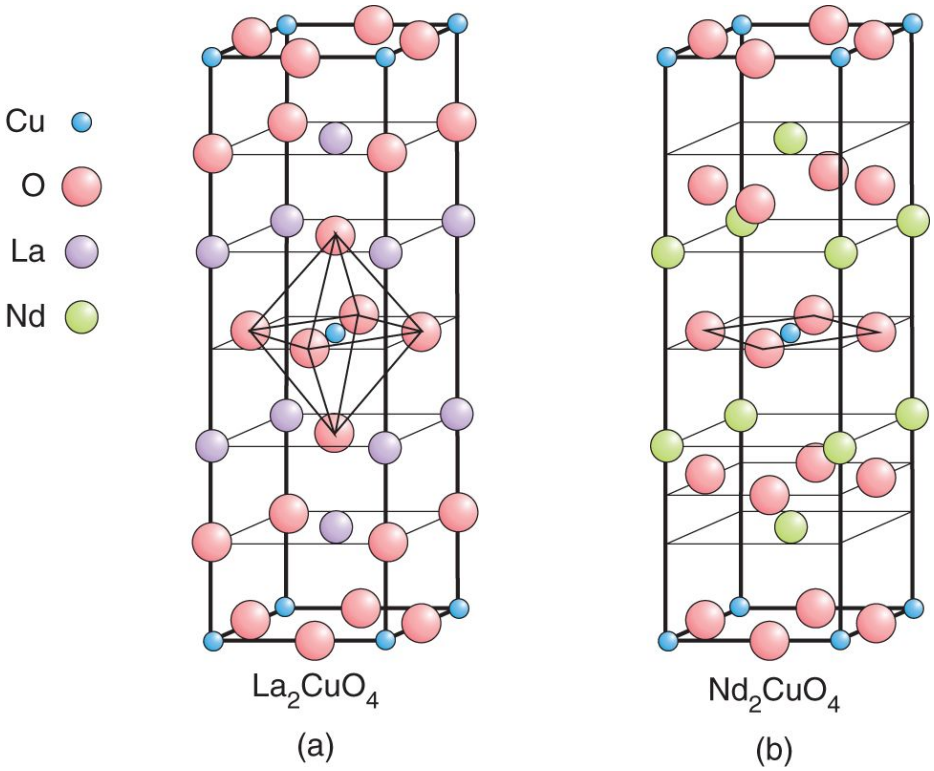


Figure 11.35 Comparison of the structures of (a) $LaCuO_4$ and (b) Nd_2CuO_4.

When prepared in air by heating the oxides CuO and La_2O_3, the compound is stoichiometric with oxygen content 4.00. Electronically, the material is an antiferromagnetic insulator. The substance can be transformed into a superconductor by acceptor doping via replacement of some of the La^{3+} cations with the alkaline earth cations A^{2+} (Ba^{2+}, Sr^{2+}, or Ca^{2+}). Charge neutrality can be maintained in one of two ways in this compound. If valence induction occurs, one Cu^{3+} forms for each A^{2+} substituent, to give a formula $La_{2-x}A_xCu^{2+}{}_{1-x}Cu^{3+}{}_xO_4$. It is also possible to generate one oxygen vacancy for every two A^{2+} added, to give $La_{2-x}A_xCu^{2+}O_{4-x/2}$. The balance between these two alternatives is very delicately poised and leads to a surprising situation.

The best understood phase is $La_{2-x}Sr_xCu^{2+}O_4$. Initially the phase is an insulator. At low dopant concentrations, the Cu^{3+} option is preferred. Because Cu^{3+} can be looked on as Cu^{2+} together with a trapped hole, it is often convenient to refer to the substitution as hole doping. The incorporation of one Sr^{2+} ion generates one hole in the structure. These are regarded as forming in the LaO layers before migrating into the CuO sheets to locate on some of the Cu^{2+} ions to form Cu^{3+} defects. Initially doping causes insulating La_2CuO_4 to become metallic and then, as the Sr^{2+} concentration passes approximately 0.07, the metallic phase becomes a *hole superconductor* when the temperature is lowered sufficiently. Continued doping causes the Cu^{3+} population to rise, peaking when x is approximately 0.16, to give a maximum superconducting transition temperature of 38 K at a composition of approximately $La_{1.84}Sr_{0.16}CuO_4$ (Figure 11.34).

Continued substitution forces the compensation mechanism to change to vacancy generation and the oxygen content of the parent phase falls below 4.0. One oxygen vacancy is generated for every two Sr^{2+} dopant ions, to give a formula $La_{2-x}Sr_xCu^{2+}O_{4-x/2}$. The number of Cu^{3+} ions decreases as oxygen vacancies form, and when the concentration of Sr^{2+} reaches approximately 0.27 all of the compensation is via vacancies and the material is no longer a superconductor. Superconductivity occurs within the phase region known as the *superconducting dome*. Outside of this phase region the electronic properties are complex and are approximated here as insulator and metal. High oxygen pressures modify the position of the dome and preparations made under these conditions allow materials to remain superconducting up to a Sr^{2+} concentration of approximately 0.32.

11.5.5.2 Neodymium Cuprate, Nd_2CuO_4

The structure of Nd_2CuO_4 is very similar to La_2CuO_4, the principle difference lying in the disposition of the oxygen atoms, as the cations in the two structures are in almost identical positions (Figure 11.35b). However, in Nd_2CuO_4 the Nd^{3+} ions are in the centres of oxygen cubes and so this region of the structure can be likened to slabs of the fluorite (CaF_2) type and the structure may be described as an intergrowth of the perovskite and fluorite structures. The Cu atoms lie at the centres of square co-ordination groups between these fluorite slabs rather than octahedra, as in La_2CuO_4. In the superconductor literature this structure is called *T′*.

The structural and chemical similarity of La_2CuO_4 and Nd_2CuO_4 suggests that Nd_2CuO_4 should become metallic following acceptor doping, but hole superconductivity does not arise in this phase when Nd^{3+} is substituted by Ca^{2+}, Sr^{2+}, or Ba^{2+}. However, Nd_2CuO_4 can be made metallic and superconducting by donor doping involving the substitution of the lanthanoid by a higher valence cation such as Ce^{4+} to form, for example, $Nd_{1-x}Ce_xCuO_4$.

Chemically, this represents the generation of one Cu^{+} in place of Cu^{2+} for every donor Ce^{4+} to give a formula $Nd_{1-x}Ce_xCu^{2+}{}_{1-x}Cu_x{}^{+}O_4$. The Cu^{+} ion is equivalent to a Cu^{2+} ion together with a trapped electron and the compounds are *electron superconductors*. Initial doping turns the insulator into a metal and then creates a superconducting state which occurs under a superconducting dome with x taking values of approximately 0.12 to 0.18. The maximum value of the superconducting transition temperature, T_c, is approximately 24 K, reached in the compound $Nd_{1.85}Ce_{0.15}CuO_4$.

11.5.5.3 Yttrium Barium Copper Oxide, $YBa_2Cu_3O_7$

The compound $YBa_2Cu_3O_7$ has been widely studied because it was the first superconductor discovered with a T_c above the boiling point of liquid nitrogen. The crystal structure of $YBa_2Cu_3O_7$ consists of three perovskite-like ABO_3 unit cells stacked one on top of the other (Figure 11.36a). The middle perovskite unit contains Y as the A atom and Cu as the B atom. The cells above and below this contain Ba as the A atom and Cu as the B atom, to give a metal formula of YBa_2Cu_3 as one would expect for a tripled perovskite cell, $A_3B_3O_9$. However, instead of the nine O atoms expected there are seven, arranged in such a way as to give the Cu atoms square pyramidal and square planar co-ordination, rather than octahedral as in the normal perovskites. The dimensions of the orthorhombic unit cell are a = 0.381 nm, b = 0.388 nm, c = 1.165 nm. If the ions are allocated the formal charges of Y^{3+}, Ba^{2+}, and O^{2-}, the Cu must take an average charge of 2.33, which can be considered to arise from the nominal presence of two Cu^{2+} ions and one Cu^{3+} ion per unit cell.

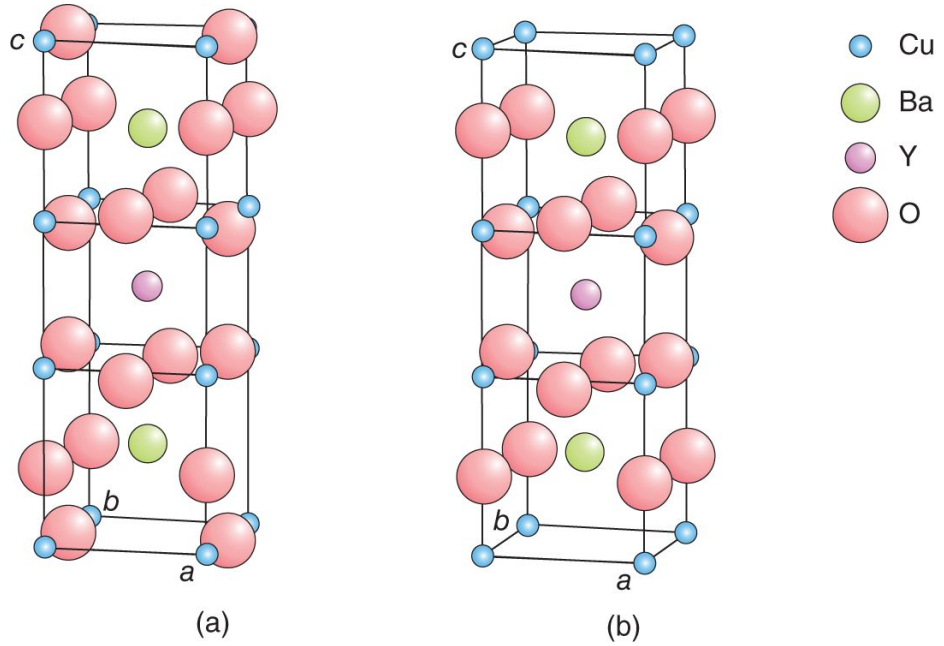


Figure 11.36 The crystal structures of: (a) $YBa_2Cu_3O_7$; (b) $YBa_2Cu_3O_6$.

$YBa_2Cu_3O_7$ is ceramic insulator, but superconductivity appears when a small amount of oxygen is lost to yield a hole superconductor. The maximum value of T_c, close to 93 K, is found at the composition $YBa_2Cu_3O_{6.95}$. As more oxygen is removed the value of T_c falls to a plateau at approximately 60 K when the composition lies between the approximate limits of $YBa_2Cu_3O_{6.7}$ and $YBa_2Cu_3O_{6.5}$ (Figure 11.37). The form of this curve suggests that there may

be two superconducting phases forming, one centred near to $YBa_2Cu_3O_{6.95}$ and the other near to $YBa_2Cu_3O_{6.50}$. Continued oxygen removal rapidly leads to a loss of superconductivity and the material again behaves as an insulating ceramic beyond $YBa_2Cu_3O_{6.35}$. Oxygen loss continues down to a composition of $YBa_2Cu_3O_{6.0}$ (Figure 11.36b).

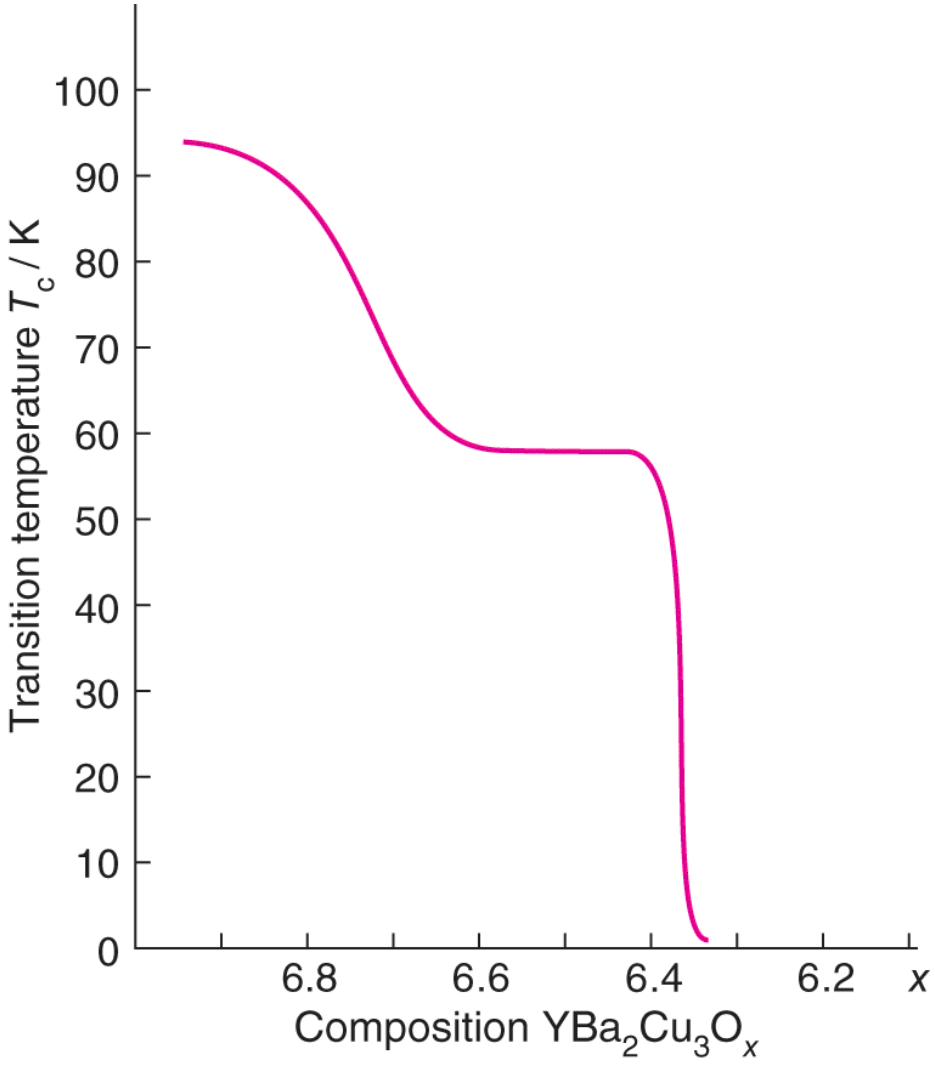


Figure 11.37 The variation of the superconducting transition temperature T_c with oxygen content x for $YBa_2Cu_3O_x$.

The oxygen atoms are not lost at random during reduction but come solely from the CuO_4 square planar units and lie along the **b**-axis. This has the effect of converting the copper co-ordination from square planar to linear, resulting in CuO_2 chains running through the structure. This structural feature appears to be of vital importance in allowing the superconducting transition to take place.

11.5.5.4 Perovskite-Related Structures and Series

The highest transition temperatures found in the cuprate superconductors are in rather complicated structures that are built from layers of an $ACuO_3$ perovskite parent n octahedra

in thickness. These, as well as the hole-conducting species described in the preceding sections, can be regarded as members of several *homologous series* of phases, each member of which differs from the preceding one by having the perovskite slab one extra octahedron in thickness (Table 11.3).

Table 11.3 Homologous series of cuprate superconductors.

Charge reservoir[a]	Charge reservoir formula	Superconducting slab formula	Idealised series formula[a]	Examples[b]
AO-AO	LaO	CuO_2 $(CuO_2)_{n-1}$	$La_2Cu_nO_{2n+2}$	$La_2CuO_{4+\delta}$ ($n = 1$) $La_{1-x}Sr_xCuO_4$ ($n = 1$)
	SrF	CuO_2 $(CuO_2)_{n-1}$	$Sr_2Cu_nO_{2n}F_2$	$Sr_2CuO_2F_{2+\delta}$ ($n = 1$)
AO-M-AO	BaO-Cu-BaO	CuO_2 (Y $CuO_2)_{n-1}$	$Ba_2Y_{n-1}Cu_{n+1}O_{2n+2}$	$CuBa_2YCu_2O_{6+\delta}$ = $YBa_2Cu_3O_{6+\delta}$ ($n = 2$)
		CuO_2 (Ca $CuO_2)_{n-1}$	$Ba_2Ca_{n-1}Cu_{n+1}O_{2n+2}$	$CuBa_2Ca_2Cu_3O_{8+\delta}$ ($n = 3$) $CuBa_2Ca_3Cu_4O_{10+\delta}$ ($n = 4$)
	BaO-Hg-BaO	CuO_2 (Ca $CuO_2)_{n-1}$	$HgBa_2Ca_{n-1}Cu_nO_{2n+2}$	$HgBa_2CuO_{4+\delta}$ ($n = 1$) $HgBa_2CaCu_2O_{6+\delta}$ ($n = 2$) $HgBa_2Ca_2Cu_3O_{8+\delta}$ ($n = 3$)
AO-MO-AO	BaO-TlO-BaO	CuO_2 (Ca $CuO_2)_{n-1}$	$TlBa_2Ca_{n-1}Cu_nO_{2n+3}$	$TlBa_2CuO_5$ ($n = 1$) $TlBa_2CaCu_2O_{7+\delta}$ ($n = 2$) $TlBa_2Ca_2Cu_3O_{9+\delta}$ ($n = 3$)
	BaO-BiO-BaO	CuO_2 (Ca $CuO_2)_{n-1}$	$BiBa_2Ca_{n-1}Cu_nO_{2n+3}$	$BiBa_2CuO_{5+\delta}$ ($n = 1$) $BiBa_2CaCu_2O_{7+\delta}$ ($n = 2$) $BiBa_2Ca_2Cu_3O_{9+\delta}$ ($n = 3$)
	SrO-GaO-SrO	CuO_2 (Y,Ca $CuO_2)_{n-1}$	$GaSr_2$(Y, Ca)$_{n-1}Cu_nO_{2n+3}$	$GaSr_2$(Y, Ca)$Cu_2O_{7+\delta}$ ($n = 2$) $GaSr_2$(Y, Ca)$_2Cu_3O_{9+\delta}$ ($n = 3$)
AO-MO-MO-AO	SrO-BiO-BiO-SrO	CuO_2 (Ca $CuO_2)_{n-1}$	$Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4}$	$Bi_2Sr_2CuO_{6+\delta}$ ($n = 1$) $Bi_2Sr_2CaCu_2O_{8+\delta}$ ($n = 2$) $Bi_2Sr_2Ca_2Cu_3O_{10+\delta}$ ($n = 3$)
	BaO-TlO-TlO-BaO	CuO_2 (Ca $CuO_2)_{n-1}$	$Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+4}$	$Tl_2Ba_2CuO_{6+\delta}$ ($n = 1$) $Tl_2Ba_2CaCu_2O_{8+\delta}$ ($n = 2$) $Tl_2Ba_2Ca_2Cu_3O_{10+\delta}$ ($n = 3$)
AO-MO-M′-MO-AO	SrO-PbO-Cu-PbO-SrO	CuO_2 (Y,Ca $CuO_2)_{n-1}$	Pb_2Sr_2(Y, Ca)$_{n-1}Cu_{n+1}O_8$	$Pb_2Sr_2Y_{0.5}Ca_{0.5}Cu_3O_{8+\delta}$ ($n = 2$)

[a] The charge reservoir layer is nonstoichiometric. This is not indicated in the idealised formulae given.

[b] The composition is variable, due to the nonstoichiometric nature of the charge reservoir slabs; δ indicates this and may be either positive or negative.

In all these compounds, the part of the structure that leads to superconductivity is the slab of CuO_2 sheets. When more than one sheet is present, they are separated by cation layers, Q (usually Ca or Y) to give a sequence $CuO_2 – (Q – CuO_2)_{n-1}$, which form the *superconducting layers* in the material (Figure 11.38). The index n is the total number of CuO_2 layers in the phase, which is equal to the formula number of Cu atoms present.

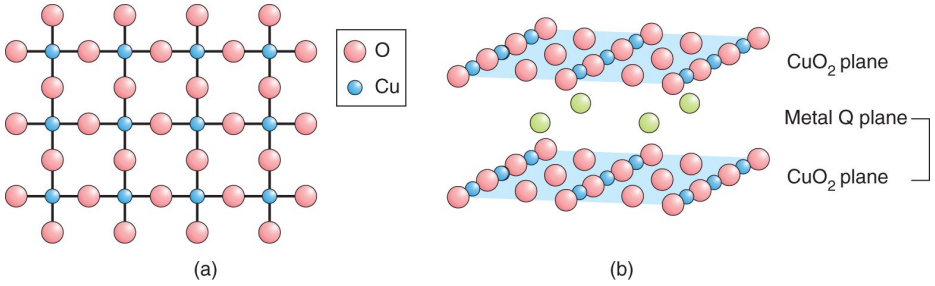


Figure 11.38 Superconducting sheets found in cuprate superconductors: (a) a CuO_2 layer; (b) a CuO_2–(Q–CuO_2) layer.

The CuO_2 layers themselves are not, as such, superconducting, and have to be doped with (usually) holes. These are provided by the *charge reservoir* layers, which separate each $CuO_2 – (Q – CuO_2)_{n-1}$ superconducting slab. The general structure of a charge reservoir sheet is $AO – [MO_x]_m – AO$, where A is a lanthanoid such as La, or alkaline earth, typically Sr or Ba, M is a metal, typically Bi, Pb, Tl, Hg, present as a non-stoichiometric oxide layer with x usually close to 0 or 1.0, and m takes values of 0, 1, 2, etc. Doping the charge reservoir layer with acceptors, mainly oxygen, favours the creation of holes, which are subsequently transferred to the CuO_2 slab to induce metallic and superconducting properties (Figure 11.39).

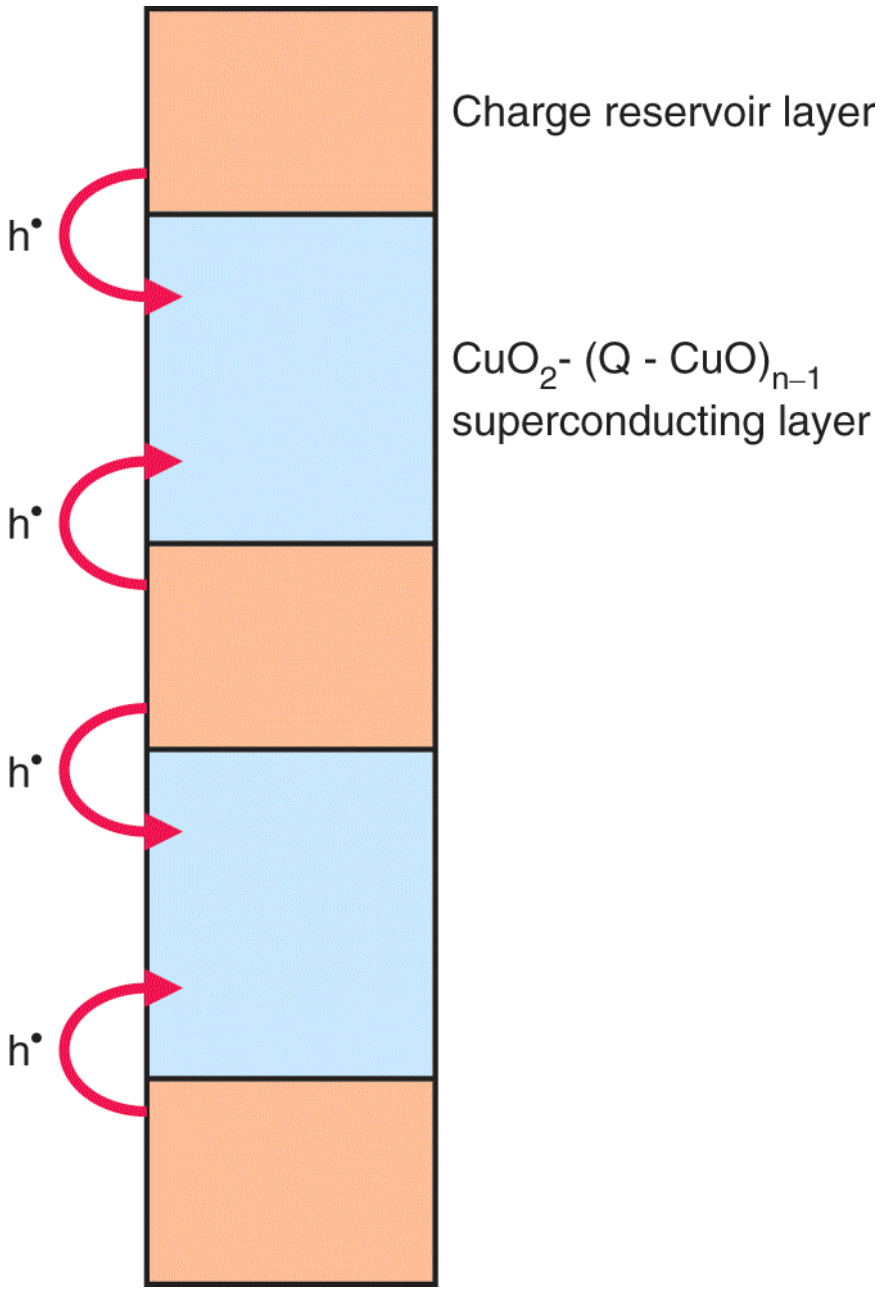


Figure 11.39 Doping into the charge reservoir layers of a cuprate superconductor results in hole transfer into the superconducting layers.

The construction principle can be illustrated with reference to compounds in which the perovskite sheets are linked by Bi_2O_2 or Tl_2O_2 layers. The formula of the Bi_2O_2 series is $Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4}$ and that of the Tl_2O_2 series is $Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+4}$, where it must be understood that the oxygen composition in the real compounds is slightly different from that given in the idealised formulae. The *idealised* structures of the first three members of these series, $Bi_2Sr_2CuO_6$ (= $Tl_2Ba_2CuO_6$), $Bi_2Sr_2CaCu_2O_8$ (= $Tl_2Ba_2CaCu_2O_8$), and $Bi_2Sr_2Ca_2Cu_3O_{10}$ (= $Tl_2Ba_2Ca_2Cu_3O_{10}$), contain perovskite layers 1, 2, and 3 octahedra thick (Figure 11.40). The single perovskite sheets in the idealised structure of $Bi_2Sr_2CuO_6$ are complete and separated by Bi_2O_2 (or Tl_2O_2) layers (Figure 11.40a). In the other compounds,

the oxygen structure needed to form the perovskite framework is incomplete. The nominal double layer of CuO_6 octahedra needed to form a sheet of idealised perovskite structure is replaced by square pyramids in $Bi_2Sr_2CaCu_2O_8$ and $Tl_2Sr_2CaCu_2O_8$ (Figure 11.40b). To make the relationship clearer, the octahedra are completed in faint outline in Figure 11.40c. ln the phases $Ba_2Ca_2Sr_2Cu_3O_{10}$ and $Tl_2Ca_2Sr_2Cu_3O_{10}$, the three CuO_6 octahedral perovskite layers have been replaced by two sheets of square pyramids and the middle layer by a sheet of CuO_4 squares (Figure 11.40d). The octahedra are completed in faint outline in Figure 11.40e.

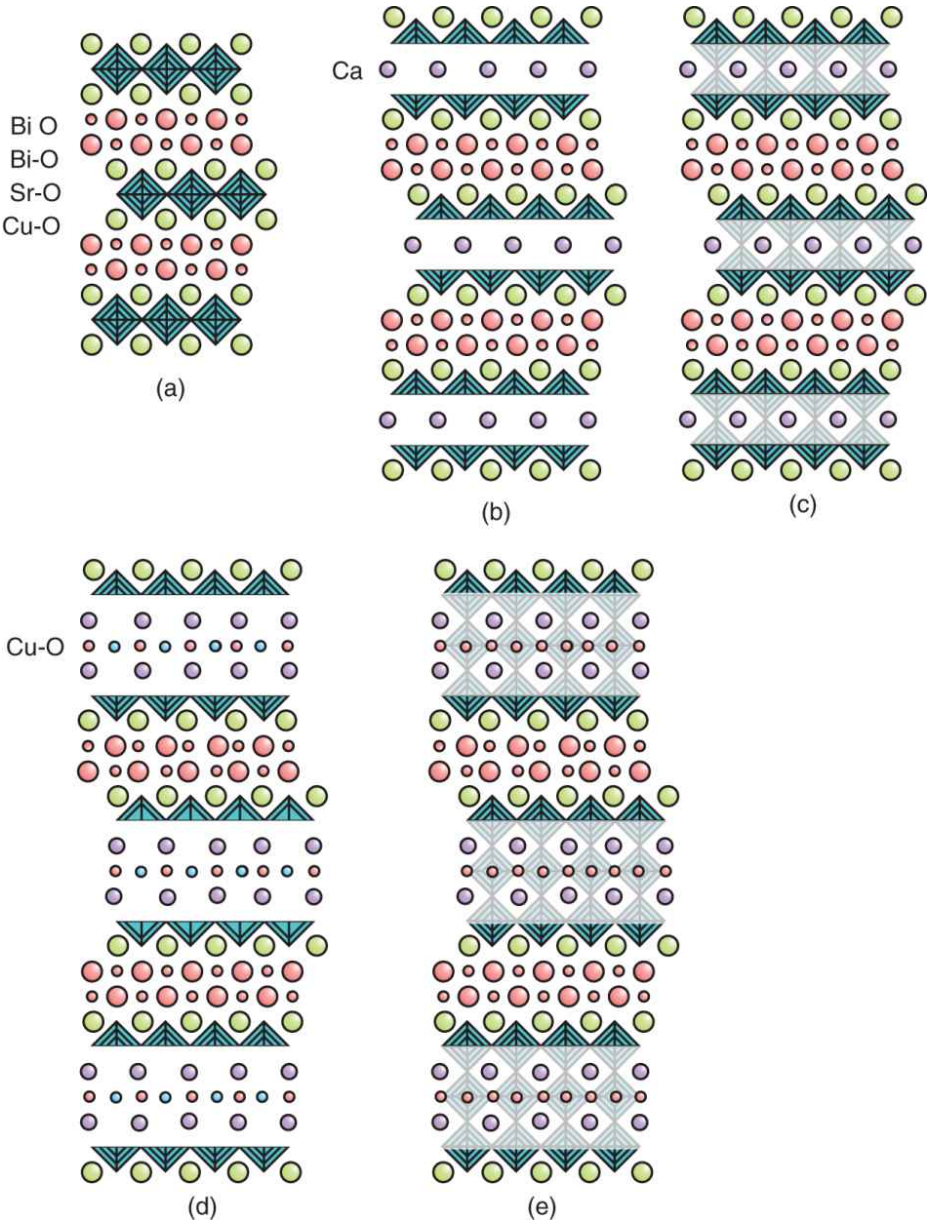


Figure 11.40 Idealised structures of the cuprate superconductors: (a) $Bi_2Sr_2CuO_6$ (=$Tl_2Ba_2CuO_6$); (b) $Bi_2Sr_2CaCu_2O_8$ (=$Tl_2Ba_2CaCu_2O_8$); (c) as (b) but with the nominal CuO_6 octahedra completed in faint outline; (d) $Bi_2Sr_2Ca_2Cu_3O_{10}$ (=$Tl_2Ba_2Ca_2Cu_3O_{10}$); (e) as (d) but with the nominal CuO_6 octahedra completed in faint outline.

Homologous series of superconductors are derived from one type of charge reservoir slab interleaved with $n = 1, 2, 3, 4$, etc. $CuO_2 - (Q - CuO_2)_{n-1}$ slabs (Table 11.3). Clearly, other series built from different ordered slab thicknesses can be envisaged, as can series with different alternating charge reservoir slabs, but synthesis of these complex structures might well pose problems.

The doping of cuprates to induce superconductivity is often described in terms of the electronic behaviour of the charge reservoir layers in these complex structures; a different approach to that described for semiconductors. This concept, using hole doping as an illustration, can best be explained with a number of examples. Initially the materials are treated as if they are ionic. The ionic charges on the charge reservoir layers and the superconducting layers are determined. If the two charge totals are equal the material as formulated is an insulator not a superconductor.

(i) Charge reservoir (AO AO), typified by La_2CuO_4.

a. Charge reservoir:

Formula: $(La^{3+}O^{2-}\ La^{3+}O^{2-})$

Charge: $2(+3 - 2) = +2$

b. Superconducting layer:

Formula: $(Cu^{2+}O_2)$

Charge: $(+2 - 4) = -2$

These charges exactly balance, and the material is expected to be an insulator; a normal ionic compound. Hole doping can be achieved by adding a lower valence cation such as Sr^{2+} or oxygen interstitials to the charge reservoir. For Sr^{2+} acceptor dopants:

a. Charge reservoir:

Formula: $(La_{1-x/2}Sr_{x/2}O)\ (La_{1-x/2}Sr_{x/2}O)$

Charge: $2[+3(1 - x/2) + 2(x/2) - 2] = +2-x$

b. Superconducting layer:

Formula: (CuO_2)

Charge: $(+2 - 4) = -2$

The charge difference between the charge reservoir and the superconducting layers, $-x$, must be achieved by the addition of balancing charges, $x\ h^{\bullet}$ in this case, which are introduced into the superconducting layers and give rise to superconductivity.

(ii) Charge reservoir AO M AO typified by $YBa_2Cu_3O_6 - YBa_2Cu_3O_7$.

a. Charge reservoir:

Formula: $(Ba^{2+}O^{2-}\ Cu^{+}\ Ba^{2+}O^{2-})$

Charge: $(+5 - 4) = +1$

b. Superconducting layer:

Formula: $(Cu^{2+}O_2{}^{2-}\ Y^{3+}\ Cu^{2+}O_2{}^{2-})$

Charge: (+4 + 3 – 8) = –1

These charges exactly balance, and $YBa_2Cu_3O_6$ is an insulator. Hole doping is achieved by adding oxygen interstitials to the charge reservoir to give a composition $YBa_2Cu_3O_{6+\delta}$:

a. Charge reservoir:

Formula: (BaO CuO_δ BaO)

Charge: (+5 – 4 – 2δ) = +1 – 2δ

b. Superconducting layer:

Formula: (CuO_2 Y CuO_2)

Charge: (+4 + 3 – 8) = –1

The charge difference between the charge reservoir and the superconducting layers, -2δ, must be balanced by the addition of 2δ $h^\bullet$. These are introduced into the superconducting layers and give rise to superconductivity.

The idea of charge reservoirs can be applied to the other members of the series.

11.5.6 Bi-layer Graphene

From among the various unconventional superconductors now known, including the cuprates described earlier, and the iron pnictides (Fe-N-P-As-Sb phases), the superconducting behaviour of bi-layer graphene is (to date) unique. Graphene consists of single sheets of hexagonal rings of carbon atoms (Figure 2.4a), which is essentially a single layer of the carbon allotrope graphite. Boron nitride, BN (Figure 2.4b), is structurally similar but has sheets composed of alternating B and N atoms in a hexagonal array. Both graphite and boron nitride are van der Waals materials, built up from layers of strongly bonded atoms lying in the plane of the sheet, these being held together by weak van der Waals interactions, allowing them to be can be readily separated into sheets.

If the two graphene sheets in a bilayer are misaligned through a twist angle, the ‘magic angle’, of 1.1°, the bilayer unit becomes superconducting. The transition temperature is approximately 1.7 K and the superconducting region is found to exist under a dome, as with the cuprates. Structurally, the two sheets in graphene can be stacked one exactly on top of the other, in an …A.A… arrangement, or with the second sheet translated to give an …A.B… arrangement. In the twisted phase a Moiré pattern occurs in which alternating regions have an essentially …AAA… packing adjoining an …AB… arrangement. The superconducting behaviour is also dependent upon the substrate that the graphene layers are linked with. The commonly used substrate is BN, with an almost identical structure, but the imposition of a layer of a different material can stabilise the superconducting region over a different range of twist angles.

At present superconductivity in bilayer graphene is still the subject of intense scrutiny.

11.6 Nanostructures and Quantum Confinement of Electrons

Quantum wells, quantum wires, and quantum dots (Section 1.1.3) have unique electronic and optical properties because of the way in which electrons are localised, or *confined*. In bulk solids, electrons are located on atom cores in ionic solids, in localised bonds in many covalent solids, delocalised over molecular orbitals in others, such as graphite, or completely delocalised, as in metals. Quantum nanostructures confine the electrons (and holes in semiconductors), at a scale different than any of these, and this gives rise to the novel properties that such structures possess.

A quantum well is constructed by laying down a thin layer of a semiconductor with a smaller band gap within a semiconductor with a larger band gap. The most studied quantum well structures are those formed from a layer of gallium arsenide sandwiched in gallium aluminium arsenide, GaAlAs (Figure 11.41a). Gallium arsenide has a band gap of about 1.42 eV, while aluminium arsenide has a band gap of about 2.16 eV. Gallium aluminium arsenide alloys have band gaps between these values. The electrons in the thin GaAs layer are effectively trapped in the 'well' formed in the conduction band (CB) of the composite material (Figure 11.41b). Similarly, the holes in the thin layer of semiconductor are trapped at the 'hill' in the valence band (VB) of the composite material. The properties of single quantum wells are enhanced when a number of these features are combined to form a *multiple quantum well* or *superlattice* (Figure 11.41c,d).

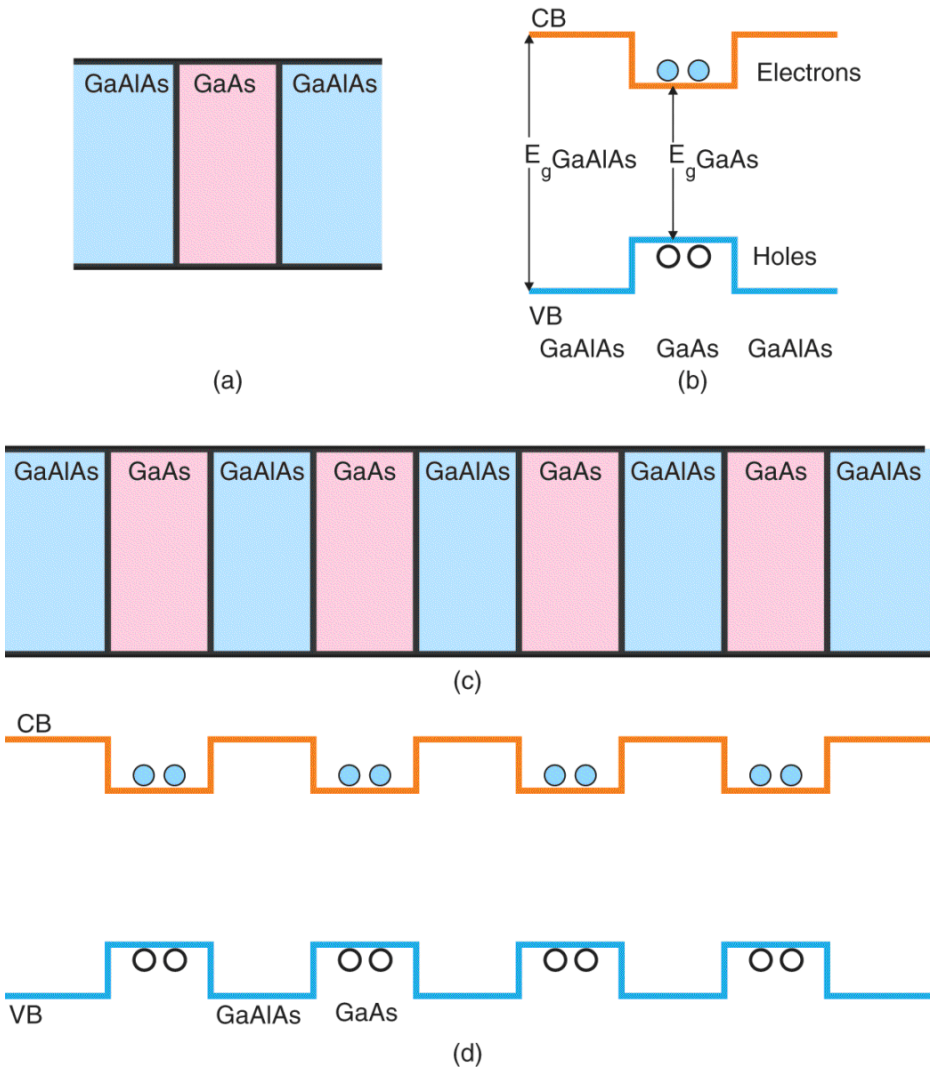


Figure 11.41 Quantum wells: (a) a single quantum well in gallium arsenide, GaAs; formed within gallium aluminium arsenide GaAlAs; (b) schematic energy band structure of the quantum well; (c) multiple quantum well superlattice of GaAs and GaAlAs layers; (d) schematic energy band structure of the superlattice. CB, conduction band; VB, valence band.

Quantum confinement becomes important at a dimension Δx, of:

$$\Delta x \approx \sqrt{(h^2/m_e^* k_B T)}$$

for electrons and

$$\Delta x \approx \sqrt{(h^2/m_h^* k_B T)}$$

for holes, where h is Planck's constant, m_e* the effective mass of the electron in the semiconductor, m_h* the effective mass of the hole in the semiconductor, k_B is Boltzmann's constant, and T the temperature (K). For quantum confinement to be important, quantum wells must be about 10 atom layers or less in thickness.

The energy of an electron in a quantum well can be calculated using classical band theory. If it is assumed that the electron is free, and trapped by an infinite boundary potential, the same

equations that apply to a free electron in a metal apply. Thus, the energy, E, of a free electron in a rectangular parallelepiped with edges a, b, and c is given by:

$$E(n_x, n_y, n_z) = [h^2/8m_e][n_x{}^2/a^2 + n_y{}^2/b^2 + n_z{}^2/c^2] \tag{11.6}$$

where h is Planck's constant, m_e is the mass of the electron, and n_x, n_y, and n_z are the quantum numbers along the three axes. Exactly the same equation will apply to a free electron confined to a slab of material, although it is better to replace the electron mass with the effective mass, $m_e{}^*$. In the case of a quantum well, the electron is confined in one dimension, say, x, and unconfined in two directions, which can be taken as y and z, so it is convenient to rewrite Eq. (11.6) as:

$$E(n_x, n_y, n_z) = [h^2/8m_e^*][n_x{}^2/a^2] + [h^2/8m_e^*] \times [n_y{}^2/b^2 + n_z{}^2/c^2] \tag{11.7}$$

The values of b and c can be taken as about 1 cm, while the value of a is about 10^{-8} m. The energy is therefore dominated by the first term in Eq. (11.7). This introduces a new set of energy levels, associated with electron waves trapped in the well (Figure 11.42). The electron energy level in the lowest, $n = 1$, state is raised by $h^2/8m_e{}^*a^2$ compared with the base of the well. These energy levels are called *electron subbands*, and when the energy levels trap the electrons they are strongly confined. Exactly the same equations apply to holes when the effective mass $m_h{}^*$ replaces or $m_e{}^*$. The energy levels that arise from trapped holes are called *hole subbands*.

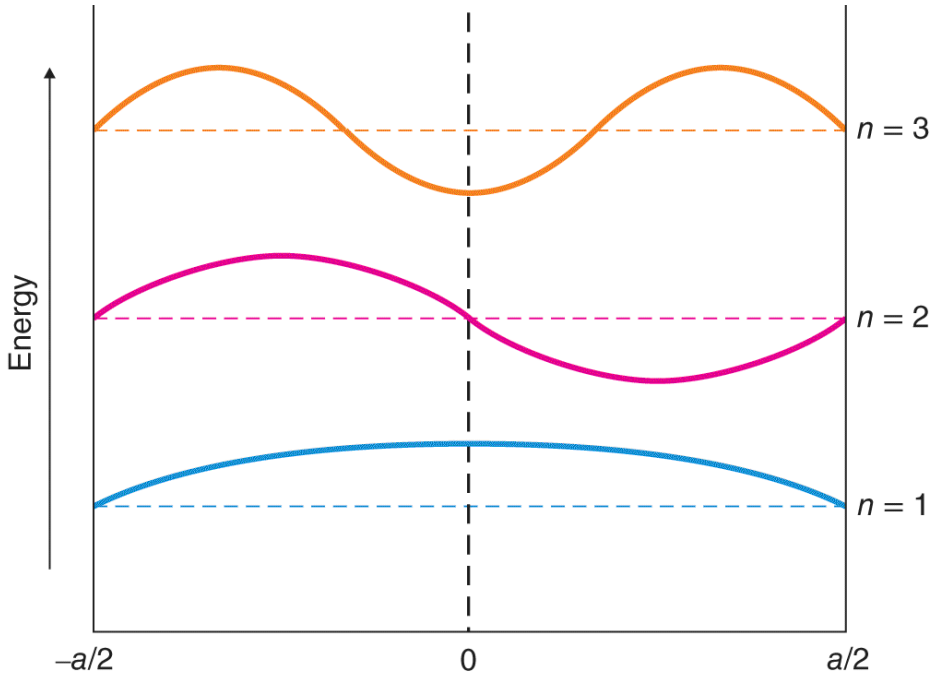


Figure 11.42 The first three energy levels for an electron trapped in a one-dimensional quantum well.

The aforementioned considerations can be applied equally well to confinement in two or three dimensions, in quantum wires and quantum dots. For a quantum wire with restricted dimensions along a and b, the free electron confined in an infinite potential well will have energy levels given by Eq. (11.6), where a and b are small and c is large. In the case of the quantum dot, Eq. (11.6) is retained, but the third dimension, c, is also small.

The optical properties of these structures are described in Section 12.10.

Further Reading

The Band Theory Definition of a Semiconductor is Due to A.H. Wilson:

Wilson, A.H. (1931). The theory of electronic semiconductors. *Proc. R. Soc. London***A133**: 458.

Conductivity of (Mainly) Inorganic Solids Due to Defects is Covered In:

Tilley, R.J.D. (2008). *Defects in Solids*. Hoboken, NJ: John Wiley & Sons, especially Chapters 7 and 8.

The Metal-Insulator Transition in VO_2:

Liu, K., Lee, S., Yang, S. et al. (2018). Recent progress on physics and applications of vanadium dioxide. *Mater. Today***21**: 875.

Polymers:

Epstein, A.P. (1997). Electrically conducting polymers: science and technology. *Mater. Res. Soc. Bull.***22**: 6.

Heeger, A.J. (2001). Semiconducting and metallic polymers: the fourth generation of polymeric materials. *Mater. Res. Soc. Bull.***22**: 900.

Superconductivity:

Errea, I., Belli, F., Monacelli, L. et al. (2020). Quantum crystal structure in the 250-kelvin superconducting lanthanum hydride. *Nature***578**: 66.

Massee, F., Huang, Y.K., and Aprili, M. (2020). Atomic manipulation of the gap in $Bi_2Sr_2CaCu_2O_{8+x}$. *Science***367**: 68.

The Following Articles in Scientific American Give a Good Overview of the Early Years of High Temperature Superconductivity:

Cava, R.J. (1990). Superconductors beyond 1 – 2 – 3. *Sci. Am.***263**: p24.

Clarke, J. (1994). SQUIDs. *Sci. Am.***271**: p36.

de Bruyn Ouboter, R. (1997). H. K. Ohnes's discovery of superconducivity. *Sci. Am.***276**: p84.

Hazen, R.M. (1988). Perovskites. *Sci. Am.***258**: p52.

Wolsky, A.M., Giese, R.F., and Daniels, E.J. (1989). The new superconductors: prospects for applications. *Sci. Am.***260**: p44.

Graphene Bilayers:

Arora, H.S., Polski, R., Zhang, Y. et al. (2018). Superconductivity in metallic twisted bilayer graphene stabilized by WSe_2. *Nature***583**: 379.

Cao, Y., Fatemi, V., Fang, S. et al. (2018). Unconventional superconductivity in magic-angle graphene superlattices. *Nature***556**: 43.

Stepanov, P., Das, I., Lu, X. et al. (2020). Untying the insulating and superconducting orders in magic-angle graphene. *Nature***583**: 375.

Quantum Hall Effect:

Tong, D. (2016). *The Quantum Hall Effect*. TIFR Infosys Lectures, January, available at www.damtp.cam.ac.uk/user/tong/qhe.html.

Michalakis, S. (2020). Quantum leap. *Sci. Am.***323**: 48.

Problems and Exercises

Quick Quiz

11.1 An insulator is a material with a full valence band, an empty conduction band, and

a. No band gap.

b. A small band gap.

c. A large band gap.

11.2 An intrinsic semiconductor is a material with a full valence band, an empty conduction band, and

a. No band gap.

b. A small band gap.

c. A large band gap.

11.3 In an intrinsic semiconductor the current is carried by

a. Electrons.

b. Holes.

c. Electrons and holes.

11.4 Donors make a semiconductor

a. p-type.

b. n-type.

c. Degenerate.

11.5 Acceptors make a semiconductor

a. p-type.

b. n-type.

c. Degenerate.

11.6 A semimetal has

a. A partly filled uppermost band.

b. Overlapping uppermost bands.

c. A small band gap between the two uppermost bands.

11.7 The conductivity of a metal

a. Increases as the temperature increases.

b. Falls as the temperature increases.

c. Is insensitive to temperature.

11.8 The conductivity of a semiconductor

a. Increases as the temperature increases.

b. Falls as the temperature increases.

c. Is insensitive to temperature.

11.9 A semiconductor crystal is transparent to radiation with energy

a. Greater than the band gap.

b. Less than the band gap.

c. Exactly equal to the band gap.

11.10 The band gap of a semiconductor

a. Does not depend on atom size.

b. Increases with increasing atom size.

c. Decreases with increasing atom size.

11.11 Atoms to the right of silicon and germanium in the Periodic Table act as:

a. Donors.

b. Acceptors.

c. Neither.

11.12 Atoms to the left of silicon and germanium in the periodic Table make the material

a. n-type.

b. p-type.

c. Neither.

11.13 When a current flows across a p–n junction under forward bias, it is made up of

a. Six components

b. Four components.

c. Two components.

11.14 Doping a transition metal oxide with a cation of lower valence will make it

a. p-type.

b. n-type.

c. Cause no change.

11.15 Conducting polymers contain

a. Metal atoms in the structure.

b. Conjugated double bonds.

c. Conjugated triple bonds.

11.16 Doping polyacetylene with sodium makes it

a. A metallic conductor.

b. A p-type semiconductor.

c. An n-type semiconductor.

11.17 A semiconductor quantum well is

a. A thin layer of a semiconductor on an insulator.

b. Alternating layers of two semiconductors on an insulator.

c. A thin layer of a semiconductor within a different semiconductor.

11.18 Electrons in a quantum wire are strongly confined

a. In one dimension.

b. In two dimensions.

c. In three dimensions.

11.19 A type I superconductor

a. Does not interact with magnetic fields.

b. Draws an external magnetic field into itself.

c. Expels internal magnetic fields.

11.20 A material in the superconducting state can be thought of as

a. A perfect diamagnetic solid.

b. A perfect paramagnetic solid.

c. A perfect ferromagnetic solid.

11.21 Superconductivity in conventional (low temperature) superconductors is due to

a. Pairs of electrons.

b. Pairs of holes.

c. Electron–hole pairs.

11.22 Ceramic superconductors mostly have structures closely related to

a. Spinel.

b. Perovskite.

c. Halite.

11.23 A Josephson junction consists of

a. A thin layer of insulator separating two superconducting regions.

b. A thin layer of superconductor separating two insulating regions.

c. A thin layer of superconductor separating two metallic regions.

11.24 A SQUID measures

a. Resistivity.

b. Superconductivity.

c. Magnetic fields.

Calculations and Questions

11.1 The electrical resistivity of gold at 273 K is 2.05×10^{-8} Ω m. Gold adopts the face-centred cubic structure with a lattice parameter of 0.4078 nm. The velocity of electrons at the Fermi surface is 1.40×10^{6} m s^{-1}. Each gold atom contributes one electron to the structure. Calculate:

a. the relaxation time.

b. the mean free path of the electrons.

c. Compare the mean free path to the interatomic spacing of gold atoms in the crystal.

11.2 The electrical resistivity of silver, at 273 K, is 1.47×10^{-8} Ω m. Silver adopts the face-centred cubic structure with a lattice parameter of 0.4086 nm. The velocity of electrons at the Fermi surface is 1.39×10^{6} m s^{-1}. Each silver atom contributes one electron to the structure. Calculate:

a. the relaxation time.

b. the mean free path of the electrons.

c. Compare the mean free path with the interatomic spacing of silver atoms in the crystal.

11.3 3 The electrical resistivity of rubidium, at 273 K, is 11.5×10^{-8} Ω m. Rubidium adopts the body-centred cubic structure with a lattice parameter of 0.5705 nm. The velocity of electrons at the Fermi surface is 8.1×10^{7} m s^{-1}. Each rubidium atom contributes one electron to the structure. Calculate:

a. the relaxation time.

b. the mean free path of the electrons.

c. Compare the mean free path to the interatomic spacing of rubidium atoms in the crystal.

11.4 The electrical resistivity of magnesium, at 273 K, is 4.05×10^{-8} Ω m. Magnesium adopts the hexagonal structure with lattice parameters of a = 0.3209 nm, c = 0.5211 nm. The velocity of electrons at the Fermi surface is 1.58×10^{6} m s^{-1}. Each magnesium atom contributes two electrons to the structure. Calculate:

a. the relaxation time.

b. the mean free path of the electrons.

c. Compare the mean free path with the interatomic spacing of rubidium atoms in the crystal.

11.5 The electrical resistivity of liquid mercury is a function of temperature is given in the following table. The velocity of electrons at the Fermi surface is 1.52×10^{6} m s^{-1}. Determine how the mean free path varies with temperature. The density of liquid mercury is 13 456 kg m^{-3}. Assume that this does not vary with temperature and that each mercury atom contributes two mobile electrons to the liquid.

ρ (10^{-8} Ω m)	**94.1**	**103.5**	**128.0**	**214.0**	**630.0**
T (°C)	0	100	300	700	1200

11.6 The resistivity of cadmium metal crystals at room temperature is 7.79×10^{-8} Ω m parallel to the **c**-axis and 6.54×10^{-8} Ω m parallel to the **a**-axis. Cadmium adopts the hexagonal structure with unit cell parameters of a = 0.2979 nm, c = 0.5620 nm. Each Cd atom contributes two mobile electrons to the crystal. Calculate the mobility of the electrons along the unit cell axes.

11.7 The resistivity of zinc metal crystals at room temperature is 6.05×10^{-8} Ω m parallel to the**c**-axis and 5.83×10^{-8} Ω m parallel to the **a**-axis. Zinc adopts the hexagonal structure with unit cell parameters of a = 0.2665 nm, c = 0.4947 nm. Each Zn atom contributes two mobile electrons to the crystal. Calculate the mobility of the electrons along the unit cell axes.

11.8 The resistivity of a sample of brass containing 70 wt% Cu and 30 wt% Zn is 6.3×10^{-8} Ω m at 0 °C, compared with that of pure copper, which is 1.54×10^{-8} Ω m at the same temperature. Determine the residual resistivity at this temperature.

11.9 The resistivity of a sample of bronze containing 90 wt% Cu and 10 wt% Sn is 1.36×10^{-7} Ω m at 0 °C, compared with that of pure copper, which is 1.54×10^{-8} Ω m at the same temperature. Determine the residual resistivity at this temperature.

11.10 The resistivity at 0 °C for a number of nickel alloys is given in the table. Plot the resistivity against the amount of nickel in the alloy. Comment on the shape of the plot in terms of the possible structures of the alloys.

Alloy	**(Nickel)**	**Alumel**	**Chromel P**	**Nichrome**	**Monel**
Resistivity (10^{-8} Ω m)	6.16	28.1	70	107.3	42.9
wt% Ni	100	95	90	77.3	67.1

11.11 The resistivity at 0 °C for a number of aluminium alloys is given in the following table. Plot the resistivity against the amount of aluminium in the alloy. Comment on the shape of the plot in terms of the possible structures of the alloys.

Alloy	**(Aluminium)**	**RR59**	**RR57**	**Alpax gamma**	**Lo-Ex**
Resistivity (10^{-8} Ω m)	2.42	3.5	3.95	3.5	3.95
wt% Al	100	93	89	87	85

11.12 The resistivity at 0 °C for a number of copper alloys is given in the following table. Plot the resistivity against the amount of copper in the alloy. Comment on the shape of the plot in terms of the possible structures of the alloys.

Alloy	**(Copper)**	**Bronze**	**Manganin**	**Brass**	**German**	**Constantan**
					silver	
Resistivity (10^{-8} Ω m)	1.54	19.8	41.5	6.3	40	49
wt% Cu	100	90	84	70	62	60

11.13 Estimate the number of intrinsic electrons, n, and holes, p, and the product np, for a crystal of silicon at 300 K taking the effective mass of electrons and holes as equal to the electron mass, m_e. E_g is 1.12 eV.

11.14 Estimate the number of intrinsic electrons, n, and holes, p, and the product np, for a crystal of gallium arsenide at 300 K taking the effective mass of electrons and holes as equal to the electron mass, m_0. E_g is 1.42 eV.

11.15 Estimate the number of intrinsic electrons, n, and holes, p, and the product np, for a crystal of gallium arsenide at 300 K taking the effective mass of electrons to be 0.067 m_e and holes as 0.082 m_e. E_g is 1.42 eV.

11.16 Estimate the number of intrinsic electrons, n, and holes, p, and the product np, for a crystal of cadmium selenide at 300 K taking the effective mass of electrons to be 0.13 m_e and holes as 0.45 m_e. E_g is 1.70 eV.

11.17 Determine the band gap of silicon from the conductivity data given in the table. What is the minimum frequency of light that will excite an electron across the band gap?

Temperature (°C)	**227**	**277**	**327**	**377**	**427**	**477**	**527**
Conductivity (Ω^{-1} m^{-1})	3×10^{-4}	9×10^{-4}	3×10^{-3}	8×10^{-3}	2.5×10^{-2}	6×10^{-2}	8×10^{-2}

11.18 Determine the band gap of germanium from the conductivity data given in the table. What is the minimum frequency of light that will excite an electron across the band gap?

Temperature (°C)	**5**	**47**	**82**	**122**	**162**	**240**	**344**	**441**
Conductivity (Ω^{-1} m^{-1})	0.0001	0.0008	0.004	0.009	0.05	0.1	0.6	1.0

11.19 Calculate the donor energy level position in silicon doped with phosphorus using the Bohr model. The effective mass of an electron is 0.33 m_e and the relative permittivity of silicon is 11.7.

11.20 Calculate the acceptor energy level position in germanium doped with aluminium using the Bohr model. The effective mass of a hole is 0.16 m_e and the relative permittivity of germanium is 16.0.

11.21 Calculate the donor energy level position in indium phosphide doped with tin using the Bohr model. The effective mass of an electron is 0.067 m_e and the relative permittivity of indium phosphide is 12.4.

11.22 Calculate the acceptor energy level position in gallium arsenide doped with zinc using the Bohr model. The effective mass of a hole is 0.082 m_e and the relative permittivity of gallium arsenide is 13.2.

11.23 The energy gap for gallium arsenide is 1.4 eV at 300 K. The effective mass of electrons is 0.067 m_e and of holes is 0.082 m_e. How near to the band gap centre is the Fermi level?

11.24 The energy gap for gallium phosphide is 2.26 eV at 300 K. The effective mass of electrons is 0.82 m_e and of holes is 0.60 m_e. How near to the band gap centre is the Fermi level?

11.25 A semiconductor containing 10^{20} holes m^{-3} and 10^{18} electrons m^{-3} has a conductivity of 0.455 Ω^{-1} m^{-1}. The ratio of the mobilities of electrons and holes, μ_e/μ_h, is 10. What are the hole and electron mobilities?

11.26 Gallium arsenide is doped with 10^{18} donor atoms. The hole mobility is 0.04 m^2 V^{-1} s^{-1} and the electron mobility is 0.85 m^2 V^{-1} s^{-1}. Using the results of question, determine the conductivity of the sample.

11.27 A 1 cm cube of n-type germanium supports a current of 6.4 mA when a voltage of 10 mV is applied across two parallel faces. The charge carriers have a mobility of 0.39 m^2 V^{-1} s^{-1}. Determine the Hall coefficient of the crystal assuming that only the majority charge carriers need be considered.

11.28

a. Estimate the Hall coefficient for intrinsic silicon, using values $n_i = 1 \times 10^{16}$ m^{-3}, mobility of electrons = 0.15 m^2 V^{-1} s^{-1}, mobility of holes = 0.045 m^2 V^{-1} s^{-1}.

b. Calculate the Hall voltage for a 1 cm cube of pure silicon at 20 °C, in a magnetic induction of 0.2 T, when a current of 10^{-3} A is applied to a cube face.

11.29 A single crystal of germanium doped with antimony to make it p-type is used in a Hall experiment. The crystal dimensions are 20 mm × 10 mm × 1 mm in the **x**-, **y**-, and **z**-directions. In a constant induction of 0.9 T along the **z**-direction, the current (along **x**)

and voltage (along **y**) readings were obtained. Calculate the Hall coefficient and the carrier density.

Current (x) (mA)	**8**	**16**	**24**	**32**	**40**
Voltage (y) (V)	0.081	0.159	0.233	0.318	0.401

11.30 Using Figure 11.24, estimate the activation energy for the conductivity of polyacetylene doped with iodine for the three concentrations shown.

11.31 The oxides TiO_2 and Ti_2O_3 both show composition ranges. The oxide TiO_2 loses a small amount of oxygen to form TiO_{2-x}. Is it likely to show p-type or n-type semiconductivity? The oxide Ti_2O_3 gains a slight amount of oxygen to form Ti_2O_{3+x}. Is it likely to show n-type or p-type semiconductivity?

11.32 The oxide $LaCoO_3$ is an insulator. What are the charges on the cations present? The compound is doped with Sr^{2+} to form $La_{1-x}Sr_xCoO_3$, in which the Sr^{2+} substitutes for the La. What are the charges on the cations present? Is the doped material a p-type or n-type semiconductor?

11.33 The compound Mg_2TiO_4 is an inverse spinel and the compound $MgTi_2O_4$ is a normal spinel (Section 3.2.2.8). Both compounds are insulators.

a. What are the charges on the ions?

b. A small amount of $MgTi_2O_4$ is doped into Mg_2TiO_4 to form $Mg_{2-x}Ti_{1+x}O_4$.

c. What are the charges on the ions in this phase?

d. How are the ions distributed in the spinel structure?

e. Will the material be an electron or hole conductor?

11.34 Gallium arsenide has an electron effective mass of 0.067 m_e and a hole effective mass of 0.082 m_e.

a. Calculate the dimension at which quantum confinement becomes significant for n-type and p-type gallium arsenide at 300 K. The crystal structure of gallium arsenide consists of layers of GaAs each 0.327 nm in thickness.

b. How many layers are needed for quantum confinement of electrons or holes?

11.35 Gallium phosphide has an electron effective mass of 0.82 m_e and a hole effective mass of 0.60 m_e.

a. Calculate the dimension at which quantum confinement becomes significant for n-type and p-type gallium phosphide. The crystal structure of gallium phosphide consists of layers of GaP each 0.315 nm in thickness.

b. How many layers are needed for quantum confinement of electrons or holes?

11.36 Compare the energy of the $n = 1$ energy level in the three directions for a quantum well of dimensions 1 cm × 1 cm × 10 nm, in a material in which the effective mass of the electron is 0.1 m_e.

11.37 Determine the energies of the lowest three states in a gallium arsenide–aluminium arsenide quantum well structure in which the potential well has a width of 9.8 nm, corresponding to 30 GaAs layers. The electron effective mass is 0.067 m_e.

11.38 Determine the energy of the $n = 1$ energy level in a fragment of n-type gallium nitride 20 nm × 393 nm × 1700 nm. The effective mass of electrons in this material is 0.19 m_e.

11.39 The critical field of the superconductor $PbMo_5S_6$ is given as 60 T. What is the critical field in A m^{-1}?

11.40 The critical field, $\mathbf{H}_{c2}$, for the type II superconductor Nb_3Sn at 15 K is given as 7 T. Estimate the value of the critical field at 0 K. The superconducting transition temperature, T_c is 25 K.

11.41 What value of the critical field, $\mathbf{H}_{c2}$, will cause the superconductivity of the type II superconductor Nb_3Ge to be lost at 20 K. The value of the critical field at 0 K is given as 37 T and the superconducting transition temperature is 23.6 K.

11.42 The superconductor $Nd_{2-x}Ce_xCuO_4$ is derived from the insulator Nd_2CuO_4 by substitution of some Nd^{3+} by Ce^{4+}.

a. What are the charges on the Cu ions present?

b. Will the superconductivity be via holes or electrons?

11.43 The compound $La_2SrCu_2O_6$ can take up oxygen to a composition $La_2SrCu_2O_{6.2}$.

a. What are the charges on the Cu ions present in each of these phases?

b. Are either potential high temperature superconductors, and if so, would the superconductivity be via holes or electrons.

Notes

1 A soliton is, in general, a single wave or pulse that keeps its shape and travels at a constant velocity. A well-known example is the Severn bore, a single wave that is created by tidal influences and travels up the River Severn against the stream at a constant velocity of between 8 and 12 miles h^{-1}.

2 Values of the critical field are often quoted in tesla, the unit of magnetic induction, not the unit of magnetic field, A m^{-1}. In these cases equate the 'field' in T to $\mu_0 H$, where H is the true field, measured in A m^{-1}.

12
Optical Aspects of Solids

- What are lasers?
- Why are thin films often brightly coloured?
- How do solar cells work?

Optical properties of materials describe the interaction of light and matter. There are two main strands to this subject: the processes that cause light to be emitted by a solid and the interactions that take place when light falls on a solid. The description of these aspects depends upon the scale of the phenomena. At ordinary dimensions, light interactions with a solid, typically involving scattering, reflection, and refraction, in which there is little or no energy loss, are most easily discussed by treating light as a wave. On the other hand, in the production and emission of light by atoms or molecules, where energy exchange is crucial, light is best regarded as a stream of particles called photons. The chapter opens with a description of these two aspects.

12.1 Light

12.1.1 Light Waves

Light, considered as a wave, forms part of the *electromagnetic spectrum*, ranging continuously from radio waves, with wavelengths of Gm, to high energy cosmic rays, with wavelengths of the order of fm (Figure 12.1). For convenience different frequency ranges are given broad-brush names, for example, X-rays or microwaves, although the boundaries between regions are blurred.

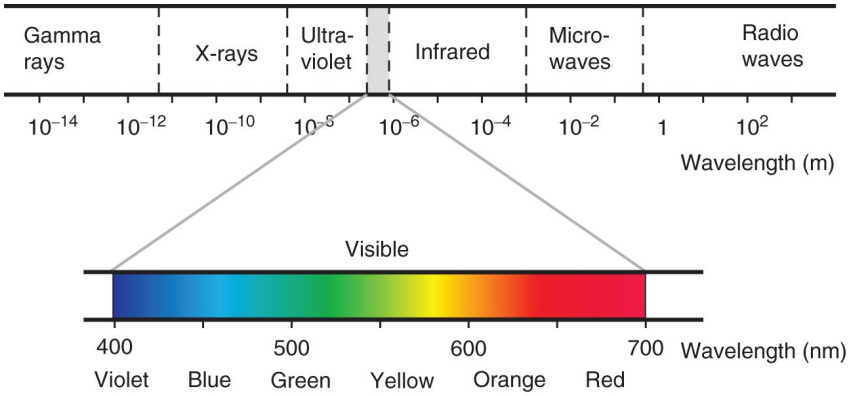


Figure 12.1 The electromagnetic spectrum; the visible portion extends from approximately 400–700 nm.

An electromagnetic wave has an electrical and magnetic component, consisting of an oscillating electric and magnetic field, each described by a vector. As far as the topics in this chapter are concerned, the magnetic component need not be considered and only the electric component, specified by an electric field vector, needs to concern us.

A light wave moving to the right can be represented by the equation:

$$\mathscr{E}_y = \mathscr{E}_0 \cos\left[(2\pi/\lambda)\,(x - v\,t) + \phi\right]$$

where $\mathscr{E}_y$ is the *magnitude* of the electric field vector along the *y*-axis at position *x* and time *t* and $\mathscr{E}_0$ is the *amplitude* of the wave, which is the maximum magnitude of the electric field. The *argument* of the cosine function $[2\pi/\lambda(x - v\ t) + \varphi]$ is called the *phase* of the wave and φ is the *initial phase* of the wave. The peaks in the wave are referred to as *crests* and the valleys as *troughs* (Figure 12.2). The separation between two adjacent peaks or two adjacent troughs is called the *wavelength* λ. The *velocity* v is called the *phase velocity* and designates the velocity at which any peak or trough moves. The phase velocity of a light wave, v, is related to the (*temporal) frequency*, *ν*, of the wave by the equation:

$$v = \lambda\nu$$

The velocity of light in a vacuum, $2.997\,92 \times 10^8\ \mathrm{ms}^{-1}$, which has a special significance in physics, is given the symbol *c*. The velocity of light in anything other than a vacuum is less than *c*.

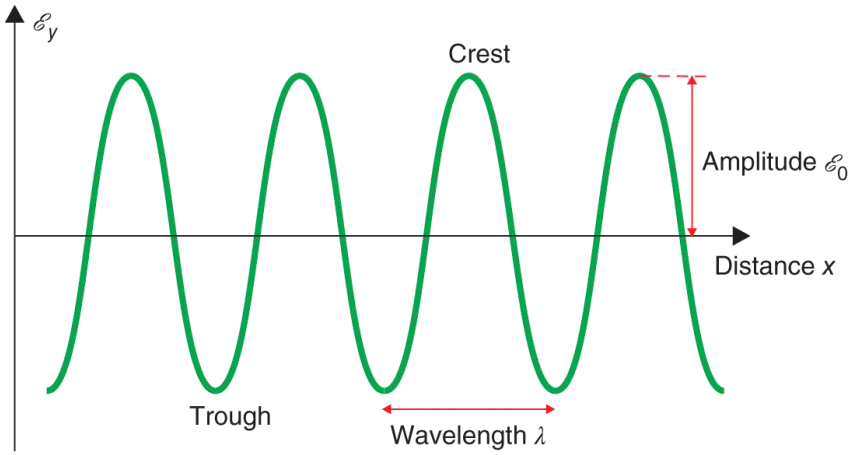


Figure 12.2 The electric field of a light wave represented as a sinusoidal wave of amplitude $\mathscr{E}_0$ and wavelength λ.

The part of the electromagnetic spectrum detected by human eyes is called the *visible spectrum*. Perception of the different wavelengths is called *colour*. The shortest wavelength of light that an average observer can perceive corresponds to the colour violet, $\lambda = 400$ nm, and the longest wavelength of light perceived by an average observer corresponds to the colour deep red, $\lambda = 700$ nm. Between these two limits, the other colours of the spectrum occur in the sequence red, orange, green, blue, indigo, and violet (Table 12.1).

Table 12.1 The visible spectrum.

Colour	λ (nm)	ν (Hz)	Energy (J)	Energy (eV)
Infrared	750	4.00×10^{14}	2.65×10^{-19}	1.65
Deep red	700	4.28×10^{14}	2.84×10^{-19}	1.77
Orange-red	650	4.61×10^{14}	3.06×10^{-19}	1.91
Orange	600	5.00×10^{14}	3.31×10^{-19}	2.07
Yellow	580	5.17×10^{14}	3.43×10^{-19}	2.14
Yellow-green	550	5.45×10^{14}	3.61×10^{-19}	2.25
Green	525	5.71×10^{14}	3.78×10^{-19}	2.36
Blue-green	500	6.00×10^{14}	3.98×10^{-19}	2.48
Blue	450	6.66×10^{14}	4.42×10^{-19}	2.75
Violet	400	7.50×10^{14}	4.97×10^{-19}	3.10
Ultraviolet	350	8.57×10^{14}	5.68×10^{-19}	3.54

The visible spectrum is bounded by the *infrared* at long wavelengths and the *ultraviolet* at short wavelengths. The longer wavelengths of infrared radiation, called *thermal infrared*, are detectable as the feeling of warmth on the skin. *Ultraviolet A* (UVA) is closest to the violet region and although invisible to humans can be seen by many animals. *Ultraviolet B* (UVB) and *ultraviolet C* are at shorter wavelengths, and are energetic enough to damage biological cells.

A beam of light is said to be *monochromatic* when it is composed of only a very narrow range of wavelengths, and *coherent* when all of the waves that make up the beam are completely *in phase*, that is, the crests and troughs of the waves are in step. Normal light is *incoherent* and laser light is coherent. The *polarisation* of the wave indicates the direction of the electric field vector of a light beam. Light emitted by most sources is *unpolarised*, as the orientation of the electric field vector changes at random every 10^{-8} seconds or so. A light beam is *plane* (or *linearly*) *polarised* if the electric field vector always remains in the same plane and *circularly polarised* when the vector tip traces out a circle. Circular and

linearly polarised light are special cases of elliptical polarisation, where the tip of the electric field vector traces out an ellipse. Laser light is usually polarised.

Many of the effects of the interaction of light with solids can be explained in terms of *interference* between light waves. If two light waves occupy the same region of space at the same time, they can add together, or *interfere*, to form a product wave (Figure 12.3). If two *identical* waves are exactly in step, then they will add to produce a resultant wave with twice the amplitude by the process of *constructive* interference. If the two waves are out of step, the resultant amplitude will be less, due to *destructive* interference. If the waves are sufficiently out of step that the crests of one correspond with the troughs of the other, the resulting amplitude will be zero.

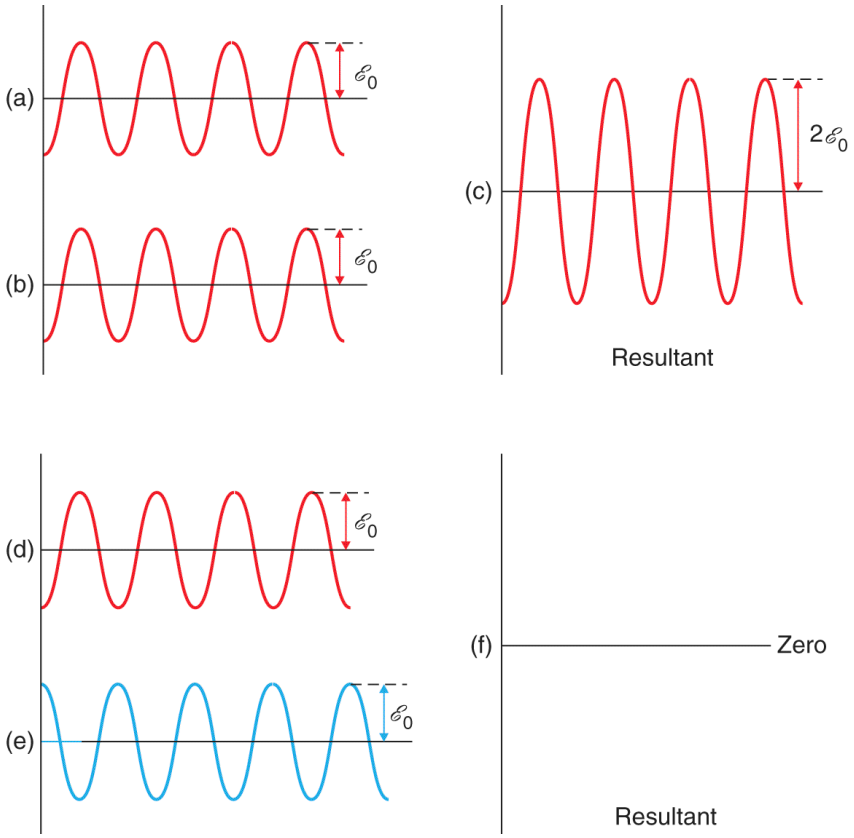


Figure 12.3 The interference of light waves: (a, b) two waves in step add to give a wave with double the amplitude of the original waves (c); (d, e) two waves out of step add to zero (f).

12.1.2 Photons

When light interacts with individual atoms or molecules, it is best regarded as a stream of particle-like *photons* of energy E given by:

$$E = h\nu = hc/\lambda$$

where h is Planck's constant and c is the velocity of light in vacuum. The wave and particle descriptions are linked by the fact that the wave equations can be considered to describe the statistical behaviour of large numbers of photons. The relationship is formalised by equating ν to the equivalent *wave frequency* and λ to the equivalent *wavelength*. The relationship between the wavelength and the frequency, for waves or photons travelling in vacuum, is:

$$\nu\lambda = c$$

The interaction of a photon and an atom is mediated by the electrons present. These occupy energy levels associated with atomic or molecular orbitals or energy bands (Chapter 1). When a photon

encounters an atom or a molecule, the energy will be absorbed only if the photon energy precisely matches the energy difference between the (generally) highest of the occupied energy levels and one of the empty higher energy levels. The energy of the absorbed photon E_a is given by:

$$E_a = h\nu = E_2 - E_1$$

where E_1 represents the energy of the initial state and E_2 the energy of the final (higher energy) state (Figure 12.4a). Under normal circumstances atoms or molecules in solids are normally in the lowest energy state, called the *ground state*, written as E_0, and the photon energy is equal to the energy between the ground state and a higher energy state. If the photon energy does not correspond exactly to the energy level separation between the occupied state and an upper energy state, it will not be absorbed, but may be scattered.

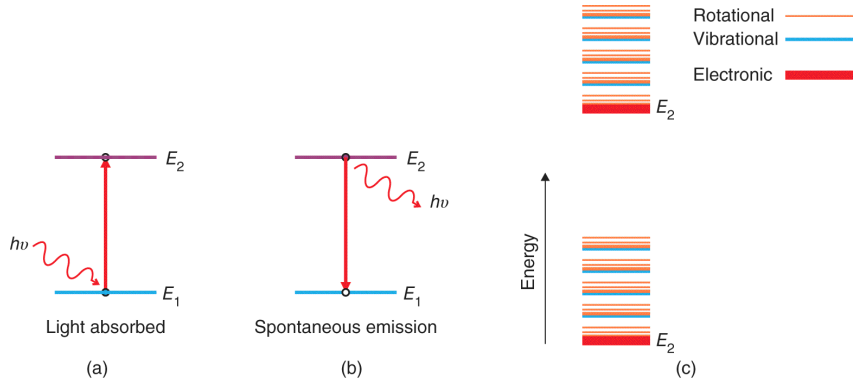


Figure 12.4 Absorption and emission of radiation: (a) light absorption; (b) light emission; (c) electron energy levels in a molecule.

When an atom or molecule in a solid is in a high-energy state, it may lose energy by dropping to a lower energy state by emitting a photon. This will carry off the excess energy. If the same energy states as in absorption are involved, the photon energy will be identical to that of the photon that was originally absorbed, so that the energy of the photon emitted, E_e, is given by:

$$E_e = h\nu = E_2 - E_1$$

(Figure 12.4b). This process is called *spontaneous emission*. Ultimately, by emitting one or more photons, an atom or molecule will return to the ground state. Because the energy levels are sharp, the photons have precise energies, and the emission spectrum will consist of one or more sharp lines. Colour arises when electrons make transitions between energy levels with a separation of approximately 2–3 eV (4×10^{-19} J, Table 12.1). Molecules have more complex energy level arrangements than atoms. Each electronic energy level is associated with energy levels due to molecular vibration, giving rise to infrared photons (approximately 0.37 eV; 6×10^{-20} J), and each vibrational energy level is associated with energy levels due to molecular rotation, giving rise to microwave photons (0.0037 eV; 6×10^{-22} J) (Figure 12.4c).

Although a photon may have the correct energy to be absorbed, the *probability* of the transition occurring is governed by quantum mechanical *selection rules*. When there is a high probability of a transition occurring, it is said to be *allowed*. Colours caused by the absorption or emission of photons in allowed transitions, such as those that occur in dye molecules, are strong. In cases where the transitions are forbidden, equivalent to a low probability of occurring, very few photons will be absorbed or emitted. Colours arising from forbidden transitions are weak.

12.1.3 Colour and Appearance

The colour of a light source, such as a display screen, depends upon the wavelength of the light emitted by the source. If all wavelengths are present in similar proportions to that in sunlight, the colour is described as white. Other colours are perceived when several wavelengths are mixed, called *additive*

colouration. The colours of display screens are produced by additive colouration. The majority of additive colours can be achieved by mixing three *additive primary colours*, red, green, and blue. When these are mixed in the appropriate amounts, white light is perceived.

The colour of a solid depends upon the light that falls on the material and then travels to the observer. When white light falls onto such a solid, some wavelengths may be absorbed, reflected, scattered, or passed through the material. Smooth surfaces reflect light uniformly, known as *specular reflection*. Rough surfaces will reflect incident light in all directions, called *diffuse* reflection. The appearance of a solid will depend upon which of these processes dominate. For example, green leaves are so coloured because the chlorophyll molecules present absorb red and blue light, allowing the green component to be reflected. However, the many different shades and textures of leaves are often centred upon the leaf surface and depends upon the proportion of specular versus diffuse reflection. Similarly, the difference in appearance of skin and a similarly coloured plastic film is one of texture. The realistic rendition of objects showing different amounts of specular and diffuse reflection is a difficult task for computer graphics, and much effort has been devoted to this objective.

Stained glass contains absorption centres, often transition metal cations, that selectively absorb some wavelengths of incident white light, allowing other wavelengths to be transmitted and to impart colour to the material. Printer inks selectively absorb some wavelengths in the incident white light using dyes, with the remainder being returned to the eye as colour. This method of generation of colours by removal of some wavelengths present in white light is called *subtractive colouration*. There are three subtractive primary colours, yellow, magenta, and cyan, which, when mixed in the correct proportions, absorb all colours to produce black. For colour printers, this tone is not dark enough for good renditions and black toner is added to the trio.

The appearance of even a simple transparent solid may vary due to scattering. For example, a transparent crystal will appear white when powdered because no light is absorbed but the numerous crystallite surfaces reflect light of all wavelengths equally. Voids within glass produce the same effect and if there are sufficient numbers, the glass will appear milky white. Opal glass is deliberately fabricated to contain large concentrations of internal surfaces and appear uniformly white. Many thermoplastic materials consist of crystalline regions embedded in amorphous material; behaviour typified by polythene. The crystalline regions scatter light, which is why films appear milky.

Metals are characterised by free electrons that are able to absorb any light falling on the surface. However, the excited electrons rapidly fall back to lower energy levels and most of the light is re-emitted. Because of this a smooth metal surface reflects light and looks shiny. Powdered metals repeatedly reflect light, but as a little is absorbed at each interaction and the surface is rough, most finely divided metals look black. The coloured metals, notably copper and gold, arise from the particular band structure of these elements, together with relativistic effects in the case of gold.

12.2 Sources of Light

When a solid emits light, energy needs to have been supplied to raise the atoms into higher energy levels before they can subsequently lose energy as light photons. The earliest way in which light emission from solids was achieved was via incandescence, where the energy was originally supplied by heating and latterly by electric current. In nature, light emission via luminescence or phosphorescence is powered by biologically important chemical reactions. More recently, artificial light emission, varying from room and street lighting to electronic displays, is powered by electricity.

12.2.1 Incandescence

Incandescence is the emission of light by a hot body. When light from an incandescent object is spread out according to wavelength the result is the continuous fan of colours listed in Table 12.1, called a *continuous spectrum*. However, the radiation emitted extends over a range of wavelengths much broader than the visible spectrum and is both incoherent and unpolarised. For a solid body a little above room temperature all the wavelengths of the emitted energy lie in the infrared and are discernible as a sensation of warmth. As the temperature increases, the overall energy of the radiation increases and the peak of the wavelength range moves towards shorter wavelengths. At a temperature of about 700 °C, the shortest wavelengths emitted creep into the red end of the visible spectrum. The colour of the emitter is seen as red and the object is said to become *red hot*. At higher temperatures, the wavelengths of the radiation given out extend increasingly into the visible region and the colour observed changes from red to orange and thence to yellow. When the temperature of the emitting object reaches about 2500 °C, all visible wavelengths are present and the body is said to be *white hot*. The most important incandescent object for us is the sun, which is the ultimate source of energy on Earth. The solar spectrum has a maximum near 560 nm. Light is perceived as *white* if it has a make-up like that of the solar spectrum.

The intensity of the radiation emitted by incandescent solids can be understood in terms of a *black body*, which is an object that absorbs and emits all wavelengths perfectly. A graph of the amount of radiation issuing from a black body as a function of wavelength is called a *black body emission spectrum* (Figure 12.5). The shape of the curve is found to be dependent only upon the temperature of the body. As the temperature increases, the peak in the curve moves to shorter wavelengths (higher energies). The solar spectrum has a form quite similar to the emission spectrum of a black body with temperature of about 5700 °C (about 6000 K).

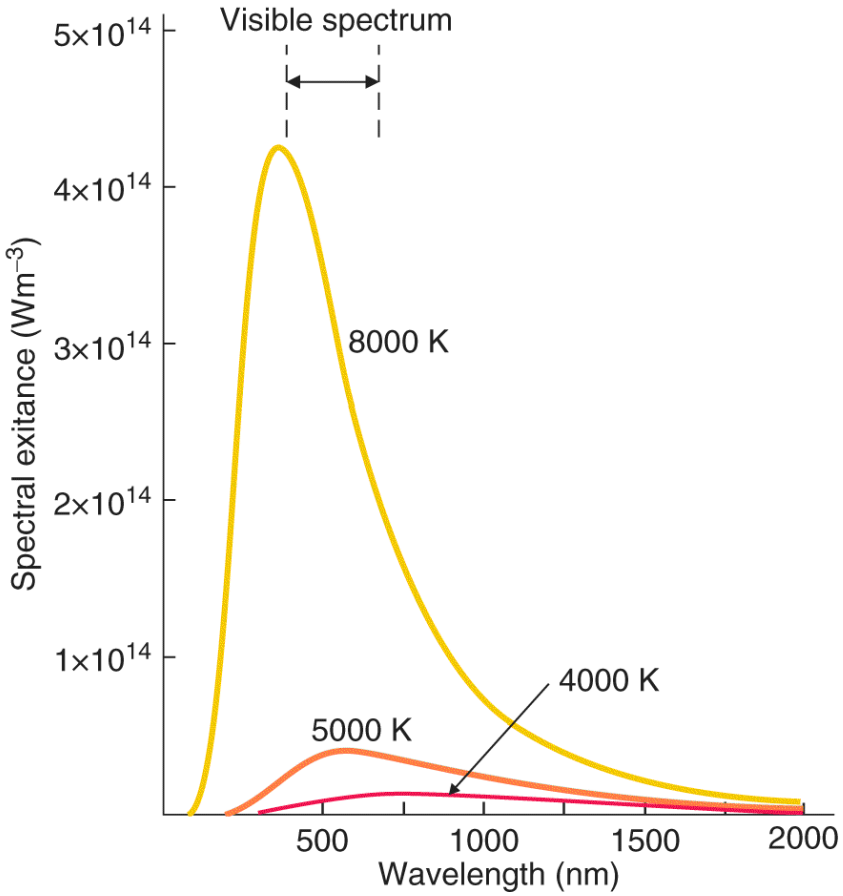


Figure 12.5 Black body emission spectra.

The form of the black body emission spectrum cannot be derived by classical physics, and its successful theoretical explanation, by Planck in 1901, initiated quantum theory. In order to reproduce the form of the curve, Planck postulated that the energy contained within the black body had to be expressed in packets or *quanta* of ε, 2ε, 3ε, ... The energy was said to be *quantised*. The relationship between the energy of a single quantum ε, and the frequency of the radiation ν, is given by:

$$\varepsilon = h\nu$$

where h is a constant (now known as Planck's constant). With this assumption, Planck was able to determine the energy density within a black body at equilibrium. It is then possible to derive a formula for the energy of the radiation emitted by a pinhole in a black body not big enough to disturb the thermal equilibrium. At a temperature T in the wavelength interval, λ to $\lambda + \delta\lambda$, the *spectral exitance*, I_λ, is given by:

$$I_\lambda = 2\pi hc^2/\{\lambda^5[\exp(hc/\lambda k_B T) - 1]\}$$

The form of this expression matched the experimental observations perfectly.

Until the end of the twentieth century, most artificial illumination was provided by incandescent sources, firstly by burning oil or fat which raised carbon particles to high temperatures (*oil lamps*, *candles*), then gas, which used a gas flame to heat a light emitting 'mantle' made from inorganic oxides (*gas light*) and most recently electric currents, which are used to heat fine tungsten filaments (*incandescent light bulbs*, *halogen* light bulbs). It is clear (Figure 12.5) that the amount of light energy emitted by these sources is only a fraction of the total energy released and these light sources were both hot (a fire risk) and inefficient. They have now largely been replaced by other forms of artificial lighting, described in the following text.

12.2.2 Luminescence

The emission of radiation by solids at relatively low temperatures, 'cold light', is called *luminescence*. For luminescence to occur energy must be supplied to the solid in some form or another and the broad-brush term 'luminescence' is divided into a number of sub-categories that mirror this dependency (Table 12.2). For example, photoluminescence is light emission brought about by the absorption of high energy photons, typically ultraviolet. The most widely utilised form of photoluminescence is *fluorescence*, in which light emission is *immediate*, taking place via allowed transitions. Laser action is a form of fluorescence. *Phosphorescence* is similar but is typified by the *slow* conversion of the exciting energy into light, so that light emission is *delayed*, often by considerable lengths of time, because the light emitting transitions are forbidden.

Table 12.2 Luminescence.

Type	Definition	Source of energy
Bioluminescence	Luminescence in a living organism	Gibbs energy of chemical reactions
Cathodoluminescence	Luminescence due to electron bombardment (cathode 'rays')	Electron kinetic energy
Chemiluminescence	Luminescence during a chemical reaction	Gibbs energy of chemical reactions
Electroluminescence	Luminescence resulting from the application of an electric field (can be equivalent to cathodoluminescence)	Electrical potential energy
Electrochemiluminescence (Electrogenerated chemiluminescence)	Luminescence via the chemical reaction of one or two molecules excited by the application of an electric potential	Electrical potential energy
Photoluminescence	Luminescence after irradiation by visible or ultraviolet light	Ultraviolet and visible photon energies
Fluorescence	Rapid decay of excited state	
Phosphorescence	Slow decay if excited state	
Radioluminescence	Luminescence as a result of radioactivity	Energetic particles and γ rays[a]
Sonoluminescence	Luminescence as a result of sound waves	Acoustic vibrations, pressure waves
Thermoluminescence	Luminescence following an increase of temperature	Thermal energy
Mechanoluminescence	Luminescence following deformation or friction	Strain energy, thermal energy, chemical bond energy
Triboluminescence	Luminescence following grinding, crushing	Chemical bond energy

[a] Materials that emit light in response to the impact of high energy particles or X rays are often termed *scintillators* rather than phosphors.

Irrespective of the source of the energetic exciting radiation, the light emitting solids are called *phosphors*. Phosphors consist of a crystal matrix that acts as a (nominally inactive) host containing a small quantity of an incorporated *activator*, often in the form of substitutional defects. The role of the host structure or of the host–activator combination is to absorb energy. The activator re-emits part of the excitation as a photon. The colour emitted is dependent upon the nature of the activator.

Sometimes it is found that the host–activator pair cannot absorb the exciting radiation directly, in which case a helper species, a *sensitiser*, is needed as well. In this case the sensitiser absorbs the exciting photons and passes the energy to the activator.

12.2.3 Fluorescent Lamps

Fluorescent lamps utilise photoluminescence for light generation. Fluorescent lighting for advertising was first used in 1925. Development of phosphors during the 1930s led to the commercial introduction of low voltage fluorescent lamps in 1939. The intensity of the luminescence is roughly proportional to the amount of phosphor that is exposed to exciting radiation. Early phosphors were not especially efficient, and the first fluorescent lamps were in the form of tubes about 1 m in length. Improvements in phosphor specification have made the efficiency greater and since the 1980s compact fluorescent light tubes were produced.

Fluorescent lamps contain an inert gas and a small quantity of mercury vapour at a low pressure. Under electron bombardment from the current passing through the lamp, the mercury atoms are excited and emit copious ultraviolet radiation. This consists mainly of the wavelengths 185, 254, and 365 nm, as well as some radiation in the visible. Conversion of the ultraviolet radiation to visible is by way of a phosphor coated onto the inside of the tube (Figure 12.6). The various forms of fluorescent lamps available differ in the type of phosphor coatings used.

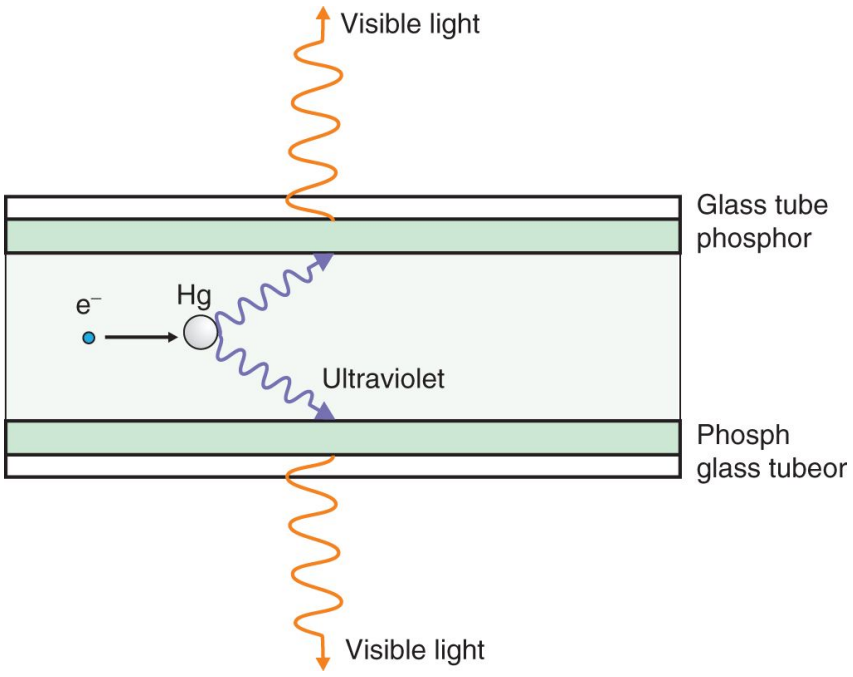


Figure 12.6 Fluorescent lamps: electrons from the cathode collide with mercury (Hg) atoms to emit ultraviolet wavelengths which are converted to visible wavelengths by the phosphor coating.

Halophosphate lamps use modified calcium fluorophosphate, $Ca_5(PO_4)_3F$ (the mineral fluorapatite) as the host matrix. When doped with Sb^{3+} ions as activator (written $Ca_5(PO_4)_3F{:}Sb$), a blue emission is produced. The Sb^{3+} ions, which exist as isolated point defects in the fluorapatite matrix, absorb via an s^2 to s^1p^1 transition centred at 254 nm, which closely matches the mercury vapour output. A minor problem with Sb^{3+} is that the blue emission gives the lamps a rather cool colour. If Mn^{2+} is also incorporated into the system as a *coactivator*, a warmer tone is produced as this ion produces an orange-red emission. Variation in the proportions of Sb to Mn varies the tone of the light.

The Mn^{2+} and Sb^{3+} ions occupy at random the Ca^{2+} positions in the host matrix. While Mn^{2+} incorporation will not pose an electroneutrality problem, as the Mn^{2+} ions have the same charge as the Ca^{2+} ions that they replace, this is not so with Sb^{3+}. The introduction of Sb^{3+} ions into the phosphate will result in an internal charge imbalance, which will degrade performance. To overcome this, charge balance is maintained by adding one F^- or Cl^- ion to the phosphate for each Sb^{3+} ion present. It has been found that an empirically derived composition for the host matrix of $Ca_{10}P_6F_{1.8}Cl_{0.2}O_{24}$ is most satisfactory.

Trichromatic (Colour 80) lamps improve the spectral balance compared with halophosphate lamps by using a phosphor mixture, which emits equal amounts of the colours red, blue, and green. The favoured red emitter in trichromatic lamps is Eu^{3+} as activator doped into a (matrix) Y_2O_3 matrix, Y_2O_3:Eu, with the Eu^{3+} ions forming point defects on the Y^{3+} sites. The main transition leads an emission wavelength near to 611 nm. The green emission is from Tb^{3+} as activator and is found at a wavelength close to 540 nm. Host matrices are La(Ce)PO_4, LaMg(Ce)$Al_{11}O_{19}$, and La(Ce)MgB_5O_{10}. In each case the Tb^{3+} ions replace La^{3+} ions and no charge compensation is needed. The blue emission is produced by Eu^{2+} ions as activator. Unlike Eu^{3+} and Tb^{3+}, in which the colour transitions are between sharp energy levels derived from the f electrons on these ions, the transition in Eu^{2+} involves f and d electron energy levels. The d electron energy levels are more influenced by the surrounding crystal structure than the f levels, and the luminescent colour of the Eu^{2+} centre will be modified by changing the site in the host lattice and the type of host structure. The emission spectrum of the usual tricolour lamp phosphor, $BaMgAl_{10}O_{17}$:Eu, has a maximum at 450 nm. As with the fluorophosphates lamps, the overall emission colour can be modified by changing the relative amounts of the three phosphors present so as to emphasise the red, green, or blue ends of the spectrum.

The colour spectrum of the fluorescent lamps described earlier, although satisfactory for many purposes, does not give an accurate impression of the colour of an object compared to that perceived when the same object is viewed in daylight. To overcome this, deluxe (*Colour 90*) lamps can be used. These employ modified phosphors so that the emissions are shifted slightly and a fourth phosphor is added to the blend. This phosphor, $Y_3Al_5O_{12}$ doped with Ce^{3+}, absorbs some of the blue-violet light emitted by Eu^{2+} and emits yellow light in its place.

Mercury lamps for street lighting use a high pressure of mercury vapour and produce an emission that is more or less continuous between the limits of 250 and 550 nm. This output is unbalanced from a visual viewpoint and it is desirable to introduce an ultraviolet absorbing phosphor that will emit in the red, so as to balance the output. A favoured phosphor for this purpose is a mixed strontium magnesium phosphate using Sn^{2+} as activator $(Sr,Mg)_3(PO_4)_2$:Sn^{2+}, which emits at 630 nm. *Sun tanning beds* also make use of phosphors, but in this case the main output is required to be in the ultraviolet. Ultraviolet A, wavelength range 320–400 nm, and ultraviolet B, wavelength range 280–320 nm, are both used for this purpose. Initially sun bed tubes used $SrMgSi_2O_7$:Pb in which Pb^{2+} is the activator. These gave a broad emission centred on 350 nm and spanning both UVA and UVB. However, this material has a low stability and was later replaced by $BaSi_2O_5$ with Pb^{2+} activator, which gave a narrower band centred at 350 nm, limited to UVA. Today sun beds often use tubes containing a mixture of $BaSi_2O_5$:Pb and $SrAl_{12}O_{19}$ containing Ce^{3+} activator, which has an emission peak centred at approximately 310 nm, thus providing some UVB output.

12.2.4 Light Emitting Diodes (LEDs)

A key feature of a p–n junction diode is the potential barrier that builds up in the junction region (Section 11.2.6). When a positive voltage, called a forward bias, is applied to the p-type side of the junction, electrons and holes enter the junction and recombine. The energy released is approximately equal to the band gap. For this to appear as light, turning the diode into a *light emitting diode* (*LED*), the band gap energy must match visible wavelengths (Figure 12.7).

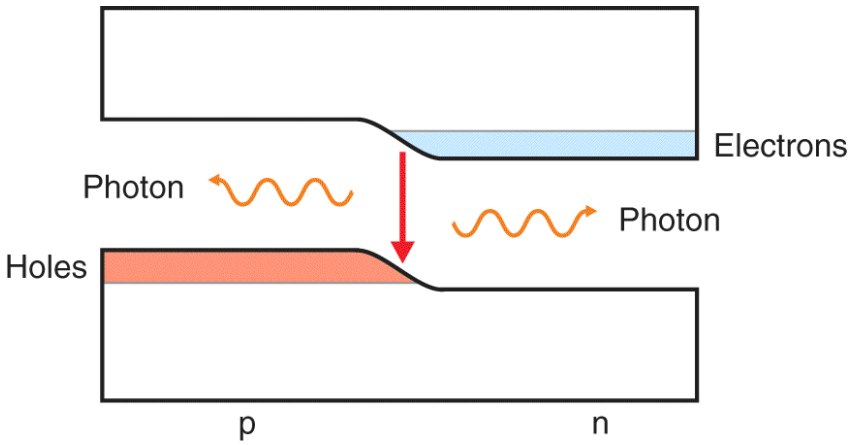


Figure 12.7 The principle of LED operation. Electrons and holes combine in the junction region and emit visible light.

To be useful, LEDs must give an adequate light output from a small voltage. To achieve this, the nature of the excitation of an electron from the valence band to the conduction band and reverse process of annihilation must be efficient. This efficiency depends upon the detailed band structure of the semiconductors, and the flat band model (Chapter 11) is not adequate. In effect the bands must be described as a series of undulating surfaces.

An important aspect for light absorption and emission by a semiconductor is the correspondence between the structures of the valence and conduction bands. When the lowest point of the valence band corresponds to the highest point of the conduction band, that is, they are at the same value of the wave vector, $\mathbf{k}_1$ (Figure 12.8a,b), an optical transition between the bands can take place without a change of wave vector (or momentum). Such a transition is called a *direct transition*, represented by the equation:

$$h\nu = E_g$$

where $h\nu$ is the energy of the photon absorbed or emitted and E_g is the optical band gap. In terms of quantum mechanics, the transition has a high probability of occurring when a photon of the correct energy hits the semiconductor, the efficiency of the process is high and direct band gap materials make good light emitters.

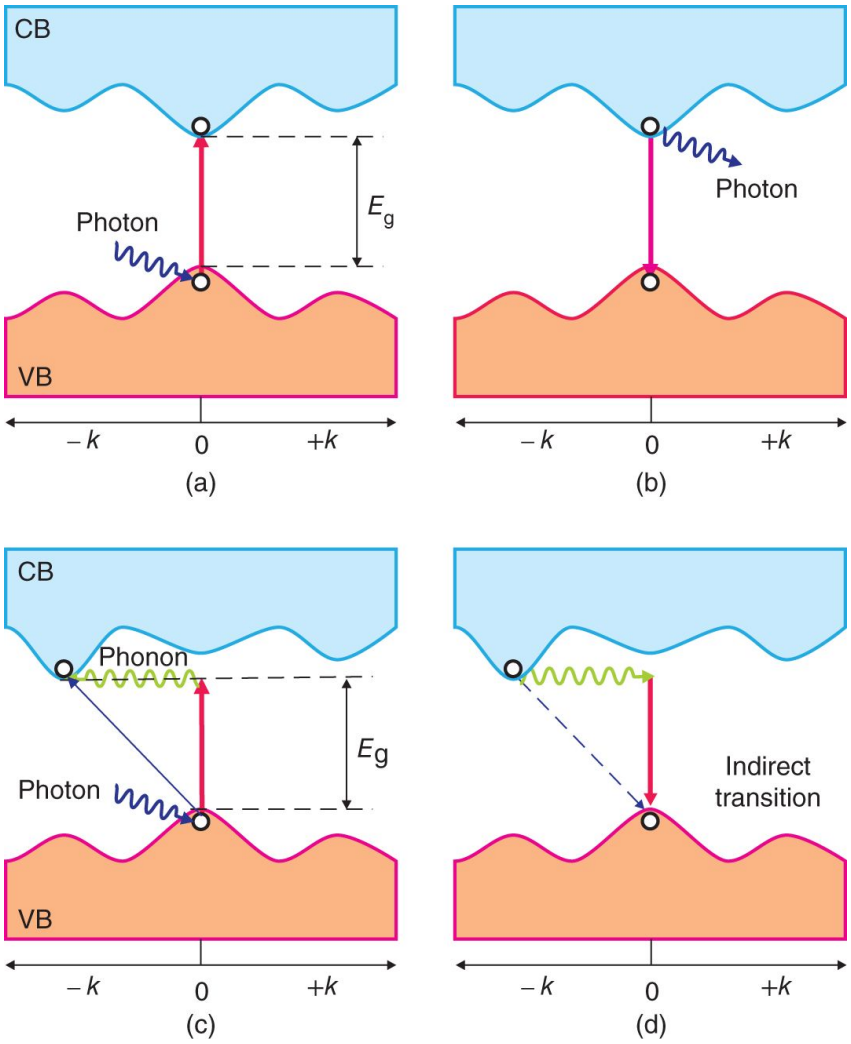


Figure 12.8 Photon absorption and emission in direct (a, b) and indirect (c, d) band gap materials.

Alternatively, the lowest point of the valence band is not at the same value of the wave vector, **k**, as the highest point in the conduction band (Figure 12.8c,d). In this case, the photon that is to promote the electron from the top of the valence band to the conduction band must also interact with the lattice to pick up (or lose) sufficient momentum to make the transition possible. To do this, the photon must interact with phonons (lattice vibrations), so as to complete this energy adjustment. Such a transition is called an *indirect transition*, represented by the equation:

$$h\nu = E_g \pm h\nu_p$$

where $h\nu_p$ is the energy of the phonon involved. The ± term depends upon whether the phonon is absorbed or emitted. The probability of an indirect transition occurring is quite low and an indirect band gap material makes a poor light emitter.

The energy band structures of crystalline silicon and germanium show that an indirect transition will occur and for this reason neither of these materials is really suitable for LEDs. (Note though, that amorphous silicon seems to show a direct transition.) The energy band structure of gallium arsenide favours a direct transition, and for this reason GaAs was the first material to be used for LEDs.

The band gap of pure GaAs, 1.42 eV, gives an emission in the infrared. However, this is moved towards the visible by alloying with a similar semiconductor with a bigger band gap. A suitable semiconductor for mixing with GaAs was found to be gallium phosphide (GaP), with a band gap of 2.26 eV. Reaction

with GaAs forms $GaAs_xP_{1-x}$ alloys. To a first approximation the band gap of the alloy is a linear function of the band gap of the two end members so that an alloy with x greater than 0.42 will emit in the visible. Unfortunately, these only preserve the direct band gap of pure GaAs over about half the composition range, which allowed red and orange-red LEDs to be fabricated but not green or blue ones. To improve efficiency, LEDs that emit in this wavelength range are now fabricated from alloys of four semiconductors, GaAs, GaP, aluminium phosphide (AlP), and indium phosphide (InP). In 1993, semiconductor LEDs made with gallium nitride (GaN) and indium nitride (InN) gave green and blue emissions. Solid solutions based upon the semiconductors GaAs, GaP, GaN, AlP, and aluminium arsenide (AlAs) have since been formulated that give light emission across the visible spectrum.

LEDs generate light of specific colours. Impressive full colour displays using LEDs have been built using millions of small red, green, and blue units. However, using three LED colours to give the impression of white light does not give a particularly good colour rendition, and four colours, red, yellow, green, and blue are better. The displays have no size limitations, and the intensity produced by each LED is also sufficient to make them easily visible in daylight. While white light generation by way of additive coloration is acceptable for displays, it is not satisfactory for most indoor lighting purposes and for this the commonest way to make white light is to coat a monochromatic LED with a phosphor. It is these devices that are normally termed 'white LEDs'. The commonest white LED configuration is to use a high intensity (Ga,In)N blue emitting LED coated with a yellow emitting phosphor. At present this is most often yttrium aluminium garnet (YAG), $Y_3Al_5O_{12}$, doped with Ce^{3+}, which gives a rather broad yellow emission. The coating is sufficiently thin to allow a certain amount of blue light to be transmitted. This, together with the yellow luminescence, creates a cool bluish white colour. The rather cold light is not entirely suitable for indoor lighting, which ideally needs to be warmer. For this, two phosphors are used, a red and a green emitter in tandem with a (Ga, In)N blue emitting LED. The orange – red phosphor is the nitrosilicide $Sr_2Si_5N_8$ doped with Eu^{2+}. The red–orange tone can be adjusted by replacing some of the Sr by Ca. The green phosphor used is often $SrSi_2O_2N_2$ doped with Eu^{2+}. Once again, colour tuning is possible by substitution of the Sr by Ca. Blue light is contributed by the LED itself.

12.2.5 Organic Light Emitting Devices/Diodes (OLEDs)

Organic molecules are generally insulators, with electrons occupying molecular orbitals confined to the spatial region of the molecule. For the purposes of understanding the optical properties, attention needs only to be focussed on the highest occupied molecular orbital (the HOMO) and the energetically nearest orbital above it, the lowest unoccupied electron orbital (LUMO). The unoccupied, higher energy orbitals, LUMOs, are generally antibonding orbitals, and the filled orbitals, HOMOs, are generally bonding orbitals. In small organic molecules these are well separated and organic electroluminescence – the generation of light by passing an electric current through an organic solid, therefore, is unlikely. However, in larger molecules and polymers the LUMO and HOMO orbitals can broaden into structures akin to a higher empty energy band and a lower energy filled band (Figure 12.9a) leading to the possibility of both electronic conduction (Section 11.4) and electroluminescence. Such materials are often called organic semiconductors.

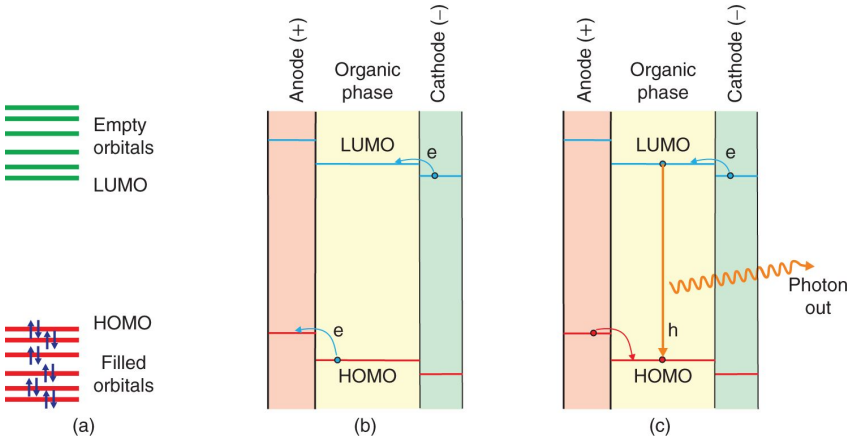


Figure 12.9 Principle of an organic electroluminescent device (OLED); (a) the relative dispositions of a filled HOMO and an empty HOMO; (b) an applied voltage introduces electrons into the HOMO and holes into the LUMO, which (c) recombine and emit light.

The molecules that can be used for electroluminescence are similar to the dye molecules that are of use in photoluminescence together with the polymers already mentioned (Section 11.4). Light emission via electroluminescence is obtained in the following way. A suitable organic film forms the light emitting layer. This is sandwiched between a cathode (the negative electrode) and an anode (the positive electrode). One electrode needs to be transparent to allow the electroluminescence to be visible. Under the influence of an applied voltage, electrons are forced into the empty LUMO, and electrons are pulled out of the filled HOMO (Figure 12.9b), this latter process often being described as saying that holes are injected into the HOMO (Figure 12.9c). The field present makes the electrons and holes drift towards each other, whence they can combine emitting a photon (Figure 12.9c).

The development of commercially successful display technologies using organic materials as the light emitting constituents took place towards the end of the twentieth century. The light emitting components are known as OLEDs (organic light emitting devices or organic light emitting diodes). The first practical OLED, using 8-hydroxyquinoline aluminium (Alq_3), was reported in 1987. This type of device, now part of a grouping of small molecule organic light emitting devices (SMOLEDs), gave out green light with a peak emission wavelength of 550 nm. In 1990 a polymer based organic light emitting device (POLED) employing poly(*p*-phenylene vinylene) (PPV) as the organic emitter was described. This gave a rather broad emission covering the yellow-green region of the spectrum. Shortly afterwards (1991) a similar device using a related polymer, poly(2-methoxy-5-2′-ethyl-hexloxy)-1,4-phenylene vinylene, abbreviated to MEH-PPV, which gave out an orange-red light was developed..

A typical early construction of an MEH-PPV OLED used a calcium cathode because the energy level of the electrons in this metal, the Fermi level, is very close to that of the LUMO in the polymer and allows easy flow of electrons across the interface. In order for light to leave the device, the other electrode must be transparent, and for this the transparent conducing material indium tin oxide (ITO) was suitable. The Fermi level of this material is similar to the HOMO of the polymer (Figure 12.10a). The completed device (Figure 12.10b) gave out an orange–red colour.

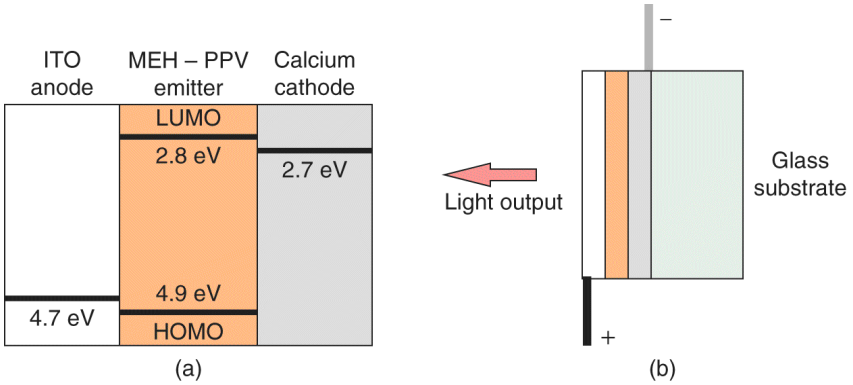


Figure 12.10 An MEH-PPV OLED: (a) energy level diagram for the anode, emitting polymer and cathode. (b) Schematic of OLED.

These early devices had a number of problems. The light output was low and device lifetimes were short. Subsequently, emitter film improvements and device architecture modifications have overcome these difficulties so that television sets that employ OLED displays are now readily available.

White organic light emitting devices (WOLEDs) for lighting purposes, in the form of panels, are also produced. Differently to displays, lighting panels need to render the perceived colour of objects 'accurately', which is to say, as perceived in outdoor sunlight. To obtain good colour rendition the output from the panel should have a spectrum similar to that of the sun itself. Moreover, the requirement of high light output means that device design, so as to optimise photon release, is crucial. As OLEDs tend to produce one colour, the strategy for producing white light is to use additive colour mixing of three primary colours. However, the flexibility of organic molecules means that alternatives are available that cannot be easily utilised by conventional lighting or LEDs. Thus, although three separate emitters can be stacked to give white light, double emitters – molecules that emit at two wavelengths – can be coupled with single emitters. As an example, a Pt-phenyl imidazole derivative has a single molecule emission of a sky-blue tone, with a peak emission wavelength of 472 nm. In addition, when the molecules are in close proximity, two adjacent excited molecules can link together to form an excimer (an excited bimolecular complex), which has a lower energy than the excited state of the monomer, with orbitals extending over both entities. The decay from these excimer units is rather broad and centred on the yellow at 592 nm. In practice, the combination of both single and double emitters produces an output which gives the appearance of white.

12.2.6 Solid-State Lasers

When an atom in an excited state makes a transition to the ground state, photons with a frequency ν will be emitted. In 1917, Einstein suggested that there should be *two* possible types of emission process rather than just one. The most obvious is that an atom in an excited state can randomly change to the ground state; a process called *spontaneous emission* (Section 12.1.2). Alternatively, a photon having energy equal to the energy difference between the two levels can interact with the atom in the excited state causing it to fall to the lower state and emit a photon at the same time; a process called *stimulated emission*. The process of stimulated emission produces light that is coherent and polarised, making it quite different to the light emitted by spontaneous emission, which is both incoherent and unpolarised. A laser is a device that produces light by stimulated emission; the word *laser* being an acronym for *Light Amplification by Stimulated Emission of Radiation*.

The key to laser action is to obtain atoms or molecules in an excited state and keep them there long enough for photons to pass and trigger stimulated emission. There are two theoretical difficulties to overcome. Under normal circumstances the number of atoms in the excited energy level, N_1, compared with those in the ground state, N_0, the *populations*, will be extremely small for energy levels, which are sufficiently separated to give rise to visible light. This may be confirmed by the Boltzmann law, which gives the relative populations under conditions of (classical) thermal equilibrium by:

$$(N_1/N_0) = \exp\left[-(E_1 - E_0)/k_B T\right]$$

where k_B is Boltzmann's constant, T is the absolute temperature, and E_1 and E_0 are the energies of the excited state and the ground state, respectively. Thus, under *equilibrium* conditions it is not possible to increase the population N_1 over N_0, that is, to achieve a *population inversion*.

A second problem compounds the difficulty and concerns the relative rates of spontaneous and stimulated emission. The rate of depopulation of an upper energy level, at energy E_1 ($-\mathrm{d}N_1/\mathrm{d}t$), by spontaneous emission, will be given by:

$$-\mathrm{d}N_1/\mathrm{d}t = A_{10}N_1$$

where the negative sign denotes that the number of atoms in the upper state, N_1, is decreasing with time. The rate is proportional to the number of atoms in the state, N_1. The rate constant, denoted here as A_{10}, is called the *Einstein coefficient for spontaneous emission*, and the subscript [10] means that the transition is from an excited state, energy E_1 to the ground state, energy E_0. The number of downward transitions due to spontaneous emission, per second, is given by:

$$A_{10}N_1$$

In similar fashion, two other rates can be defined for the cases of stimulated emission and for absorption. These can be expressed in terms of two rate constants (or Einstein coefficients), one for stimulated emission and one for absorption. The rates are proportional to the numbers of atoms in the relevant state and the number of photons present. Thus, the rate at which atoms in the ground state energy E_0 are excited to the upper state energy E_1 is given by:

$$-\mathrm{d}N_0/\mathrm{d}t = B_{01}\rho(\nu_{01})N_0$$

where N_0 is the number of atoms in state E_0, $\rho(\nu_{01})$ is the radiation density responsible for absorption, which is the number of quanta incident per second at the correct excitation frequency ν_{01}, and B_{01} is the *Einstein coefficient for absorption of radiation*. Similarly, the rate of depopulation of state E_1 by stimulated emission is given by:

$$-\mathrm{d}N_1/\mathrm{d}t = B_{10}\rho(\nu_{10})N_1$$

where N_1 is the number of atoms in state E_1, $\rho(\nu_{10})$ is the radiation density responsible for depopulation, which is the number of quanta incident per second at the correct frequency ν_{10}, and B_{10} is the *Einstein coefficient for stimulated emission of radiation*. The frequency for excitation will be the same as that for depopulation, so that $\nu_{10} = \nu_{01}$, which can be written ν, and the radiation density will be the same in each case, hence:

$$\rho(\nu_{10}) = \rho(\nu_{01}) = \rho(\nu)$$

The number of stimulated downward transitions per second will be given by:

$$N_1B_{10}\rho(\nu)$$

while the total number of upward transitions in the same time will be given by:

$$N_0B_{01}\rho(\nu)$$

At equilibrium, the total number of transitions in each direction must be equal, hence:

$$N_0B_{01}\rho(\nu) = N_1A_{10} + N_1B_{10}\rho(\nu)$$

so

$$\rho(\nu) = N_1A_{10}/(N_0B_{01} - N_1B_{10})$$

At equilibrium, the Boltzmann distribution applies, thus:

$$N_1/N_0 = \exp(-h\nu/k_B T)$$

and by making this substitution:

$$\rho(\nu) = A_{10}/[\exp(-h\nu/k_B T)B_{01} - B_{10}]$$

This expression represents the net radiation emitted or absorbed by the material. The radiation density should be equal to the radiation density in a black body, derived by Planck:

$$\rho(\nu) = 8\pi h\nu^3/c^3[\exp(-h\nu/k_B T) - 1]$$

leading to the conclusion that:

$$B_{01} = B_{10}$$

which will be replaced by the single symbol B, and

$$A_{10}/B = 8\pi h\nu^3/c^3$$

The *ratio* of the rate of spontaneous emission to stimulated emission under conditions of thermal equilibrium, given by:

$$R = A_{10}/\rho(\nu)B = \exp(h\nu/k_B T) - 1$$

At 300 K, at visible wavelengths, R is much greater than 1. This shows that for light, stimulated emission will be negligible compared to spontaneous emission. On the other hand, if the wavelength increases beyond the infrared into the microwave and radio wave regions of the electromagnetic spectrum, R becomes much less than 1 and all emission will be stimulated. Hence, radio waves and microwaves arise almost entirely from stimulated emission and are always coherent.

The solution to the difficulty of obtaining a population inversion that will apply to optical frequency emission lies in recognising that the transition from one energy level to another is associated with a transition probability. If atoms can be excited into energy levels from which the probability of a transition is small, the atoms will remain in this state for long enough to produce a population inversion and so are available for stimulated emission. The resulting population inversion is *not* an equilibrium situation. How this is achieved in practice is described for three solid state lasers, the ruby laser, the neodymium laser, and the semiconductor laser.

12.2.6.1 The Ruby Laser: Three-Level Lasers

The ruby laser, invented in 1960, was the first device to put the ideas just described into practice. Rubies are crystals of Al_2O_3 containing about 0.5% Cr^{3+}. These impurity ions occupy octahedral sites as substitutional defects in place of Al^{3+}. The laser action only involves the 3d orbitals on the Cr^{3+} ions.

The ground state electron configuration of a free Cr^{3+} ion is $3d^3$ with all electron spins parallel (↑ ↑ ↑). The multiplicity of this configuration is:

$$2S + 1 = 2\ (1/2 + 1/2 + 1/2) + 1 = 4$$

and the ground state term symbol is ^4F (Section 1.1). The Cr^{3+} ion in ruby is in an octahedral site and the crystal field (Section 10.2.2) splits this term into three energy levels, which are, in ascending order, labelled $^4A_{2g}$, $^4T_{2g}$, and $^4T_{1g}$. Electron transitions involved in colour (optical transitions) are only allowed between energy levels with the same multiplicity and are called *spin-allowed transitions*. In the case of ruby there are thus two important spin-allowed transitions,2

$$^4T_{2g} \leftarrow {}^4A_{2g}$$

(at 556 nm, absorbs yellow/green)

$$^4T_{1g} \leftarrow {}^4A_{2g}$$ (at 400 nm, absorbs violet)

In addition, there is a higher energy term due to an electron configuration in which two electrons are spin-paired (↑ ↓ ↑). The multiplicity of this level is:

$$2S + 1 = 2\ (\ ½ + ½ - ½) + 1 = 2$$

and the corresponding free ion term is ^2G. In the crystal field of ruby this gives rise to four energy levels, of which one, ^2E_g, lies between the ground state energy level $^4A_{2g}$ and the first spin-allowed energy level $^4T_{2g}$ (Figure 12.11a). The colour of ruby is due to the transitions from the ground state to the two spin-allowed higher energy states, giving rise to a double peaked absorption spectrum (Figure 12.11b). A transition to or from $^4A_{2g}$, $^4T_{2g}$, and $^4T_{1g}$ to ^2E_g level is forbidden under the electron spin rule. However, in ruby, excited Cr^{3+} ions in states $^4T_{2g}$ or $^4T_{1g}$ can lose energy to the crystal structure and drop down to level ^2E_g. This process operates under different selection rules to the optical transitions and is independent of spin. The energy loss is taken up as phonons in a *radiationless* or *phonon assisted transition* and the ruby crystal warms up.

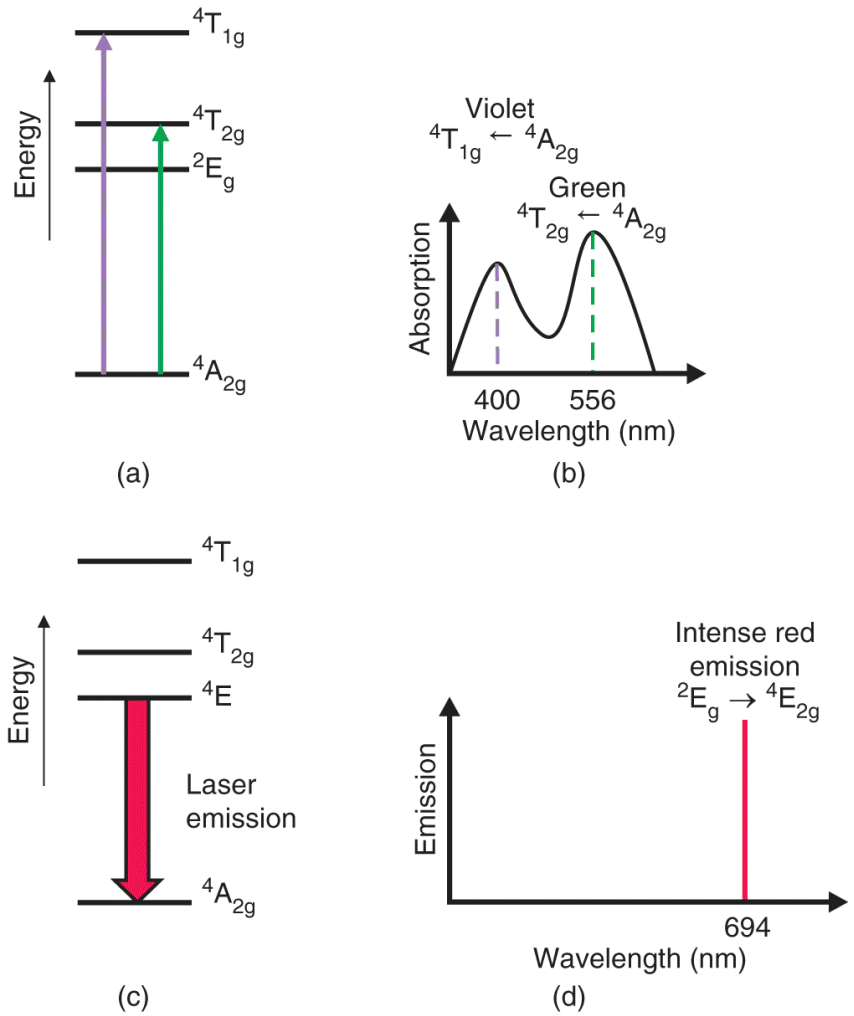


Figure 12.11 The energy levels involved in the ruby laser: (a) spin-allowed transitions; (b) the absorption spectrum of ruby; (c) the laser transition; (d) the emission spectrum of a ruby laser.

Typical rates of the transitions are:

$$^4T_{2g} \rightarrow {}^4A_{2g},\ 3 \times 10^5\ s^{-1}$$
$$^4T_{2g} \rightarrow {}^2E_g,\ 2 \times 10^7\ s^{-1}$$

The second of these two transitions is about 100 times faster than the first. The rates of the transitions from the $^4T_{1g}$ energy level to $^4A_{2g}$ and ^2E_g levels are of a similar magnitude. This means that on irradiating the ruby with white light, Cr^{3+} ions will be excited to energy levels $^4T_{2g}$ and $^4T_{1g}$, and then a significant number end up in the ^2E_g state rather than returning to the ground state. The transition from ^2E_g to the ground state is not allowed because of the spin rule and so atoms in the ^2E_g state have a long lifetime. (The spontaneous emission rate is $2 \times 10^2\ s^{-1}$.)

Laser operation takes place in the following way. An intense flash of white light is directed onto the crystal – a process called *optical pumping*. This excites the Cr^{3+} ions into the $^4T_{2g}$ and $^4T_{1g}$ states. These then lose energy by radiationless transitions and 'flow over' into state ^2E_g to produce a population inversion between ^2E_g and $^4A_{2g}$. About 0.5 ms after the start of the pumping flash, some

spontaneous emission will occur from ^2E_g. In order to prevent these first photons from escaping from the crystal without causing stimulated emission from the other excited ions one end is coated with a mirror and the other with a partly reflecting mirror. In this arrangement, the photons are reflected back and forth, causing stimulated emission from the other populated ^2E_g levels. Once started, the stimulated emission rapidly depopulates these levels in an avalanche. There will be a burst of red laser light of wavelength 694.3 nm, which emerges from the partly reflecting surface (Figure 12.11c,d).

Following light output, the upper levels will be empty and the process can be repeated. The ruby laser generally operates by emitting energy in short bursts, each of which lasts about 1 ms, a process referred to as *pulsed* operation. The ruby laser is called a *three-level laser*, because three energy levels are involved in the operation. These are the ground state ($^4A_{2g}$), an excited state reached by optical absorption or pumping ($^4T_{2g}$ or $^4T_{1g}$), and an intermediate state of long life-time (^2E_g) reached by radiationless transfer and from which stimulated emission (laser emission) occurs to the ground state.

It is energetically costly to obtain a population inversion in a three-level laser because one must pump more than half the population of the ground state to the middle level. Very little of the electrical energy supplied to the flash lamp ends up pumping photons, and carefully designed reflectors are essential. The energy lost in the transitions from $^4T_{2g}$ and $^4T_{1g}$ to ^2E_g ends up as lattice vibrations, which cause the crystal to heat up considerably. To make sure that the ruby does not overheat and shatter, it is necessary to cool the crystal and to space the pulses to allow the heat to dissipate. Although the ruby laser was the first laser made, the three-level mode of operation makes it inefficient.

12.2.6.2 The Neodymium (Nd^{3+}) Solid State Laser: Four-Level Lasers

Laser operation in *four-level lasers* takes place in the following sequence of steps and results in a more energy efficient device. Atoms in the ground energy state E_0 are excited to a rather high energy level, E_1 by optical pumping (Figure 12.12). Subsequently, the atoms in E_1 lose energy without radiating light, to an intermediate state I_1. Both steps should be fast and efficient. However, once in I_1, atoms should have a long lifetime and not lose energy quickly. When another intermediate state, I_0, is present and sufficiently high above the ground state to be effectively empty, a small population in I_1 gives a population inversion between I_1 and I_0. Ultimately a few photons will be released as some atoms drop from state I_1 to I_0. These can promote stimulated emission between I_1 and I_0, allowing laser action to take place. Atoms return from I_0 to E_0 by a step that needs to be rapid. If the energy corresponding to the transitions from E_1 to I_1 and I_0 to E_0 can be easily dissipated, the laser matrix does not become too hot and continuous operation rather than pulsed operation is possible.

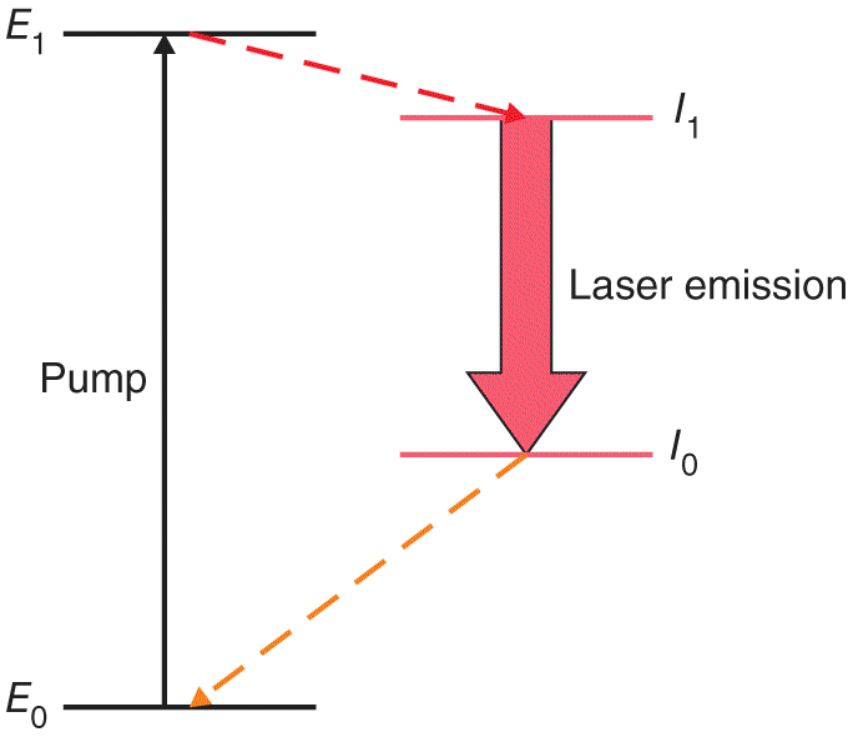


Figure 12.12 The main transitions in four-level laser operation.

The most important four-level solid-state laser uses neodymium (Nd^{3+} ions) as the active centres and the important transitions are between the f-electron energy levels. These are rather sharp and can be approximated to free ion energy levels because the f orbitals are shielded from the effects of the surrounding crystal lattice by 5d and 6s orbitals. Above the f-electron energy levels lie energy bands of considerable width derived from the interaction of these same 5d and 6s orbitals (Figure 12.13). Optical pumping excites the ions from the ground state to these wide bands. This process is very efficient because broad energy bands allow a wide range of wavelengths to pump the laser and because the transitions are allowed. In addition, loss of energy from the excited state down to the f-electron energy levels is fast. The energy loss halts at the ^4F pair of levels. The principal laser transition is from these to the $^4I_{11/2}$ level. The emission is at approximately 1060 nm in the infrared. This is a useful wavelength as it coincides with a reasonably low-loss region of silica based optical fibres (Section 12.8).

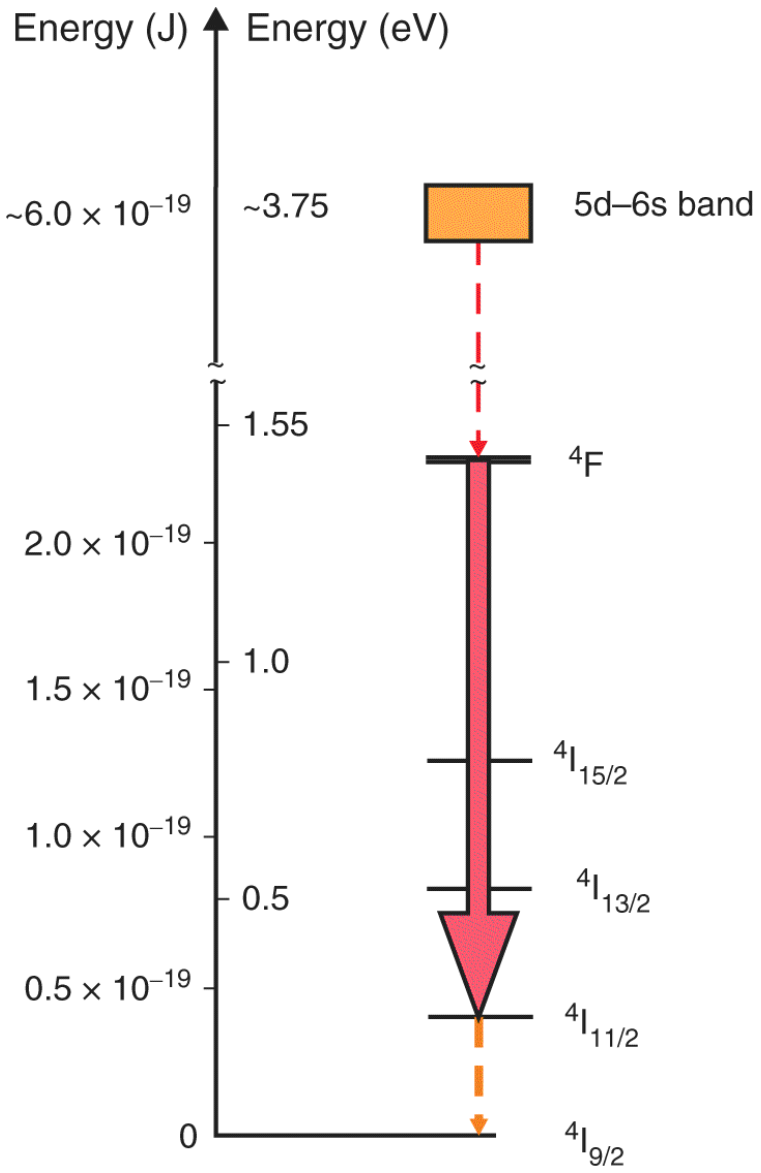


Figure 12.13 The energy levels of most importance in the neodymium laser.

Practical lasers contain about 1% Nd^{3+} and have high power outputs. The most common host materials are glass, yttrium aluminium garnet (YAG), and calcium tungstate, $CaWO_4$. They can be operated continuously or pulsed. At higher Nd^{3+} concentrations the lifetime of the ^4F levels drops from about 200 μs in a typically 1% doped material to about 5 μs at higher dopant concentrations, due to Nd–Nd interactions and associated changes in lattice vibration characteristics. Under these conditions, laser operation is no longer possible.

12.2.6.3 Semiconductor Lasers

Semiconductor lasers are, in essence, identical to LEDs, although the physical structure of lasers tends to be more complex. In principle, a semiconductor laser consists of a p–n junction, also called the *active layer*, in which one component has been heavily doped. The device is placed under forward bias and a high current is passed across the active layer. Initially the unit functions as an LED, and electrons recombine at random with holes to give out light by way of spontaneous emission. If the current is high enough, at some point in the junction region the number of electrons in the conduction band (from the n-type region) exceeds the number in the valence band (from the p-type region). When this population

inversion is achieved, stimulated emission can occur and light leaving the active layer is coherent laser light (Figure 12.14). To increase the chance of stimulated emission occurring and to make the beam directional, two ends of the device are polished, and the whole is constructed of several carefully engineered layers with varying electronic characteristics.

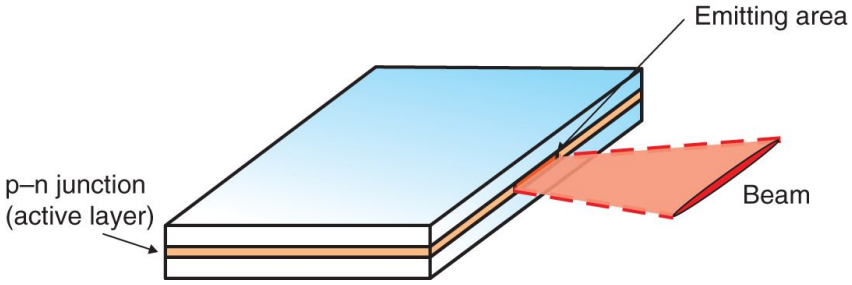


Figure 12.14 Semiconductor laser, schematic.

An advantage of semiconductor lasers is that they are very efficient. Moreover, because the emission comes from the host material itself (not a small quantity of dopant as in ruby or neodymium lasers), these lasers are very powerful for their size.

12.3 Refraction

12.3.1 The Refractive Index

Refraction is the apparent bending of a ray of light when it enters a transparent material such as water or glass (Figure 12.15). The magnitude of the deviation is given by the *index of refraction* or *refractive index*, n, where:

$$n = \sin\theta_1 / \sin\theta_2$$

θ_1 being called the *angle of incidence* and θ_2 the *angle of refraction* (Table 12.3). This equation is known as Snell's law (even though the originator was named Snel). This equation is a special case of the more general relation that applies to light passing from a medium of refractive index n_1 to one of refractive index n_2:

$$\sin\theta_1 / \sin\theta_2 = n_2/n_1$$

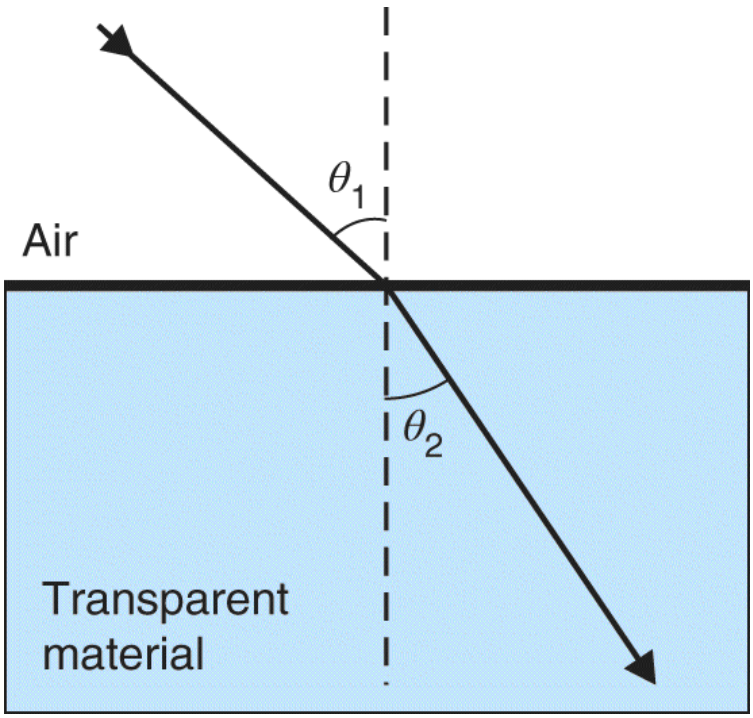


Figure 12.15 Refraction of a light beam on entering a transparent solid.

Table 12.3 Refractive indices.

Substance	Refractive index[a]	Substance	Refractive index[a]
(Vacuum)	1.0 (definition)	Dry air, 1 atm, 15 °C	1.00027
Water	1.3324	Na_3AlF_6 (cryolite)	1.338[b]
MgF_2	1.382[b]	Fused silica (SiO_2)	1.4601
KCl (sylvite)	1.490	Crown glass	1.522
Extra light flint glass[c]	1.543	NaCl (halite)	1.544
Flint glass[c]	1.607	MgO (periclase)	1.735
Dense flint glass[c]	1.746	Al_2O_3 (corundum)	1.765[b]
ZrO_2 (zirconia)	2.160[b]	C (diamond)	2.418
$CaTiO_3$ (perovskite)	2.740	TiO_2 (rutile)	2.755[b]

[a] A value appropriate to the yellow light emitted by sodium atoms, the sodium D-lines, with an average wavelength 589.3 nm, is given.

[b] The refractive index varies with direction; the average value is given.

[c] The flint glasses contain significant amounts of lead oxide, PbO, as follows: extra light flint, 24 mass% PbO; flint, 44 mass% PbO; dense flint, 62 mass% PbO.

In amorphous materials such as glass, or in crystals with a cubic symmetry, the index of refraction is the same in all directions. These are called *optically isotropic* solids. In the majority of crystals, the index of refraction varies with direction. These are called *optically anisotropic* materials.

In effect, the refractive index is a manifestation of the fact that the light is *slowed* on entering a transparent material. This is due to the interaction of the light with the electrons around the atoms that make up the solid. It is found that the refractive index of a transparent substance is given by:

$$n = c/v$$

where c is velocity of light in a vacuum and v the velocity of light in the medium. The frequency of the light does not alter when it enters a transparent medium and because of the relationship between the velocity of a light wave and its frequency:

$$v\lambda = v$$

$$n = c/v = \lambda_v/\lambda_s$$

where λ_v is the wavelength of the light wave in a vacuum and λ_s is the wavelength in the transparent substance. Thus, light has a smaller wavelength in a transparent material than in vacuum (Figure 12.16).

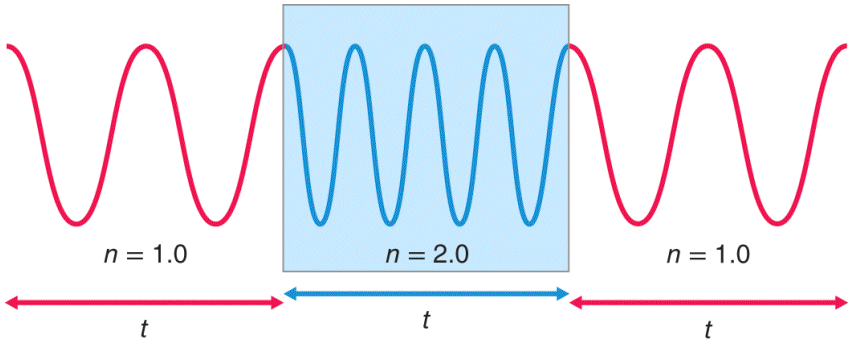


Figure 12.16 The effect of refractive index on the wavelength of light.

This can introduce confusion when a light ray traverses several different materials. To overcome this, it is useful to define the *optical path* or *optical thickness* [d], and distinguish it from the *real* or *physical thickness* of a material, t. The relationship is given by:

$$[d] = nt$$

where n is the refractive index. The optical thickness of a material is frequently quoted as a number of wavelengths. Thus, a thin film with an optical thickness of λ has a real thickness given by the wavelength of the light involved divided by the refractive index. For several transparent materials traversed in sequence

$$[d] = n_1 t_1 + n_2 t_2 + n_3 t_3 + \cdots$$

When light passes from a higher refractive index material such as glass to one of lower refractive index such as air, the refraction causes the emerging ray to bend towards the interface. As the angle, θ, at which the ray approaches the surface increases, the angle of the emerging ray becomes closer to the surface, until, at the *critical angle*, θ_c the emerging ray actually travels exactly along the surface (Figure 12.17). If θ_c is exceeded, then *no light escapes* and all behaves as if it were reflected from the under surface. This effect is called *total internal reflection*. The critical angle follows from the general relation given above when θ_i is equal to 90°:

$$\sin \theta_c = n_{\text{low}}/n_{\text{high}}$$

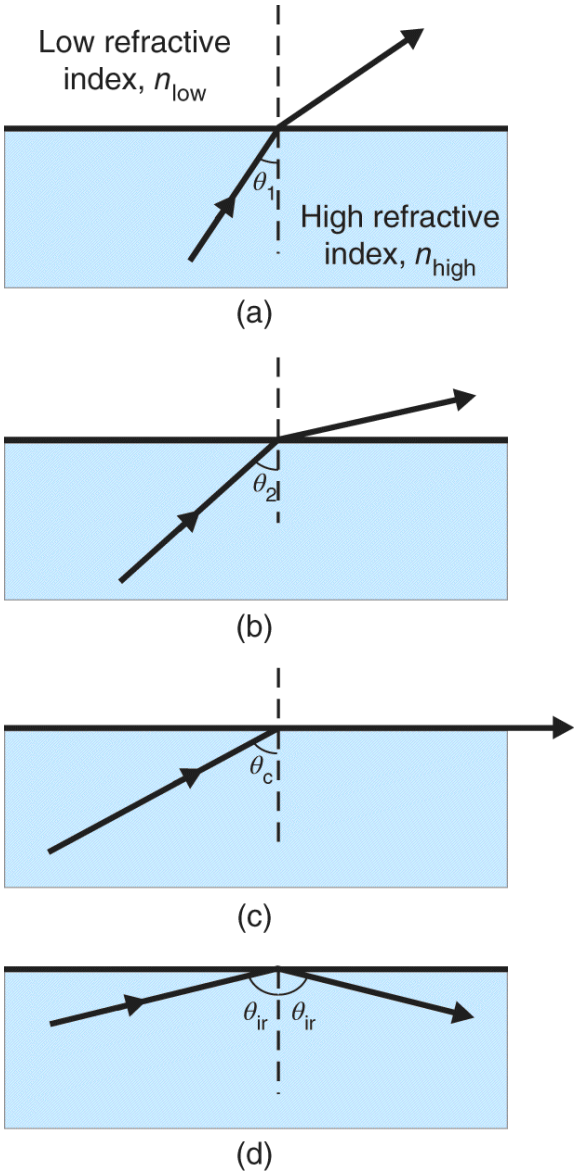


Figure 12.17 Total internal reflection of a beam of light: (a), (b) θ less than *critical angle*, θ_c; (c) θ equals θ_c; (d) θ greater than θ_c.

Note that this is not an all or nothing phenomenon. At angle θ_1 a small amount of the light is reflected into the solid, but most escapes. At angle θ_2 more is reflected back into the solid, and ultimately at θ_c all the light is reflected back into the solid and none escapes. At greater angles, θ_{ir}, the surface acts as a mirror.

The refractive index of a transparent solid varies with wavelength. This is called (normal) *dispersion*. In general, the index of refraction of transparent materials increases as the wavelength decreases so that the refractive index of red light in a material is less than that of violet light. The *dispersion* can be formally defined as $dn/d\lambda$, which is the slope of the refractive index vs. wavelength curve. Although the dispersion of many materials is rather small, it is important to include it when calculating the optical properties of optical components. Note that when the energy of the light is sufficient to promote an electron from one energy level to another, the smooth trend in dispersion described earlier no longer applies and the solid is said to exhibit *anomalous dispersion*.

The refractive index of metals and many semiconductors is influenced by the electrons that occupy energy bands rather than discreet orbitals. The refractive index in these cases is expressed as a complex number:

$$N = n + ik$$

where N is the refractive index of the solid, n is the real part of the refractive index, and k is called the *absorption index*, *absorption coefficient*, *attenuation coefficient*, or *extinction coefficient*. The pair of terms n and k are called the *optical constants* of the material, although they vary considerably across the electromagnetic spectrum and are by no means constant.

12.3.2 Refractive Index and Structure

The refractive index of a transparent material reflects the interaction of the electric field of a light wave with the electrons present – the electronic polarisability of the medium (Section 9.1.2). In general, strongly bound electrons have a low polarisability, and this leads to a low refractive index. Loosely bound electrons, outer electrons on large atoms, or lone pair electrons are highly polarisable and so will yield materials with a larger refractive index.

This qualitative link can be quantified via the *Gladstone–Dale formula*, which allows refractive indices of materials to be estimated. It is especially useful for complex oxides, for which the Gladstone–Dale formula can be written:

$$n = 1 + \rho(p_1 k_{r1} + p_2 k_{r2} + p_3 k_{r3} + \cdots)$$

or

$$n = 1 + \rho \sum p_i k_{ri}$$

where ρ is the density of the oxide. The factor k_r, called the *refractive coefficient*, is an empirically determined constant (Table 12.4). The amount of each oxide is taken into account by multiplying the refractive coefficient by its *weight fraction* in the compound, p.

Table 12.4 Refractive coefficients for some oxides.[a]

Oxide	k_r	Oxide	k_r	Oxide	k_r	Oxide	k_r	Oxide	k_r	Oxide	k_r	Oxide	k_r	Oxide	k_r
H_2O	0.34														
Li_2O	0.31	BeO	0.24							B_2O_3	0.22	CO_2	0.22	N_2O_5	0.24
Na_2O	0.19	MgO	0.20							Al_2O_3	0.21	SiO_2	0.21	P_2O_5	0.18
K_2O	0.20	CaO	0.21			TiO_2	0.40			Ga_2O_3	0.17	GeO_2	0.17	As_2O_3	0.16
Rb_2O	0.13	SrO	0.15	Y_2O_3	0.17	ZrO_2	0.21	Nb_2O_5	0.27	In_2O_3	0.13	SnO_2	0.14	Sb_2O_3	0.15
Cs_2O	0.12	BaO	0.13	La_2O_3	0.15			Ta_2O_5	0.15			PbO	0.13	Bi_2O_3	0.14

[a] These refractive coefficients apply when the density of the material is measured in g cm^{-3}. The use of density in kg m^{-3} requires that the coefficients above are each multiplied by 1×10^{-3}.

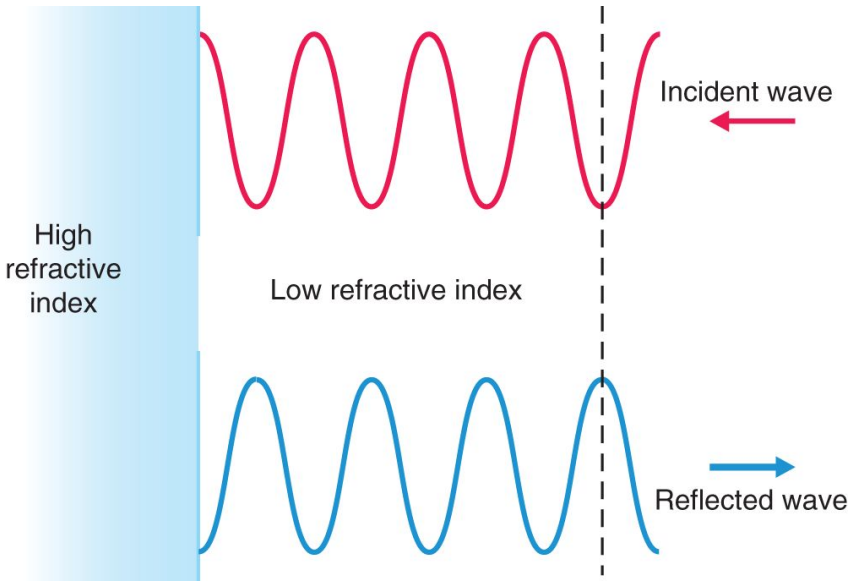


Figure 12.19 Phase change on reflection at a surface of higher refractive index.

The assumption underlying the formula is that the refractive index of a complex oxide is made up by adding together the contributions from a collection of simple oxides for which optical data are known. The rule works well and usually gives answers within about 5%. Note, however, that the value obtained is an average refractive index. The Gladstone–Dale relationship ignores the fact that many oxides have refractive indices that vary according to crystallographic direction.

12.4 Reflection

12.4.1 Reflection from a Surface

Light is reflected from smooth surfaces following the *law of reflection*: the angle of incidence, θ_i, is equal to the angle of reflection, θ_r. The *amount* of light reflected from a single surface at *normal incidence* (i.e. perpendicular to the surface) is given by the *coefficient of reflection*, r:

$$r = (n_0 - n_1)/(n_0 + n_1)$$

where n_0 is the refractive index of the entrance medium and n_1 is the refractive index of the material making up the reflecting surface (Figure 12.18). The coefficient of reflection for an insulating solid such as glass is defined such that if a wave of *amplitude* a_0 falls upon the surface, then the *amplitude* of the reflected wave is ra_0. For reflection at a surface between a substance of low refractive index and a substance of high refractive index, r is *negative*. This signifies a phase change of π radians on reflection, which means, in terms of a light wave, that a peak turns into a trough on reflection (Figure 12.19).

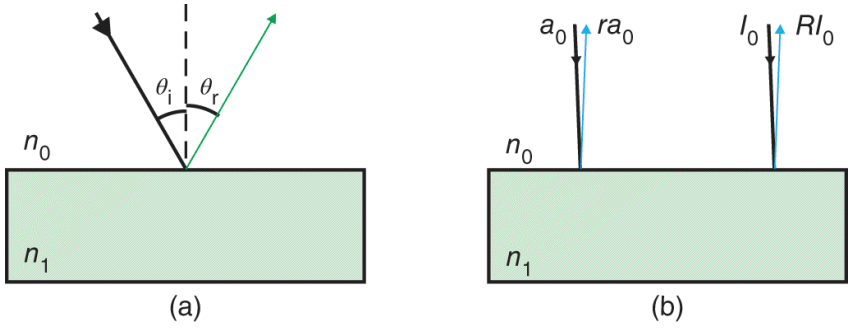


Figure 12.18 The reflection of light from a surface: (a) angles of incidence and reflection; (b) coefficient of reflection.

The eye detects *irradiance* changes rather than amplitude changes, and so it is more convenient to work with the *reflectivity* or *reflectance*, R:

$$R = r^2 = [(n_0 - n_1)/(n_0 + n_1)]^2$$

This is because the irradiance, I_0, is proportional to the square of the amplitude $(a_0)^2$. The reflected irradiance is proportional to $r^2(a_0)^2$. The reflectivity, R, for a plate of a transparent material of refractive index n in air is:

$$R = [(1 - n)/(1 + n)]^2 = [(n - 1)/(n + 1)]^2$$

As n depends upon wavelength, the reflectivity will vary across the spectrum.

In general, and especially when the reflecting surface is a metal, it is necessary to use the complex refractive index, $N = n_1 + ik$. In this case, the reflectivity of a surface at *normal incidence* is

$$R = [(n_1 - n_0)^2 + k^2]/[(n_1 + n_0)^2 + k^2]$$

For the case of a surface in air this becomes:

$$R = [(n_1 - 1)^2 + k^2]/[(n_1 + 1)^2 + k^2]$$

12.4.2 Reflection from a Transparent Thin Film

Monochromatic light travelling through air falling upon a homogeneous transparent thin film will be reflected from the top surface to give a reflected ray. The light transmitted into the film will be repeatedly reflected from the bottom surface and the underside of the top surface. At each reflection, some of the light will escape to produce additional reflected and transmitted rays (Figure 12.20). Because of the difference in the paths taken by the repeatedly reflected rays, the waves that leave the film will interfere with each other and so contribute to the reflectivity. The reflectivity of a thin film for light at normal incidence is given by:

$$R = [r_1^2 + 2r_1r_2\cos 2\delta + r_2^2]/$$
$$[1 + 2r_1r_2\cos 2\delta + r_1^2r_2^2]$$

where

$$r_1 = (n_0 - n_1)/(n_0 + n_1)$$

$$r_2 = (n_1 - n_0)/(n_1 + n_0)$$

$$\delta = 2\pi[d]/\lambda$$

and n_0 is the refractive index of the surrounding medium, n_1 the refractive index of the film, and $[d]$ is the optical thickness of the film. For a transparent film of refractive index n in air:

$$r_1 = (1 - n)/(1 + n)$$

$$r_2 = (n - 1)/(n + 1)$$

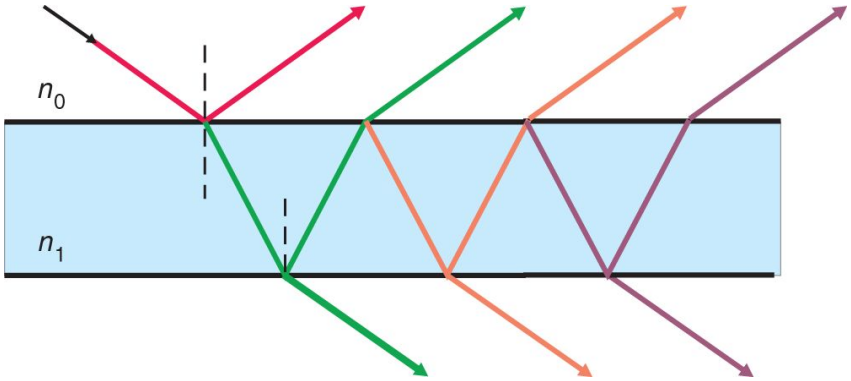


Figure 12.20 Multiple reflections at the upper and lower surfaces of a thin film in air.

The reflectivity varies in a cyclic fashion, with a minimum for values of $[d]$ equal to 0, $\lambda/2$, λ, etc. and a maximum for values of $[d]$ given by $\lambda/4$, $3\lambda/4$ (Figure 12.21).

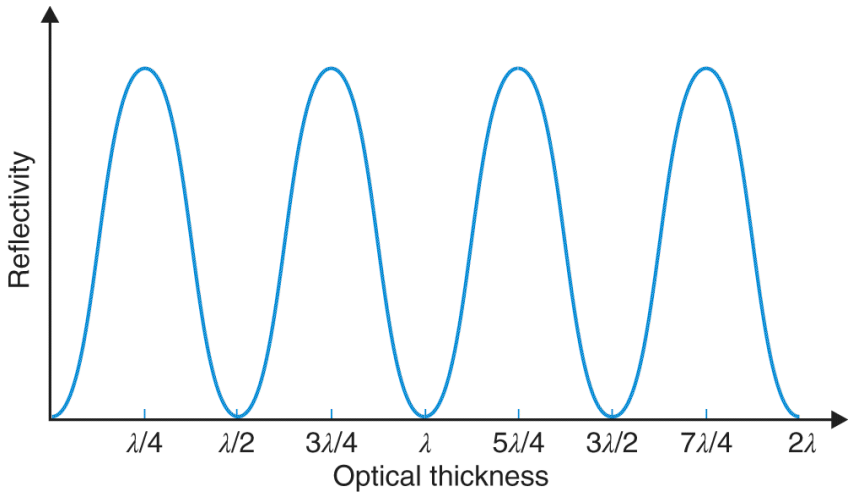


Figure 12.21 The sinusoidal variation of the reflectivity of a thin film.

The interference contribution may be approximated by considering interference between two waves, wave 1, reflected initially from the upper surface and wave 2 reflected the lower surface of the film of refractive index n in air. Recall that because wave 1 is reflected at a surface of higher refractive index a wave peak will turn into a trough. For an observer looking down onto the film in a perpendicular direction (Figure 12.22a), the path difference between waves 1 and 2, often called the *retardation* will be p:

$$p = 2[d] = 2nt$$

where t is the physical thickness, $[d]$ the optical thickness, and n is the refractive index of the film. If the path difference is equal to an integral number of wavelengths, when the phase change of wave 1 on reflection is added in, the two waves will be a half wavelength out of step and the film will appear dark due to destructive interference.

$$p = 2nt = m\lambda \, (m = 1, 2, 3, \ldots)$$

minimum (dark)

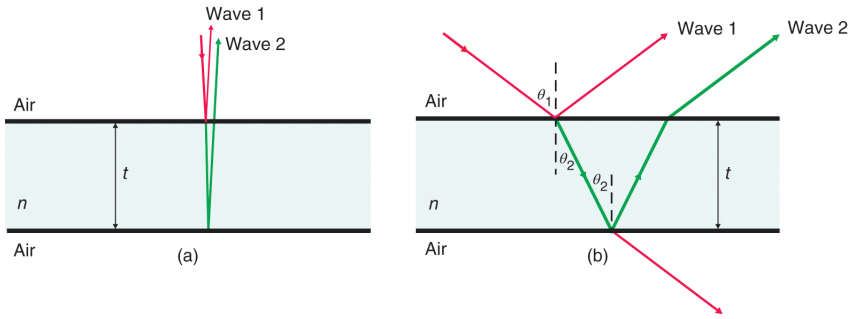


Figure 12.22 Reflection at a thin film in air: (a) perpendicular incidence (angle displayed for clarity); (b) incidence at an angle.

If the path difference, p, is equal to a half-integral number of wavelengths, when the phase change on reflection of wave 1 is added in, the two waves will be exactly in step and the film will then appear bright, because constructive interference will occur:

$$p = 2nt = (m + \tfrac{1}{2})\lambda \quad (m = 1, 2, 3, \ldots)$$

maximum (bright)

At other path differences, the film will appear to have an intermediate tone, depending upon the exact phase difference between the rays.

Should the light beam fall on the surface at an angle of incidence θ_1, producing and angle of refraction θ_2 (Figure 12.22b), the path difference, [p], between waves 1 and 2 is:

$$p = 2nt \cos\theta_2$$

provided that the angle of incidence is not too large (in which case polarisation effects dominate). In this case the dark and bright bands are given by

$$p = 2nt \cos\theta_2 = m\lambda$$

minimum (dark)

$$p = 2nt \cos\theta_2 = (m + {}^1/_2)\lambda$$

maximum (bright)

When a thin transparent film is viewed in white light, the same reflection and interference discussed earlier will occur, except that we have to add the contributions of all of the different wavelengths present. For any particular film thickness some of the colours will be reinforced by constructive interference, while others will be diminished by destructive interference. The result of this is that the film will appear coloured. As the film thickness increases or decreases, the sequence of colours seen varies in a cyclical fashion as certain colours are either reinforced or cancelled. Each sequence of spectral colours is called an *order*, which starts with the *first order* for the thinnest of films. A new order begins every 550 nm of retardation (Table 12.5, Figure 12.23).

Table 12.5 The colour of a thin film in white light.

Path difference[a] (nm)	Colour reflected[b]	Colour transmitted[c]
	Start of first order	
0	Black	Bright white
40	Iron grey	White
97	Lavender grey	Yellowish white
158	Grey-blue	Brownish white
218	Grey	Brownish yellow
234	Green-white	Brown
259	White	Bright red
267	Yellow-white	Carmine red
281	Straw yellow	Deep violet
306	Bright yellow	Indigo
332	Yellow	Blue
430	Yellow-brown	Grey-blue
505	Orange-red	Blue-green
536	Red	Green
551	Deep red	Yellow-green
555	End of first order; start of second order	
565	Magenta-purple	Bright green
575	Violet	Green-yellow
589	Indigo	Gold
609	Dark blue	Yellow
664	Sky blue	Orange
680	Blue	Orange-brown
728	Blue-green	Brown-orange
747	Green	Carmine red
826	Bright green	Purple-red
843	Yellow-green	Violet-purple
866	Green-yellow	Violet
910	Yellow	Indigo
948	Orange	Dark blue
998	Orange-red	Green-blue
1050	Crimson-violet	Yellow-green
1100	Dark violet-red	Green
1110	End of second order; start of third order	
Beyond this point the film colour is less intense and ultimately becomes pale pink or pale green in reflection.		

[a] The path difference p is often called the *retardation*.

[b] This colour is seen in *reflection* from a thin film in air when illuminated by white light at normal incidence.

[c] This colour is the complementary colour to that reflected and is the same as that shown in transmission by a thin film in air when illuminated by white light at normal incidence. In addition, these colours are seen in *reflection* when a thin transparent film on a substrate with a greater refractive index is viewed at normal incidence in white light.

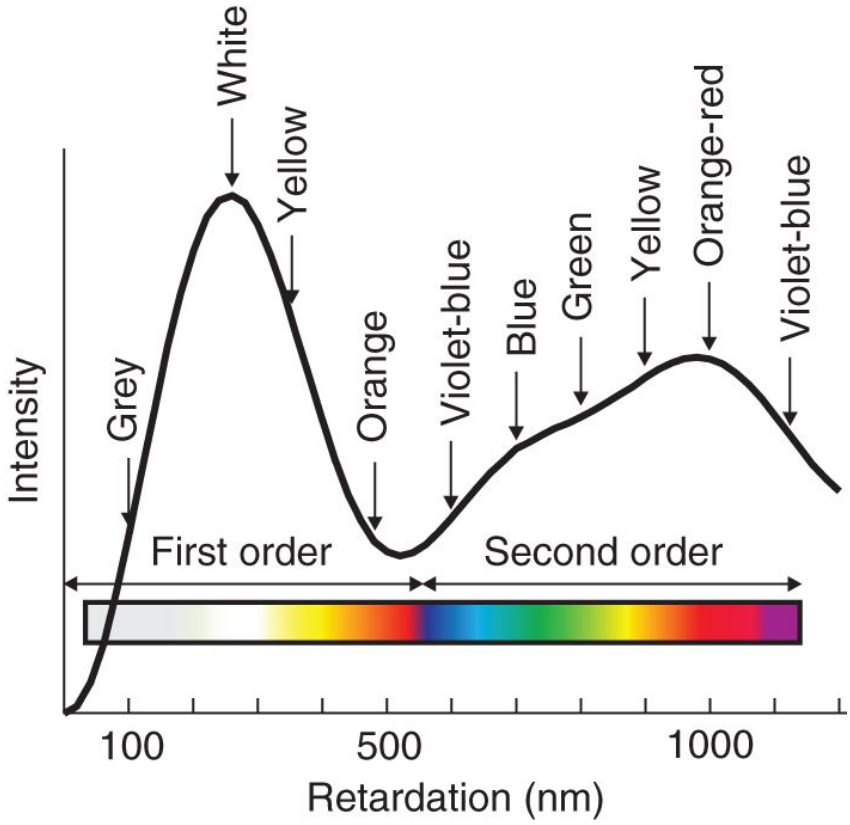


Figure 12.23 The approximate colour of a thin film when viewed from directly above as the film thickness varies.

Since the fraction of incident white light that is reflected is coloured, it follows that the transmitted light will be depleted in this colour. The transmitted colour seen will therefore be the *complementary colour* to that strongly reflected (Table 12.5).

If the angle of viewing is not perpendicular to the film, the optical path difference is given by:

$$[p] = 2[d]\ \cos\theta_2 = 2n\,t\ \cos\theta_2$$

where θ_2 is the angle of *refraction*. This formula indicates that as the viewing angle moves away from perpendicular the film the colour observed will move towards that appropriate to a lower optical path length. For example, second order orange red will change towards green and blue. In addition, because the value of n depends on wavelength, the reflectivity will also vary across the spectrum.

The behaviour of a single thin film on a substrate is similar to that of a single thin film in air. Now, however, it is necessary to take into account any change of phase that might occur on reflection at the back surface of the film. If the substrate has a *lower refractive index* than the film on the surface, then the treatment will be identical to that for a thin film in air. If the refractive index of the substrate is *greater* than that of the film, then a phase change will be introduced at both the air–film interface and the film–substrate interface. In this case, the reflected colour seen at normal incidence when viewed in white light will be the *complementary* colour to that given, equivalent to the transmitted colour in Table 12.5.

12.4.3 Low-Reflectivity (Antireflection) and High-Reflectivity Coatings

The reflectivity, R, of a homogeneous non-absorbing thin film on a substrate depends upon the optical thickness $[d]$ of the film. For values of $[d]$ given by $\lambda/2$, λ, $3\lambda/2$, etc.

$$R = [(n_0 - n_s)/(n_0 + n_s)]^2$$

This equation is interesting, because if n_0 is set as 1.0, for air, and n_s is set as the refractive index of the material, n, it is *identical* to the equation for an uncoated surface. Thus, a layer of optical thickness $\lambda/2$ can be considered to be *optically absent*.

For values of [d] given by $\lambda/4$, $3\lambda/4$, etc.

$$R = [(n_f^2 - n_0 n_s)/(n_f^2 + n_0 n_s)]^2$$

The reflectance will be either a *maximum* or a *minimum*. When the film has a *higher* refractive index than the substrate, the reflectivity will be a *maximum*. When the film has a *lower* refractive index than the substrate, the reflectivity will be a *minimum*.

These equations indicate that to make a non-reflective coating (*antireflection* [*AR*] coating) on a glass surface in air the value of n_f must lie between that of air and the glass. The reflectivity will be a minimum for a $\lambda/4$ film. Putting R equal to zero yields a value of the refractive index of the film that will give no reflection at all:

$$n_f = \sqrt{(n_s)}$$

For glass, n_s, is about 1.5, so the film must have a refractive index of approximately 1.225. Very few solids have such a low index of refraction, and a compromise material often used is magnesium fluoride, for which n in the middle of the visible is 1.384. Because the refractive index of a solid varies with the wavelength of the incident light, any antireflecting film that is suitable for a particular wavelength, the *design wavelength*, will not be perfect across the whole of the spectrum.

A similar strategy can be used to optimise the reflectivity of a surface by coating it with a high reflection coating. A film of thickness $\lambda/4$ will increase the reflectivity in the case when the refractive index of the film is greater than the refractive index of the substrate. Two materials frequently used are silicon monoxide ($n \approx 2.0$) or titanium dioxide ($n \approx 2.90$). A titanium dioxide film of thickness $\lambda/4$ on glass will have a reflectivity of about 0.48 (48%). As R for a single glass surface in air is about 0.04 (4%), almost 48% represents a great improvement. The effect is used in costume jewellery. Rhinestones are made of glass coated with an approximately $\lambda/4$ thickness film of TiO_2.

12.4.4 Multiple Thin Films and Dielectric Mirrors

Multiple thin films of transparent materials can be laid down one on top of the other in such a way as to form perfect mirrors, called *dielectric mirrors*. The simplest formulae for the reflectance of such a mirror refer to the specific case in which all layers are $\lambda/4$ thick and of alternating high (H) and low (L) refractive indices, n_H and n_L, called a *quarter-wave stack* (Figure 12.24). For such a stack deposited on a substrate in the sequence: substrate; L, H; L, H; L, H; ...; L, H; air and illuminated by light falling *perpendicular* to the surface, *maximum* reflectance is given by:

$$R = [(n_s f - n_0)/(n_s f + n_0)]^2$$

where f is equal to $(n_H/n_L)^{2N}$, n_0 is the refractive index of the surrounding medium, usually air ($n_0 = 1.0$), n_s is the refractive index of the substrate, usually glass ($n_s \sim 1.5$), and N is the number of (LH) pairs of layers in the stack. For a stack in air the equation reduces to:

$$R = [(n_s - (n_L/n_H)^{2N})/(n_s + (n_L/n_H)^{2N})]^2$$

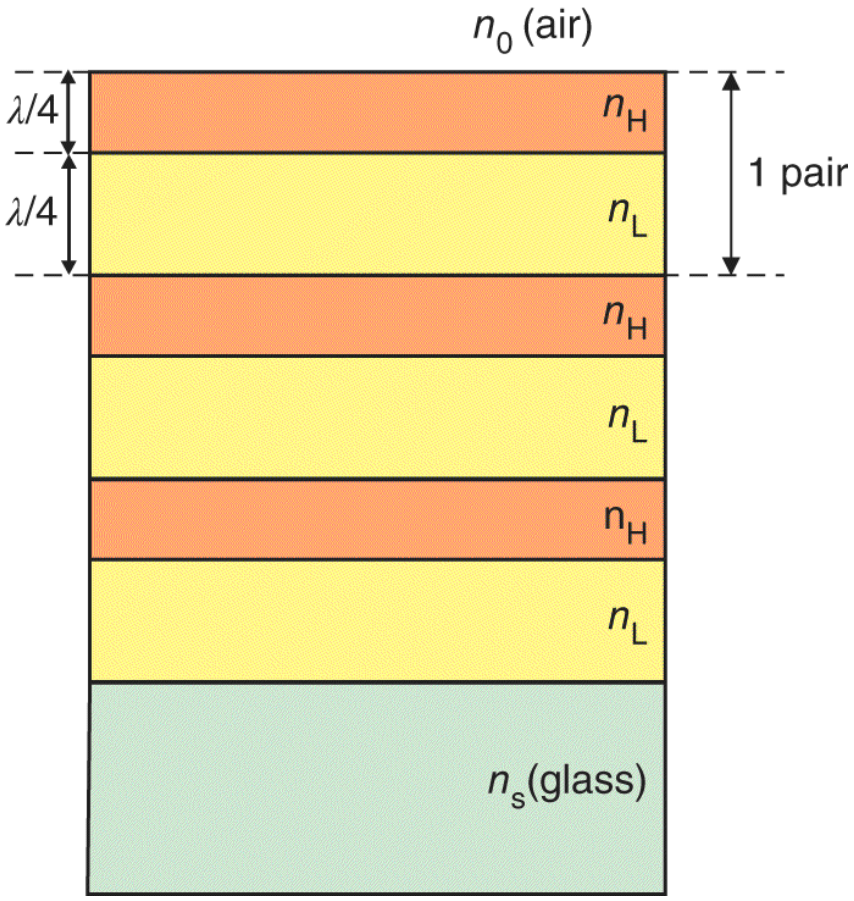


Figure 12.24 A quarter-wave stack of alternating layers of high and low refractive index solids.

The general approach used to make a dielectric mirror is to calculate the reflectivity of a stack of films using computer routines to determine the optimal thickness and refractive index of each layer so as to produce virtually perfect mirrors. The same method is employed to make antireflection coatings and a variety of *optical interference filters* that will allow certain wavelength bands to be transmitted while other bands are reflected. These fall into three categories. *Shortpass filters* transmit visible wavelengths and cut out infrared radiation. They are often used in surveillance cameras to eliminate heat radiation. *Longpass filters* block ultraviolet radiation and transmit the visible. *Bandpass filters* transmit a limited section of the electromagnetic spectrum (Figure 12.25).

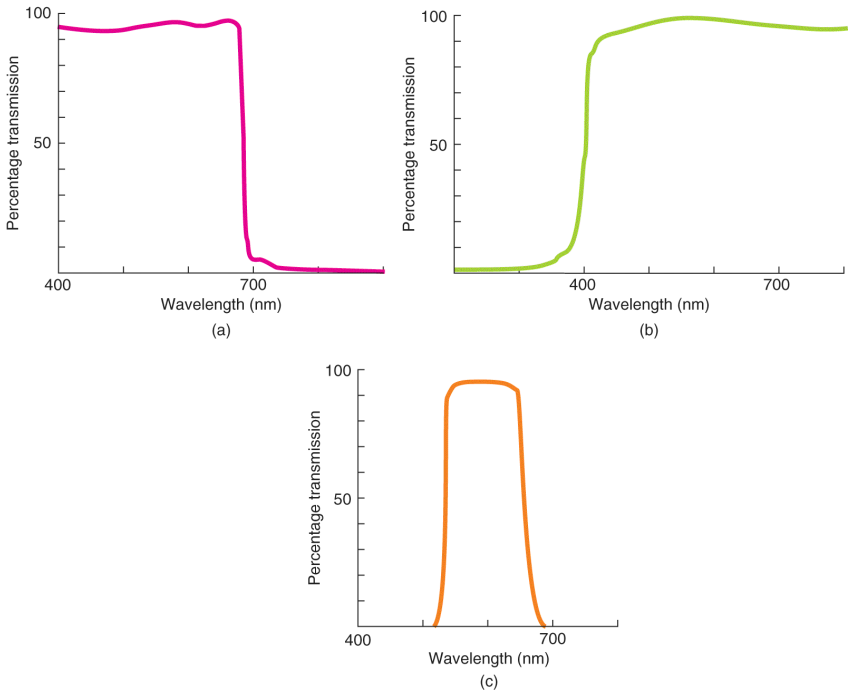


Figure 12.25 Interference filters: (a) shortpass; (b) longpass; (c) bandpass.

The reflectivity of a stack of transparent thin films that are of varying thickness or refractive index will give a reflection that appears to be metallic silver. This can be seen in a less than perfect fashion with a stack of microscope slides or a roll of thin transparent plastic film, such as ‘cling film’.

12.5 Scattering and Attenuation

12.5.1 Scattering

If a transparent solid contains *scattering centres*, the irradiance of light traversing the medium in the incident direction will gradually fall as the light is scattered into other directions. The distribution of scattered irradiation around a scattering centre is described by two models, *Rayleigh scattering* and *Mie scattering*. Rayleigh scattering applies to spherical insulating particles with a diameter less than about one tenth of the wavelength of the incident light and applies to scattering from small compositional inhomogeneities in the solid. The irradiance of the scattered light, I_s, at a distance d from the scattering centre is:

$$I_s = I_o(9\pi^2 V^2/2d^2\lambda^4)$$
$$[(m^2-1)/(m^2+2)]^2(1+\cos^2\theta)$$

where I_o is the incident irradiance, V is the volume of the scattering particle, λ is the wavelength of the light, θ is the angle between the incident beam and the direction of the scattered beam, and m is the relative refractive index of the particle:

$$m = n_p/n_m$$

where n_p is the refractive index of the particle and n_m the refractive index of the surrounding medium. As much light is scattered backwards as forwards and only half as much is scattered normal to the beam direction (Figure 12.26a). All wavelengths scatter in this pattern, but the shorter wavelengths are more strongly scattered than the longer wavelengths because the scattering is proportional to $1/\lambda^4$, so violet light is scattered far more than red light.

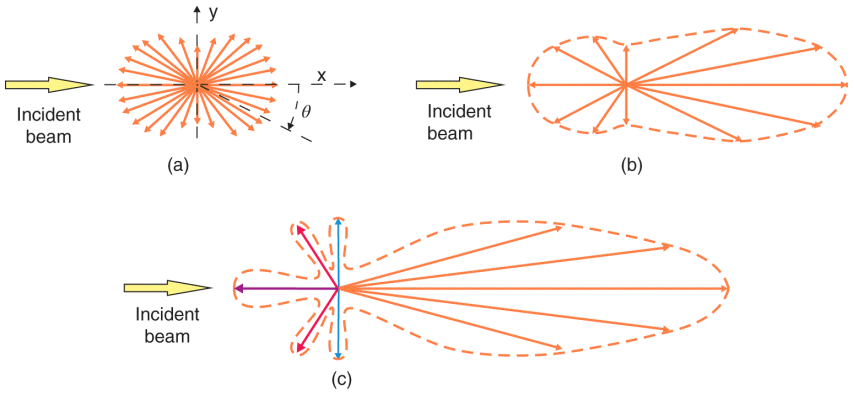


Figure 12.26 Scattering: (a) Rayleigh scattering; (b) Mie scattering from relatively small particles; (c) Mie scattering from larger particles.

The term Mie scattering is reserved for scattering by particles that are larger than those for which Rayleigh scattering is valid, about one third the wavelength of light or more. It is also applicable to scattering from both metallic and insulating particles. The colour of ruby glass is due Mie scattering from a dispersion of gold nanoparticles in the glass. As the size of the inhomogeneity increases from that appropriate to Rayleigh scattering, forward scattering begins to dominate over backward scattering (Figure 12.26b). As particle size passes the wavelength of light, the forward scattering lobes increase further, and side bands develop, due to maxima and minima of scattering at definite angles (Figure 12.26c). The position of these lobes depends upon the wavelength of the scattered light and so they are strongly coloured. These bands, referred to as *higher-order Tyndall spectra*, are dependent upon the particle size. With even larger particles, white light becomes reflected (rather than scattered in the sense used here). This situation holds in opaque glasses.

12.5.2 Attenuation

As a beam of light passes through a transparent material, it gradually loses intensity due to scattering or absorption, a process generally called *attenuation* (or sometimes *extinction*). When attenuation takes place in a homogeneous solid, the amount of light transmitted by a plate of thickness x is given by:

$$I_x = I_0 \exp(-\alpha_e x)$$

where I_x is the irradiance leaving the plate, I_0 is the incident irradiance, and α_e is the (*Napierian*) *linear attenuation coefficient* (*extinction coefficient*). This equation is known as *Lambert's law* or *Beer's law* (although first clearly set out by *Bouguer*). The *attenuation length* is defined as $1/\alpha_e$. The amount of light removed from the beam is thus:

$$\begin{aligned} I_r &= I_0 - I_x = I_0 - I_0 \exp(-\alpha_e x) \\ &= I_0[1 - \exp(-\alpha_e x)] \end{aligned}$$

If the attenuation of the beam is solely due to absorption, the attenuation coefficient is replaced by the (*Napierian*) *linear absorption coefficient*, α_a. Similarly, if the attenuation is solely due to scattering the attenuation coefficient is replaced by the (*Napierian*) *linear scattering coefficient*, α_s. For non-homogeneous solids these coefficients may vary with direction. Note that the degree of attenuation will vary significantly across the spectrum and the attenuation coefficient is not a constant.

Attenuation is often associated with the presence of chemical or physical centres, which may be atoms, molecules, or larger particles, distributed throughout the bulk of a material. In this case, the degree of attenuation is often a function of the concentration of the centres, expressed via the *Beer–Lambert* or *Beer–Lambert–Bouguer law*:

$$\log(I_x/I_0) = -\varepsilon c x$$

where I_x is the irradiance after passage through a length of sample x, I_0 is the incident irradiance, c is the *molar concentration* of the absorbing centres, and ε is the *molar (decadic) attenuation coefficient*. The dimensionless product ($\varepsilon\, c\, x$) is called the *absorbance* or *optical density* A and the ratio I_x/I_0 the *transmittance* or *transmissivity*, T. Thus:

$$\log T = -A$$

The Beer–Lambert Law finds use in the measurement of concentrations.

The incident irradiance I_0 can be equated to the amount of light reflected, I_r, scattered, I_s, absorbed, I_a, and transmitted I_t:

$$I_0 = I_r + I_s + I_a + I_t$$

or

$$1 = R + S + A + T$$

where R is the *fraction* of light reflected, I_r/I_0, S the *fraction* of light scattered, I_s/I_0, A the *fraction* of light absorbed, I_a/I_0, and T the *fraction* of light transmitted, I_t/I_0. In good quality optical materials, the amount of light scattered and absorbed is small and it is often adequate to write:

$$I_0 = I_r + I_t$$

$$1 = R + T$$

In a pure liquid the Beer–Lambert law is often written in the form:

$$\log(I_x/I_0) = -ax$$

where $a = \varepsilon c$ is the *molar (decadic) attenuation (or absorption) coefficient*.

12.6 Diffraction

Diffraction is a form of light scattering that occurs when waves interact with objects having a size similar to the wavelength of the radiation. Here the term object can be widely interpreted, from simple apertures in an opaque solid to, for example, refractive index changes in a transparent solid. In general, two regimes have been explored in most detail: diffraction quite close to the object which interacts with the light, called *Fresnel diffraction*, and diffraction far from the object which interacts with the light, called *Fraunhofer diffraction*. The result of diffraction is complex set of overlapping fringes, called a *diffraction pattern*. Here the discussion of diffraction of light is limited to Fraunhofer diffraction, as this is more often observed. However, the same diffraction fundamentals underlie structure solution via X-ray diffraction and electron microscopy, as well as more recent aspects, including holograms and diffraction from nanostructured surfaces.

12.6.1 Diffraction by an Aperture

If a long narrow slit is illuminated by monochromatic light, the irradiance pattern observed far from the slit (the *Fraunhofer diffraction pattern*) is given by the expression:

$$I_x = I_0\left[\frac{\sin x}{x}\right]^2$$

where

$$x = (\pi w \sin \theta)/\lambda$$

w is the width of the slit, θ is the angular deviation from the 'straight through' position, and λ is the wavelength of the light. This produces a set of bright and dark fringes with *minima* given by:

$$\sin \theta_{\text{min}} = m\lambda/w$$

where m takes values 1, 2, 3, etc. For θ_{min} to be appreciable, w must be close to λ and as the spacing between the minima will be proportional to the reciprocal of the slit width, narrower openings give wider fringe spacing (Figure 12.27). The positions of the maxima between these dark bands are not given by a simple formula but are approximately midway between the minima.

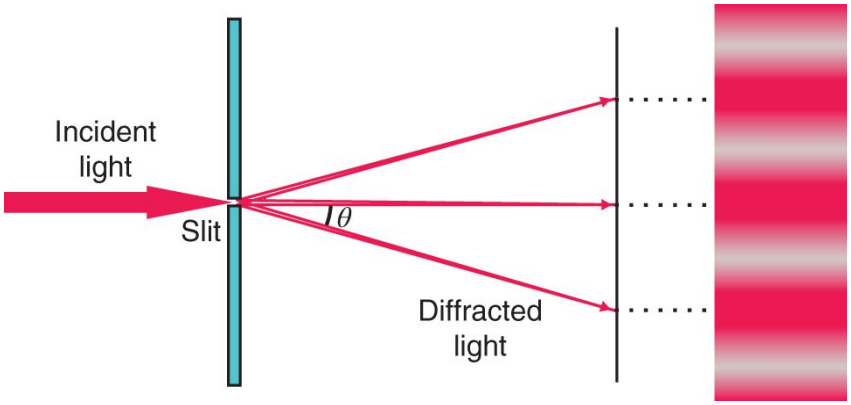


Figure 12.27 The diffraction pattern from a thin slit.

The sine of the angle through which a ray is diffracted is related to its wavelength. Thus, each wavelength in white light will be diffracted through a slightly different angle with red light deviated most. In this way, white light will produce a set of diffraction patterns, each belonging to a different wavelength. These patterns look like and are called *spectra*. They are referred to as *first order*, *second order*, and so on as they are recorded further and further from the undeviated beam.

When the slit is shortened to form a rectangular aperture, the diffraction maxima will take the form of small rectangular spots running in two perpendicular directions. White light will produce coloured

spots.

The form of the diffraction pattern produced by a circular aperture consists of a series of bright and dark circles concentric with the original aperture. The spacing of the maxima and minima is given by:

$$\sin\theta = m\lambda/d$$

where θ is the angle between the directly transmitted ray and the diffraction ring, λ is the wavelength of the light, and d the diameter of the aperture. The parameter m takes the values 0 (central bright spot), 1.220 (first dark ring), 1.635 (first bright ring), 2.333 (second dark ring), 2.679 (second bright ring), and 3.238 (third dark ring). The dependence of the diffraction angle upon wavelength means that a circular aperture illuminated with white light will produce a set of coloured rings, rather like miniature circular rainbows. Each ring will have a violet inner edge and a red outer edge.

Diffraction by a circular aperture limits the performance of optical instruments such as telescopes and microscopes as well as the eye. The resolution of such instruments, which is, roughly speaking, equivalent to the separation of two points that can just be distinguished as separate objects, is controlled by diffraction. It is of the order of the wavelength of the observing radiation. Because of this constraint optical microscopes are unable to image atoms. Electron microscopes, using radiation with a wavelength of the order of 0.002 nm, are able to do so.

12.6.2 Diffraction Gratings

Planar diffraction gratings consist of a set of parallel lines with spacing similar to that of the wavelength of light. A *transmission grating* has alternating clear and opaque lines, and diffraction effects are observed in light transmitted by the clear strips. A *reflection grating* consists of a set of grooves or blazes and diffraction effects are observed in the light reflected from the patterned surface (Figure 12.28). The effectiveness of a grating is the same whether light is transmitted through it or reflected from it.

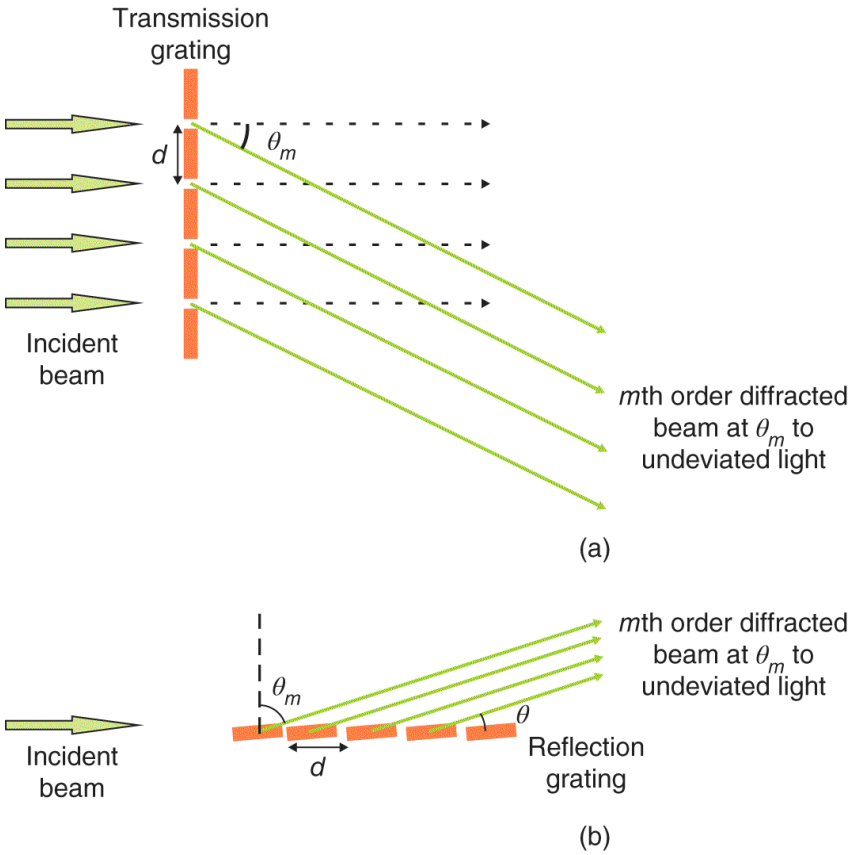


Figure 12.28 Diffraction gratings: (a) transmission; (b) reflection.

The position of the diffraction maxima from a transmission grating of repeat spacing d illuminated by monochromatic light normal to the surface is given by:

$$\sin\theta_m = m\lambda/d$$

where λ is the wavelength of the radiation and θ_m is the angle through which the beam has been diffracted. The positions of the maxima for light at grazing incidence to a reflection grating of repeat spacing d are given by:

$$(1 - \sin\theta_m) = m\lambda/d$$

$$(1 - \cos\theta) = m\lambda/d$$

where λ is the wavelength of the radiation, θ_m is the angle between the grating normal and the diffracted beam, and θ is $(90 - \theta_m)$.

The term m in both formulae can take integer values of 0, ±1, ±2, and so on. Each of these corresponds to a different diffraction maximum, called an *order*. When illuminated by white light, each wavelength will be diffracted through a slightly different angle so that each order will consist of a spectrum, similar to those produced by a long narrow slit, but because each line on the grating acts as a contributing slit, they are of much greater intensity. The colours seen reflected from CDs or DVDs are caused by the tracks on the disc surface, which act as curved reflection gratings.

12.6.3 Diffraction from Crystal-like Structures

The atomic structures of crystals were initially determined by recording the X-ray diffraction patterns obtained when crystals were irradiated by an X-ray beam. The positions of the beams of X-rays diffracted from the atoms in a crystal are given by *Bragg's Law:*

$$m\lambda = 2d \sin\theta_B$$

where d is the separation of the atom planes, λ is the wavelength of the radiation, and θ_B the angle between the beam and the scattering planes (Figure 12.29a). The theory holds for any three-dimensional array no matter the size of the 'atoms', provided that the spacing of the diffracting centres is similar to that of the wavelength of the incident radiation. Thus, an arrangement of particles or voids that are spaced by distances similar to the wavelength of light will diffract light according to Bragg's Law. When white light is used, each wavelength will diffract at a slightly different angle and colours will be produced.

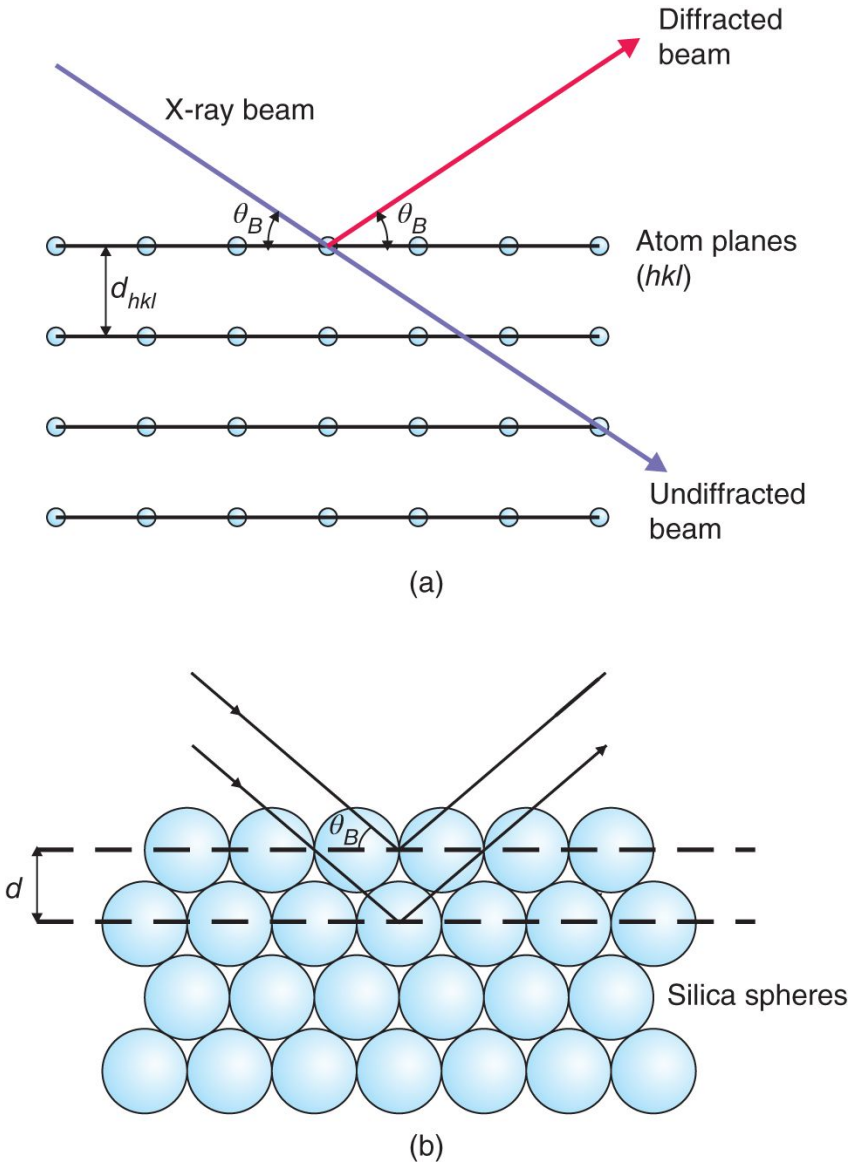


Figure 12.29 (a) Diffraction of X-rays by a crystal; (b) diffraction of light by an ordered array of silica spheres.

The colour of precious opal arises in this way. The regions producing the colours are made up of an ordered packing of silica spheres, which resemble crystallites, embedded in amorphous silica (Figure 12.29b). The spacing of the spheres is similar to that of the wavelength of light. However, because the diffraction takes place within a silica matrix, it is necessary to substitute the optical thickness $n_s d$ for the vacuum path d in Bragg's law. The refractive index of the silica in opal, n_s, is about 1.45 giving:

$$m\lambda = 2n_s d \sin\theta_B \sim 2.9d \sin\theta_B$$

The longest wavelength diffracted back to the viewer by the opal is when light falls perpendicular to the diffracting layers:

$$\lambda_{max} = 2n_s d \sim 2.9d$$

The relationship between the radius of the spheres, r, and the distance between the layers, d, will depend upon the exact geometry of the packing. If each layer of spheres is arranged in cubic or hexagonal closest packing the relationship between the sphere radius and the layer spacing is:

$$d = (2\sqrt{2}r)/\sqrt{3} = 1.633r$$

A useful general relationship is that the radius of the spheres is given, to a reasonable approximation, by one fifth of the wavelength of the colour observed at normal incidence.

Photonic crystals are artificial opal-like structures that diffract light in specified ways. The dimensions of the diffracting centres in the 'crystals' are approximately the same as the wavelength of light and the diffraction can generally be understood in terms of the Bragg equation. However, the terminology employed to describe diffraction in photonic crystals is that of semiconductor physics. The transition from a diffraction description to a physical description can be illustrated with respect to a one-dimensional photonic crystal.

One common form of one-dimensional photonic crystal is a stack of transparent volumes of alternating refractive indices similar to dielectric mirrors. They are called *Bragg stacks*, or, when built into an optical fibre, *fibre Bragg gratings* and they behave like selective mirrors that can pass or reflect specific wavelengths of the incident light. The simplest model is that of a transparent material containing regularly spaced air voids (Figure 12.30). When a beam of light is incident on such a structure, a single wavelength, λ, will be diffracted when Bragg's law is obeyed:

$$\lambda = 2\,nd\,\sin\,\theta_B$$

where n is the refractive index of the material and d the repeat spacing. For a beam normal to the voids, $\sin\theta = 1$ hence:

$$\lambda = 2\,nd$$

and the beam will be diffracted back on itself.

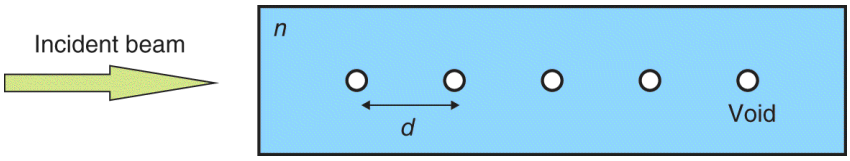


Figure 12.30 A one-dimensional photonic crystal.

The failure of the solid to allow light waves to propagate through the material is analogous to the fact that electron waves cannot propagate across an energy band gap. Because of this, the array of voids is said to have opened a *photonic band gap* (*PBG*), in the material that blocks transmission of the light wave with an appropriate wavelength. In real materials, the voids have thickness, and a small range of wavelengths is blocked, rather than just one. The result is similar to that described for multilayer interference filters, and the range of wavelengths blocked increases as the difference between the refractive indices of the voids and the surrounding medium increases.

Two-dimensional photonic band gap crystals can be constructed from a two-dimensional array of voids or particles in a transparent medium, and opal is an example of the three-dimensional photonic band gap crystal. Many animals use natural photonic crystal structures for the production of vivid colours including iridescence in butterfly wings, beetles, and feathers.

12.6.4 Holograms

12.6.4.1 Hologram Formation

A hologram is an interference pattern. Holograms were originally made using a monochromatic coherent light wave, generally from a laser source. The beam is divided into two. One part illuminates an object and this portion is modified by this interaction to create the *object wave* or *signal wave*. The other part of the beam traverses an identical distance but without having encountered the object, to

form the *reference wave* or *carrier wave*. The interference pattern that constitutes the hologram is created by subsequently overlapping the object wave with the unchanged reference wave.

The completed hologram is viewed by illuminating the hologram with a play-back light wave, the *reconstruction beam*. Play-back, in fact, gives rise to three output waves moving away from the hologram in different directions. The first can be thought of as a reconstruction of the reference wave and the second is a reconstruction of the object wave. The third wave is similar to the object wave, but with an inverted phase. The hologram (image) of the object is seen by viewing the object wave.

Rainbow holograms are the brightly coloured holograms seen on credit cards, security documents, and so on, and viewed in ordinary daylight not laser light (Figure 12.31a–c). The result of this is the formation of a multiplicity of reconstructed images, each arising from a separate wavelength. The image seen and the colours visible depend upon both the viewing angle and the viewing distance, so that a variety of colours and patterns can be picked out as the hologram is tilted and moved. The confusing images that this would ordinarily produce are constrained by arranging the image reconstruction as if it were viewed through a narrow slit. Thus, as the observer moves up and down, different slices of the composite images are seen. Each wavelength of light used in the reconstruction will create its own window, with red corresponding to most diffracted light and violet the least. Thus, an observer on a level with the hologram may see a green reconstruction, while if the observer moves up, the image colour will move through yellow and orange to red, or if the observer moves down to green, then blue, indigo, and violet (Figure 12.31d). There will still be considerable overlap of colours in parts of the image dependent upon the viewing distance and angle.

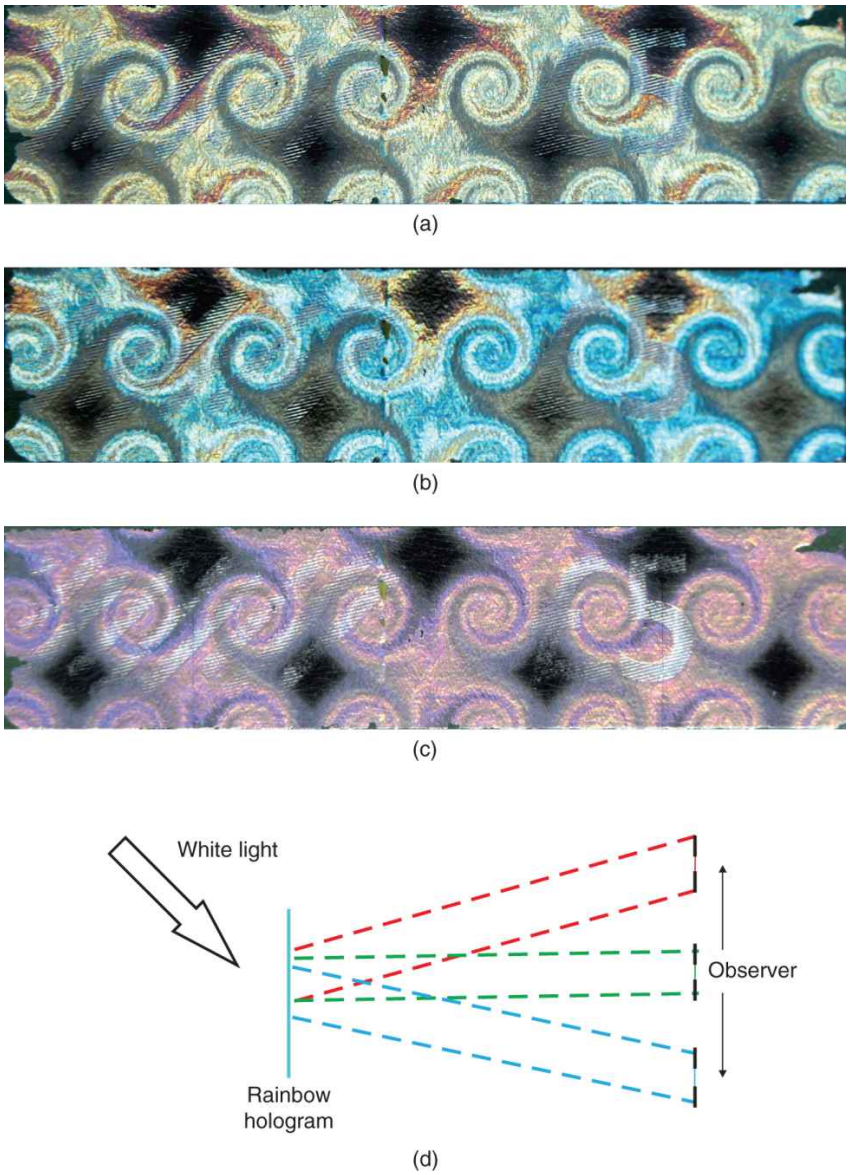


Figure 12.31 Rainbow hologram: (a–c) a security hologram viewed at slightly different angles; (d) hologram formation, schematic. Only three colours are shown for clarity.

12.6.4.2 Hologram Recording Media

The interference pattern that is the information content of a hologram is stored in a photosensitive material. The three principal classes of photosensitive materials that have been used for hologram storage employ photorefractive compounds (materials whose refractive index changes upon light irradiation), photopolymers (materials that polymerise under light irradiation), and photochromic materials (materials that change colour under light irradiation). To some extent, the type of photoresponsive medium used depends on the purpose of the hologram.

The first holographic recording medium used was fine grain photographic film deposited on a glass plate for maximum dimensional stability or else on a stable polymer base. Exposure to the object and reference beams causes silver crystallites to form in the emulsion, as in traditional photography. The emulsion, after development, contains blackened lines corresponding to the peaks in the interference pattern and transparent areas corresponding to the troughs. In this form the processed emulsion acts as an *amplitude* hologram, as it influences the amplitude of the reconstruction beam. These primary holograms can be 'bleached' to remove the silver grains. After this process the refractive index of the

emulsion varies in a mirror of the amplitude variation. In this form the emulsion acts as a *phase hologram* as the phase of the reconstruction beam is modified.

Photorefractive materials depend upon the hologram pattern to generate a refractive index modulation in the recording medium. A number of inorganic glasses, especially of the chalogenides (S, Se, and Te), are utilised to record phase holograms. In these solids, light of sufficiently high irradiance is used to break some of the bonds in the glass. This leads to refractive index changes and hence to a phase hologram.

Photopolymers contain a monomer that is able to polymerise under the influence of the holographic light pattern. In areas that show high intensity the monomer polymerises quickly. In areas where light intensity is low there is little polymerisation and the monomers that are not consumed diffuse to the strongly illuminated region, adding to this selective polymerisation. Eventually, the monomer is consumed, giving the highly polymerised regions a higher refractive index than the unexposed volumes, resulting in a refractive index distribution, which mirrors the holographic irradiance pattern.

Photochromic molecules show a reversible transformation between two forms, A and B, initiated by light. The absorption spectra of both molecular forms are different, and they can be converted from one form to the other by irradiation with the appropriate frequencies ν_1 and ν_2:

$$\begin{aligned} \mathrm{A} + h\nu_1 &\rightarrow \mathrm{B} \\ \mathrm{B} + h\nu_2 &\rightarrow \mathrm{A} \end{aligned}$$

An advantage of this type of information storage is that it can be erased and rewritten many times by using appropriate light signals. A common group of photochromic molecules used for holographic storage are the azobenzenes, typically methyl red and methyl orange. On irradiation, these change from the *cis*-form to the *trans*-form. Apart from a change in the absorption spectrum of the molecule due to the conformational change, a change in the relative permittivity and refractive index is also often encountered. In addition, the same group can be used to make polarisation holograms. These are formed by attaching modified azobenzene molecules as side groups on polymer chains. These molecular fragments have a dipole moment, and when such a material is irradiated with linearly polarised light, the side chains orient parallel or perpendicular to the electric field vector of the light, causing the initially isotropic polymer matrix to become birefringent with an accompanying large change in refractive index. Irradiation with circularly polarised light can erase the hologram. Rotation of the plane of polarisation of the light can then be used to erase or rewrite data. Because there are no unalterable changes such as bond breaking or polymerisation, these materials are favoured for erase and rewrite storage.

12.7 Fibre Optics

The transmission of light along thin fibres of glass, plastic, or other transparent materials is referred to as *fibre optics*. Data is carried by a series of pulses of light encoded so that information can be stored and retrieved. The transparent optical wave carrier used for communications is silica glass. The light pulses launched into the fibre are constrained to stay within the fibre by total internal reflection. Thus, the core of the fibre, along which light travels, must possess a higher refractive index than the outer surface of the fibre. Moreover, the surface at which the total internal reflection occurs is easily damaged, and needs protection. Both of these objectives are met by providing a *surface cladding* of lower refractive index glass, compared with the core of the fibre. The core and the cladding make up a single glass fibre (Figure 12.32). The cladding should not be confused with an outer plastic protective covering, which has no optical role to play.

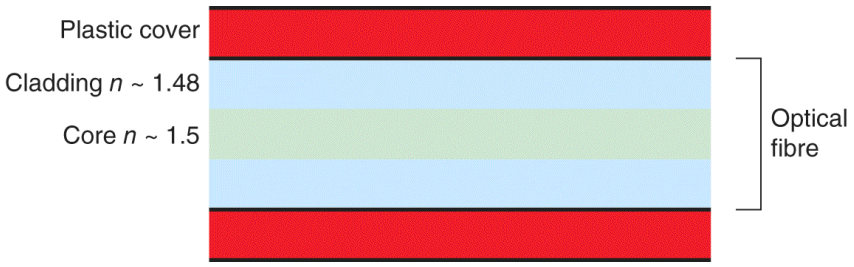


Figure 12.32 The structure of a silica glass optical fibre.

12.7.1 Attenuation in Glass Fibres

Attenuation describes the loss of light as the signal is transmitted along the fibre. This is of major concern, as any degradation of the signal must be minimised. The loss is defined as:

$$\text{loss} = -10 \log_{10}[P(x)/P(0)]$$

where $P(0)$ is the power input, at $x = 0$, and $P(x)$ is the power at a remote point, x. The unit of loss is the decibel, dB. The attenuation is defined as the loss per kilometre, thus:

$$\text{Attenuation} = \{-10 \log_{10}[P(x)/P(0)]\}/x$$

For a material showing an attenuation of 1 dB km^{-1}, an input power of 10 W would give an output power of 7.9 W after 1 km. Ordinary window glass has an attenuation of about 10 000 dB km^{-1}. Attenuation, like dispersion, varies with wavelength. The *spectral response* of a fibre defines the way in which the fibre attenuation changes with the frequency of the radiation being transmitted.

Attenuation is caused by a combination of *absorption* and *scattering* within the glass. *Extrinsic* attenuation arises during processing and may be due to artefacts such as bubbles, particles, impurities, and variable fibre dimensions. Attenuation in early fibres was mainly due to Fe^{2+}, the ion that imparts a greenish tint to window glass. Even a concentration as low as 1 part per million of Fe^{2+} can result in an attenuation of 15 dB km^{-1}. The presence of this and other metal cations is avoided by preparing silica from very high purity chemicals made available via the semiconductor industry. The most important impurity in silica fibres today is hydroxyl (–OH), which arises from water or hydrogen incorporation into the glass during fabrication. An impurity level of 1 part per million can give an attenuation of 10^4 dB km^{-1} at a 1.4 μm signal wavelength.

Intrinsic attenuation is a property of the pure material itself, and cannot be removed by processing. It is the ultimate limit on the performance of the fibre and mainly arises from two factors, Rayleigh scattering and lattice vibrations. Rayleigh scattering (Section 12.5) is due to small inhomogeneities in the glass, which cause changes in refractive index. This variation is an inevitable feature of the non-crystalline state and cannot be removed by processing. Rayleigh scattering is more important at short wavelengths as scattering is proportional to λ^{-4}, where λ is the wavelength of the optical pulse. For any particular glass, most of the factors affecting Rayleigh scattering are constant and cannot be easily

changed. However, materials with a low refractive index and glass transition temperature tend to exhibit low Rayleigh scattering.

Absorption of energy by lattice vibrations, referred to as *phonon absorption*, occurs when the energy of the radiation matches vibration frequencies. This occurs for infrared wavelengths, and converts the signal energy into heat. It is a function of the mass of the atoms in the glass and the strength of the chemical bonds between them and results in a decrease in the transparency of the glass at long wavelengths. Absorption due to *electronic transitions*, mostly at high energies and associated with ultraviolet wavelengths, does not figure significantly in present day applications, but may become important if shorter signal wavelengths are to be used in the future.

By 1979, the best silica fibres showed only intrinsic attenuation and had a loss of about 0.2 dB km^{-1} at 1.5 μm wavelength.

12.7.2 Dispersion and Optical Fibre Design

A short pulse of light launched into a fibre will tend to spread out due to dispersion. In optical fibres the dispersion is defined as the delay between the arrival time of the start of a light pulse and its finish time relative to that of the initial pulse. It is measured at half peak amplitude. If the initial pulse has a spread of t_i seconds at 50% amplitude and the final pulse a spread of t_f seconds at 50% amplitude after having travelled x kilometres, the dispersion is given by:

$$\text{Dispersion} = (t_f - t_i)/x$$

The units of dispersion in optical fibres are ns km^{-1}.

Dispersion will result if the light source is not strictly monochromatic. An initially sharp pulse consisting of a group of wavelengths will spread out as it travels down the fibre, because the refractive index depends on wavelength. Thus, different wavelengths will travel at different speeds. This effect is known as *wavelength dispersion*.

Even with completely monochromatic light, pulse spreading can still occur, because the radiation can take various paths, or *modes*, through the fibre (Figure 12.33). It is apparent that a ray that travels along the axis of a fibre will travel less than one that is continually reflected on its journey. (In fact, the dispersion that results cannot be properly understood in terms of the transmission of light rays, and the various modes are better described in terms of the allowed wave patterns that can travel down the fibre.) The resultant pulse broadening, due to the various modes present, is called *modal* (or *intermodal*) *dispersion*. In order to overcome modal dispersion, a number of different fibre types have evolved.

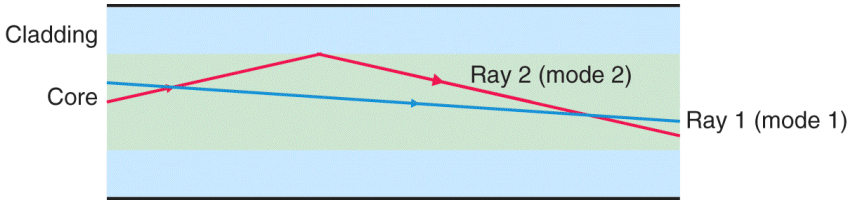


Figure 12.33 Alternative paths that light beams can take in the core of an optical fibre, schematic.

The earliest fibres were called *stepped index multimode* fibres. These fibres have a large core region, allowing many modes to propagate. The ray labelled H (Figure 12.34a) is known as a *high order mode* awhile the ray L is a *low order mode*. Stepped index multimode fibres are easy to make and join, but have a lower performance compared with those now described. The first advance on stepped index fibres was the *graded index* fibre. In this design, the refractive index of the fibre varies smoothly from high at the centre to low at the periphery of the core region (Figure 12.34b). The refractive index gradient means that light travels faster as it approaches the edge regions of the fibre. The velocity of mode B will vary smoothly from lowest at the fibre centre to greatest near to the fibre edge and be similar, on average to mode A. Modal dispersion is thus minimised. For best results *monomode fibres* (Figure 12.34c) are now used. The number of possible modes is reduced by reducing the diameter of the core. When the core diameter reaches 10 μm or less, only one mode can propagate and, in principle,

modal dispersion is zero for these fibres. Monomode fibres have a high performance but are harder to make and join.

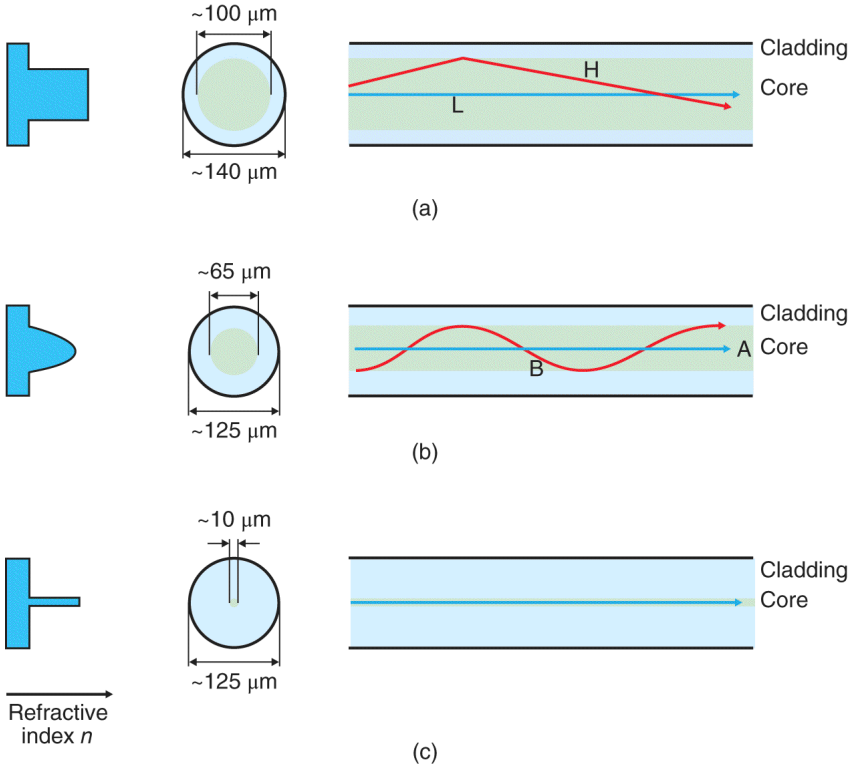


Figure 12.34 Types of optical fibre: (a) stepped index; (b) graded index; (c) monomode.

The material used for optical communications fibre is highly purified silica glass. The cladding and core regions are created by doping with carefully chosen impurities. A commonly used production method starts with a tube of pure silica glass. Layers of germanium dioxide are laid down in the centre of the tube. Germanium dioxide is chemically and physically very similar to silica and readily forms a solid solution with the silica glass. As the germanium atoms are heavier than silicon atoms, they increase the refractive index of the doped inner region relative to the undoped outer region of the tube. When a sufficient amount of GeO_2 has been laid down, the tube is heated until it collapses into a solid rod called a *preform* with a germanium-rich higher refractive index core and a lower refractive index periphery. The preform is drawn into a fibre, and because of the nature of the way in which glass flows at elevated temperatures, the refractive index profile of the preform is maintained in the fibre.

Although fibres in commercial use are made of silica glass, it is not perfect. The dispersion is lowest at 1.3 µm, but the minimum attenuation occurs at 1.5 µm, leading to some sacrifice of performance irrespective of the signal wavelength chosen. The search for new materials to resolve this conflict continues in many laboratories.

12.7.3 Optical Amplification

The amplification of signals in fibre optic transmission is achieved in a stretch of about 30 m of monomode fibre core containing just a few hundred parts per million of Er^{3+} (Figure 12.35a). This section of the fibre is illuminated by a semiconductor diode laser at a wavelength related to that of the carrier signal; the commonest being 980 and 1480 nm. The erbium ions have the ability to transfer energy from the laser to the signal pulses as they traverse this section of fibre. Illumination of the erbium containing section of fibre with energy of wavelength 980 nm excites the ions from the ground state ($^4I_{15/2}$) to the upper state ($^4I_{11/2}$) from whence they rapidly decay to the $^4I_{13/2}$ level (Figure 12.35b). The use of radiation of 1480 nm wavelength excites the Er^{3+} ions directly from the ground state to the $^4I_{13/2}$ level. This illumination process is referred to as *pumping* and the laser

involved as the *pump*. The excited state has quite a long lifetime, and so a passing light pulse, with a wavelength close to 1480 nm, can empty it via *stimulated emission* (Section 12.2.6). This achieves signal amplification while retaining the coherence of the pulse constituting the signal.

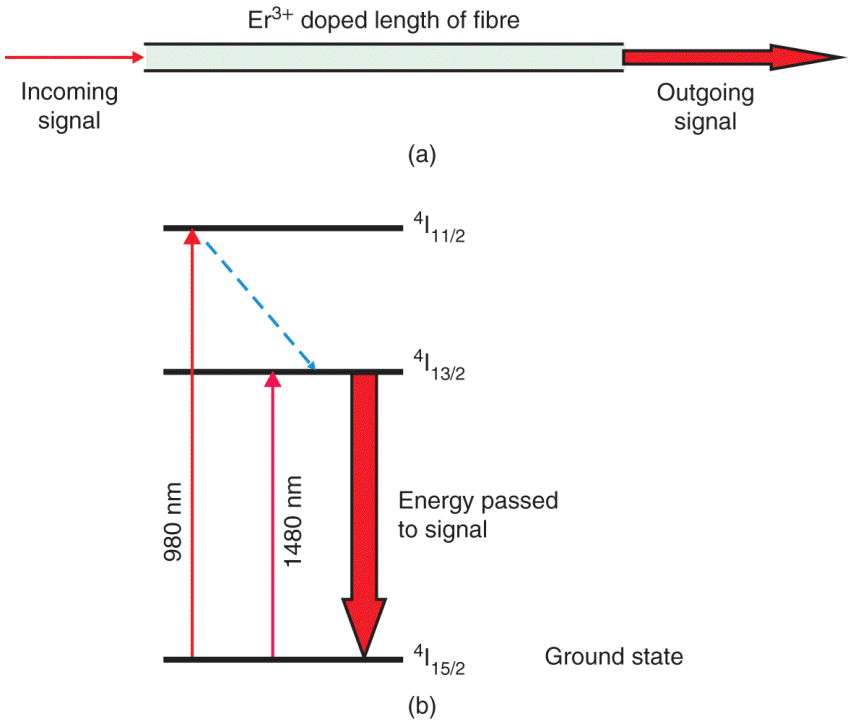


Figure 12.35 Signal amplification: an incoming weak signal is amplified on passage through a length of fibre doped with Er^{3+}; (b) energy levels and radiation transitions involved in amplification.

12.8 Energy Conversion

12.8.1 Photoconductivity and Photovoltaic Solar Cells

If radiation of a suitable wavelength falls on a semiconductor, it will excite electrons across the band gap giving rise to a voltage and a related increase in conductivity. Solids that behave in this way are called *photovoltaic materials*. The magnitude of this *photovoltaic* or *photoconductive effect* is roughly proportional to the light intensity. It is used in light metres, exposure metres, and automatic shutters in cameras. The first photographic exposure metres used selenium, cadmium sulfide, or silicon. In the case of selenium, the photovoltage is large enough to be measured directly and converted to an exposure value. Cadmium sulfide and silicon need voltage amplification, and they require a power source, usually a battery, to give a reading. In these materials a voltage is applied across the semiconductor and illumination levels are measured as an increase in conductivity.

A p–n junction can act in a similar way to a single piece of semiconductor. However, the control afforded by the junction makes the device, a *photodiode*, far more flexible. As a result, photodiodes are widely used, especially in *solar cells*. A solar cell is a specialist large area p–n junction with a depletion region approximately 500 nm thick. (Solar cells must have a large area, to collect as much sunlight as possible.) In addition, the normal built-in potential that exists across the junction, due to the space charge, is engineered to be high. The junction is not connected to any external power source. Holes and electrons produced in the junction region by sunlight are swept across the depletion region by the high built-in space charge present, the electrons going from the p to n region and the holes from the n to p region. This process, called *drift*, makes the p region more positive and the n region more negative, produces a photovoltage, and causes a photocurrent I to flow (Figure 12.36a). Should an external load, R, be connected, some current can pass through it, and so do useful work.

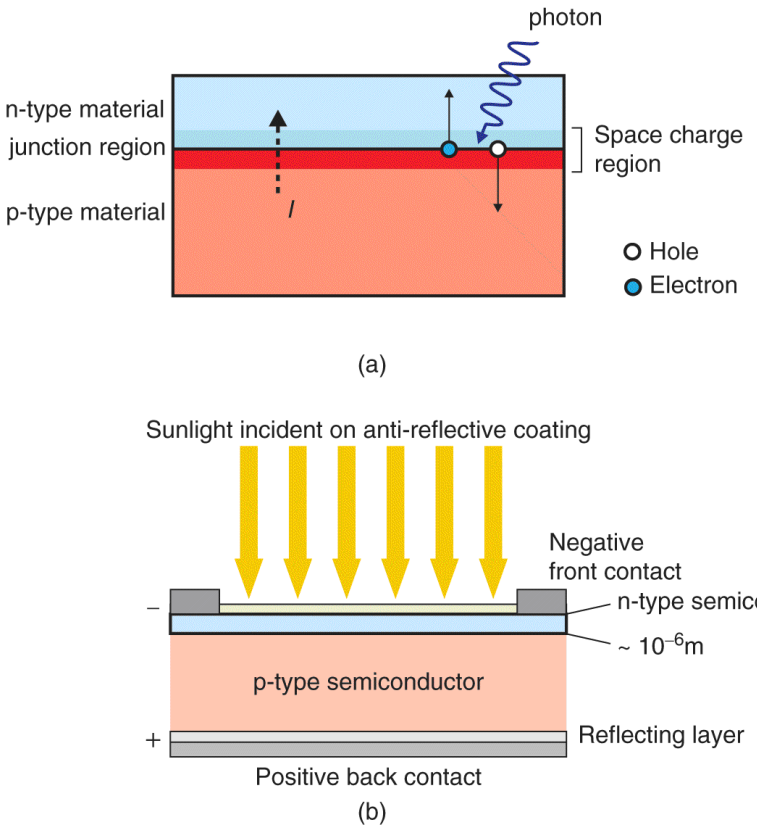


Figure 12.36 Solar cells: (a) a photon incident upon a p–n junction generates an electron–hole pair; (b) construction of an operating cell, schematic.

Impurities and defects in the semiconductor layers, such as point defects, dislocations, and grain boundaries, are detrimental to performance and high purity materials, although costly, are mandatory.

It is also clear that the band gap of the solar cell materials chosen must utilise as much of the wavelength spread available in sunlight as possible. At sea level, the energy available in sunlight amounts to approximately 1000 W m^{-2} and has a wavelength spread is approximately from 400 to 2500 nm, with a peak in the yellow-green at 550 nm. Moreover, indirect band gap materials have a lower efficiency than direct band gap materials and are usually avoided.

The cells that currently show the highest efficiency and are commonly seen on rooftops as solar energy generators use doped silicon as the active material. Single crystal silicon, an indirect band gap solid, is unsuitable and is replaced by amorphous silicon, which behaves as a direct band gap material. The construction of a working device, such as a typical silicon solar cell, is designed to optimise efficiency in light collection and voltage generation. An obvious requirement is the provision of an antireflection surface on the sunlight-receiving side, to maximise the number of photons reaching the semiconductor. It is also necessary that the photons are absorbed and create electron–hole pairs close to the p–n junction. (Electron–hole pairs created elsewhere have to diffuse to the junction region, and unless the materials are of high purity and defect free, recombination is likely.) To achieve this, the first semiconductor layer encountered by the photons is thin, so that an optimum number of photons reach the junction layer. Finally, an underlying reflective layer redirects any photons that pass straight through the semiconductor layers back towards the junction region is added (Figure 12.36b).

Solar concentrators, mirrors, or lenses that focus the sunlight onto the photoactive layers are widely used to increase efficiency, as are mobile systems that are able to follow the motion of the sun throughout the day. Efficiency can also be increased by using stacked arrays of cells to absorb as much as possible of the solar spectrum. For example, a stack of cells using $GaInP_2$, GaAs, and Ge is able to utilise photons from 590 to 1200 nm.

12.8.2 Dye-Sensitised Solar Cells

The method of conversion of sunlight to energy in a conventional solar cell is quite different to that of most importance on the Earth, photosynthesis, where the central reactions are oxidation and reduction. *Photoelectrochemical cells*, of which *dye-sensitised solar cells* (DSSCs), also called *Grätzel cells*, are an important example, aim to mimic this process. The task of harvesting the light is left to a sensitiser, which is a dye molecule, and the carrier transport task is allocated to a semiconductor. Because the charge separation takes place in the dye, the purity and defect structure of the semiconductor are not crucial to satisfactory operation.

The reactions in the cell are:

(i) Excitation of the sensitiser dye, S, by a photon:

$$S + h\nu \rightarrow S^*$$

(ii) The excited sensitiser, S*, loses an electron, which moves into the conduction band of the semiconductor:

$$S^* \rightarrow S^+ + e^-(\text{semiconductor})$$

(iii) The electron moves through the conduction band of the semiconductor to the transparent conducting anode, also called the *working electrode*, which acts as the electron collector. Thereafter, electrons traverse the external circuit to arrive at the cathode, also called the *counter electrode*.

(iv) Electrons arriving at the counter electrode reduce a redox couple, R/R^-, usually in a liquid electrolyte:

$$R(aq) + e^- \rightarrow R^-(aq)$$

where (aq) represents an aqueous solvent.

(v) The sensitiser is regenerated by reaction with the reduced half of the redox couple:

$$S^+ + R^- \rightarrow S + R$$

(Figure 12.37a).

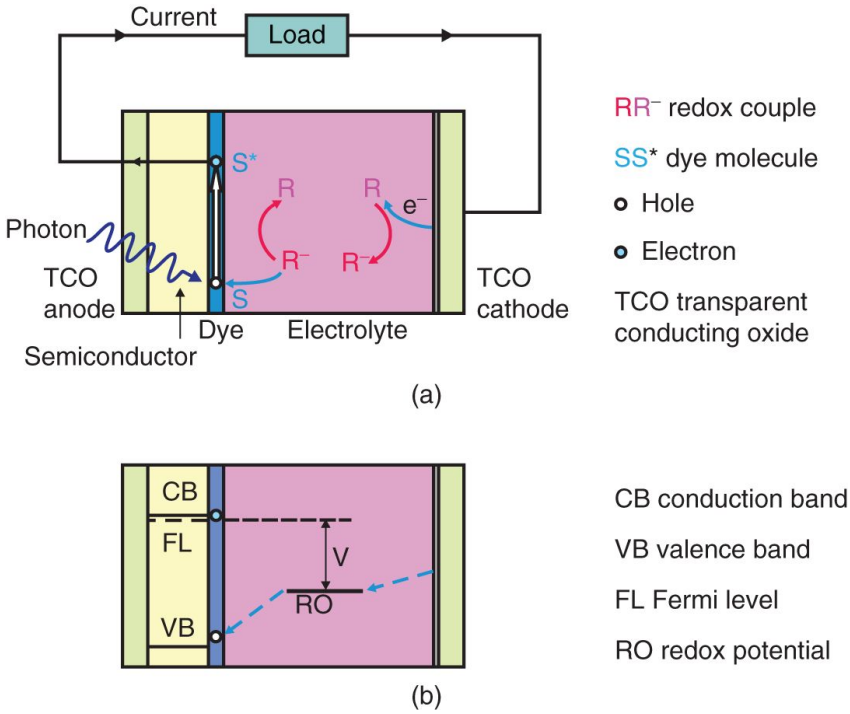


Figure 12.37 Dye-sensitised solar cell, schematic: (a) reactions taking place in a cell; (b) energy levels in cell.

A large number of different dyes have been tried in the role of sensitiser, in conjunction with a variety of inorganic oxides, including ZnO and Nb_2O_5 as the semiconductor. Currently the highest efficiency is obtained with cells using dye molecules containing Ru(II) such as *cis*-dithiocyanatobis-2,2′-bipyridine-4-COOH, 4′-COO–Ru(II) in combination with nanocrystalline anatase (TiO_2) as the semiconductor.

The charge states on the dye correspond formally to the conversion of Ru^{2+} to Ru^{3+} via photon interaction:

$$Ru^{2+} + h\nu \rightarrow Ru^{3+} + e^-$$

The dye is absorbed onto the surfaces of the crystallites to give a large surface area whist maintaining compact electrode geometry. The transparent conducting oxide electrodes are usually tin oxide doped with fluorine, SnO_2:F. The main redox couple chosen is iodide–triiodide in solution:

$$I^{3-}(aq) + 2e^- \rightarrow 3I^-(aq)$$

In order to catalyse the oxidation- (oxidation-reduction) reduction equilibria in the electrolyte, the cathode electrode is coated with a thin layer of platinum.

The cell output depends upon the relative positions of the energy levels in the adjoining components, which must be matched for optimum efficiency. In the present cell design, this is given by the difference between the redox potential of the oxidation–reduction couple chosen and the Fermi level of the semiconductor (Figure 12.37b). For the triiodide–iodide couple the redox potential is +0.54 V. The Fermi level of the nanostructured anatase is about −0.4 V, so that the cell voltage is approximately 0.54 V + 0.4 V, i.e. 0.94 V.

There is much current research directed towards improving the dyes used in these cells. Similarly there is interest in replacing the expensive platinum cathode with an organic conductor such as PEDOT, poly(3.4-ethylenedioxythiophene).

12.8.3 Perovskite Solar Cells

An important component of later types of DSSCs was a thin layer of nanocrystalline anatase, a polymorph of TiO_2, as an electron conductor, which was deposited upon a thin film of a transparent conducting oxide acting as the anode on a glass substrate. A large number of different dyes have been tried in the role of sensitiser and a considerable breakthrough occurred in 2009 when the dye was replaced by a layer of the organic–inorganic perovskite methylammonium lead iodide, $CH_3NH_3PbI_3$ (Figure 12.38). This has the typical ABX_3 perovskite structure (Figure 3.11), consisting of a framework of PbI_3 octahedra, with the nominal ionic states Pb^{2+} and I^-, but the inorganic A cation is replaced by the organic methylammonium ion $CH_3NH_3^-$. The conversion efficiency of these earliest cells, which had the same architecture as the DSSCs (Figure 12.39a), was approximately 4%, but it was immediately clear that this preliminary figure could be substantially improved by device engineering.

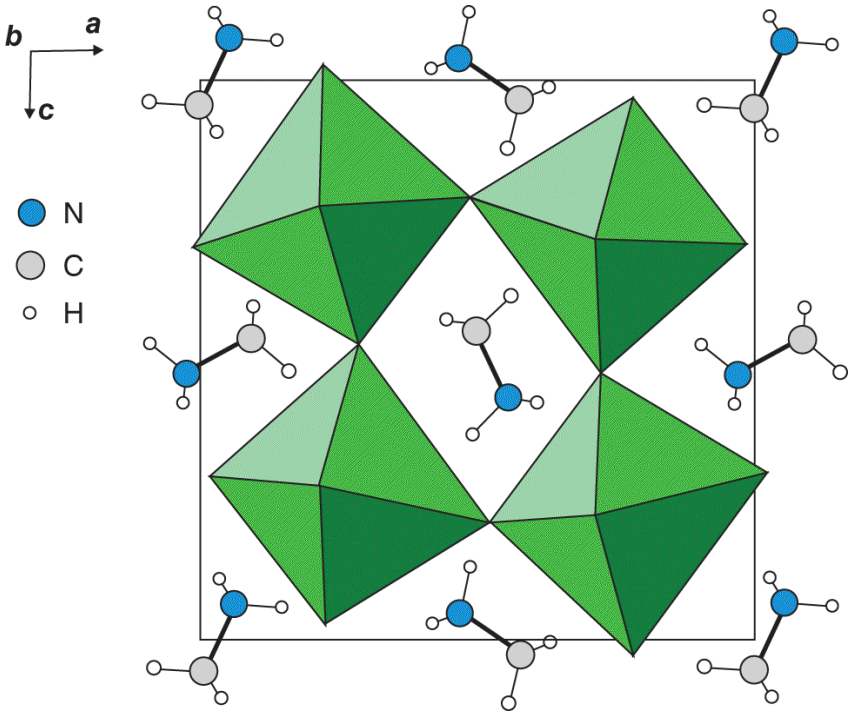


Figure 12.38 The structure of $(CH_3NH_3)PbCl_3$ ($MAPbCl_3$). In the layers preceding and succeeding the layer shown, the methyl ammonium molecules are rotated by 180°.

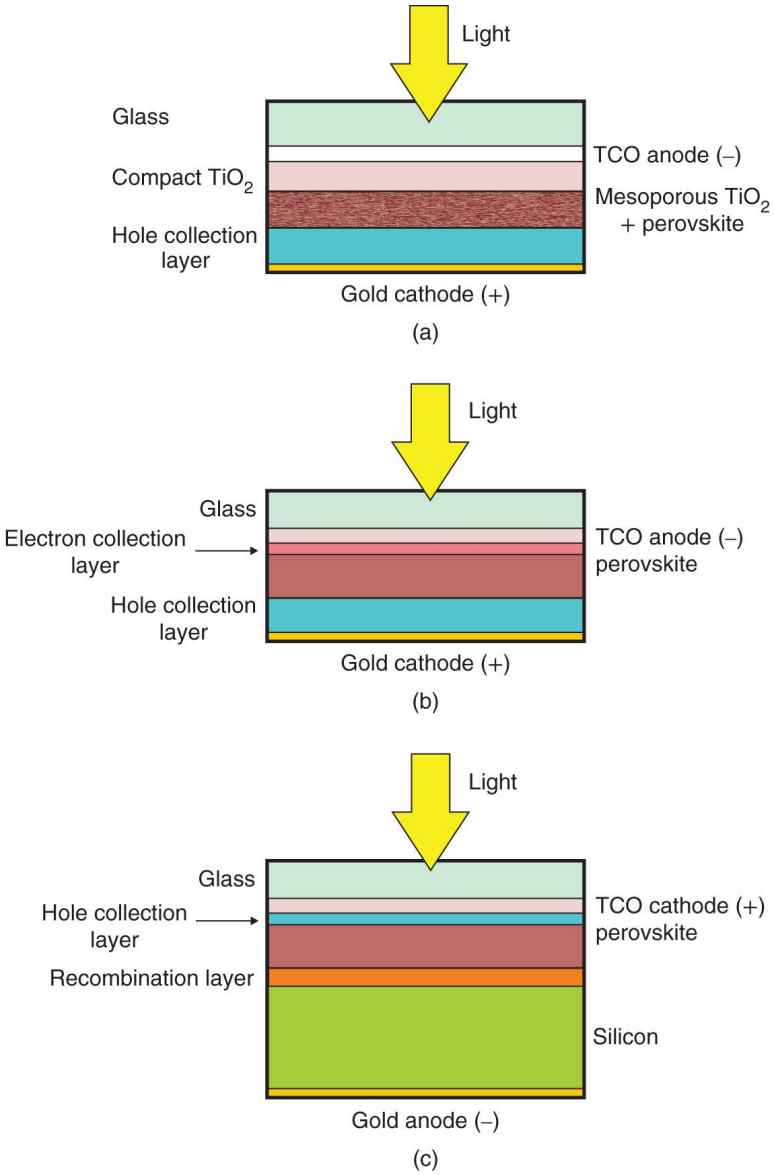


Figure 12.39 Perovskite solar cells, schematic: (a) dye-sensitised solar cell architecture; (b) conventional solar cell architecture; (c) tandem silicon/perovskite architecture. TCO = transparent conducting oxide.

The efficiency of these perovskite sensitised cells is superior to the dye alternatives, and the perovskite component offers a number of other advantages. The main one of these is that the electron and hole conducting ability of the perovskite is high enough to allow for the omission of the mesoporous TiO_2 layer, allowing a traditional photovoltaic cell design to be employed (Figure 12.39b). In addition, thin films of the active phase can be laid down from water solution, making processing relatively inexpensive. In these cells, efficiency of over 25% has been achieved to date.

In view of the flexibility of the perovskite structure, many modifications are possible. The band gap of these phases, and thus the absorption spectrum of the perovskite and its light harvesting efficiency can be tuned by doping the organic and the halogen sites. The perovskites $CH_3NH_3Pb(I_{1-x}Br_x)_3$ have a band gap variation given by:

$$E_g = 1.57 + 0.39x + 0.33x^2$$

giving a smooth band gap variation from 1.57 (x = 0) to 2.29 (x = 1) and hence a smooth change in the optical absorption spectrum, from infrared into blue, over this composition range. Similarly the mixed methylammonium $(CH_3NH_3)^+$ and formamidinium ((HN□$CHNH_3)^+$ perovskites $(CH_3NH_3)_x$(HN□$CHNH_3)_{1-x}PbI_3$ show a decrease in light harvesting at red wavelengths as the formamidinium amount increases. Other related phases, such as the methylenediammonium $[CH(NH_2)_2{}^+]$, perovskites, and mixed Pb^{2+}, Sn^{2+}, I^-, Br^-, and Cl^- phases are all under investigation.

It is important to note, at this stage, that the microstructures of these materials in working cells have not yet been fully understood. In particular, the perovskite film is sensitive to water vapour and humidity changes cause a significant degradation in properties and performance. Thus, these cells must be sealed from atmospheric water if they are to have a commercially viable lifetime. As with conventional solar cells, defects in the perovskite layers are detrimental and studies on the reduction of defects and the use of single crystal perovskite layers is ongoing. In addition, the construction of stacks of differing types of solar cells will give an improved output. The most investigated of these at present is a perovskite/silicon solar cell tandem (Figure 12.39c), which is anticipated to have better light harvesting potential and correspondingly higher efficiencies.

12.9 Nanostructures

The exploitation of the optical aspects of nanostructures, thin films, and particles is a rapidly expanding area of research and application. These materials are being explored for photocatalysis, solar cells, flat lenses, anticounterfeiting and security marking, displays, and a myriad of other applications. The optical properties of thin films are introduced via a description of quantum wells, a continuation of Section 11.6. The optical properties of nanoparticles depend upon two factors: (i) the size of the particle and (ii) the pattern of organisation of particles, usually on a surface. In this section, the way in the optical properties of materials with nano-dimensions differs from those of the bulk solid are briefly outlined. References to review articles detailing a plethora of applications are given in the Further Reading.

12.9.1 The Optical Properties of Quantum Wells

In a quantum well, which approaches nanodimensions in one direction, the electrons and holes occupy electron and hole subbands (Figure 12.40). When electrons in the upper energy levels drop to the lower levels in *interband transitions*, a photon is emitted. The energy separation of the subbands is greater than the energy gap of bulk material so the photons will be of shorter wavelength than those associated with the bulk semiconductor and are said to be *blue shifted* compared to the bulk.

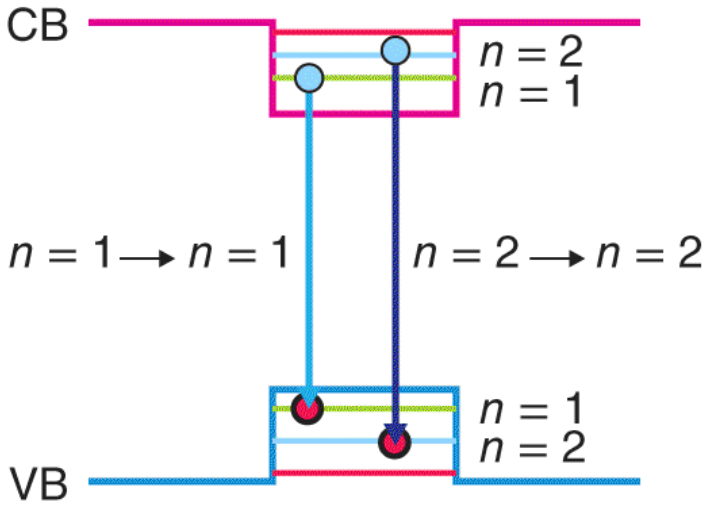


Figure 12.40 Interband transitions (schematic) between electron (upper) and hole (lower) subbands.

The photon energy derived from an interband transition is:

$$E(\text{photon}) = E_g + E_e + E_h$$

where E_g, E_e, and E_h represent the bulk band gap energy and the energies of the electron and hole subbands. The photon energy is given by:

$$E(\text{photon}) = E_g + [h^2/8a^2][n^2/m_e^* + n^2/m_h^*]$$

where h is Planck's constant, a is the dimension of the quantum well, m_e^* is the electron effective mass, and m_h^* the hole effective mass. In the approximation that the effective mass of the electron and the hole are identical:

$$E(\text{photon}) = E_g + [n^2h^2/4a^2m^*]$$

The selection rule for the transition is $\Delta n = 0$; that is, transitions only take place between levels with the same quantum number. (As with all selection rules, these are never perfectly obeyed, and

transitions between levels with differing n values do occur infrequently, giving rise to weak lines in the emission spectrum.)

Electrons can also be excited from one electron level, say, $n = 1$, to another electron level, say $n = 2$, both levels lying in the *electron subband*. Holes can make similar transitions between levels in the *hole subband*. These transitions, which give rise to extra peaks in the emission spectrum, are known as *intersubband transitions*.

Because the dimensions of the quantum well can be changed, the emission spectrum can be varied or *tuned*. This feature is called *band gap engineering*.

Quantum well structures are widely used in LEDs and laser diodes to improve device performance. They do this in a number of ways; by confining electrons and holes in a limited space, so that recombination is more likely, and by guiding the output photons by virtue of the differing refractive indices of the materials. Typical of these device structures is the single quantum well (SQW) structure used in the first green emitting LEDs (Figure 12.41). A change in the composition of the SQW active layer allows the colour emission to vary between 450 nm, blue to 600 nm yellow.

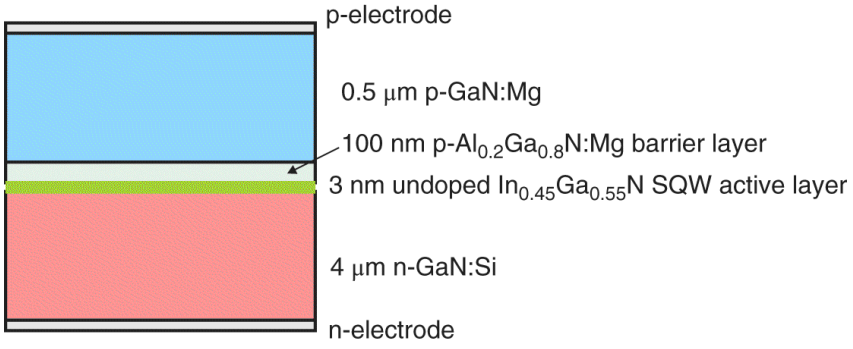


Figure 12.41 Green-emitting single quantum well (SQW) active layer LED.

12.9.2 The Optical Properties of Nanoparticles

The optical properties of photoluminescent nanoparticles, or quantum dots, have been extensively investigated because they emit fluorescent light that is a precise function of the dimensions of the particle. For example, CdSe quantum dots of radius 2.9 nm emit at approximately 555 nm, of radius 3.4 nm emit at approximately 580 nm and of radius 4.7 nm emit at approximately 625 nm (Figure 12.42).

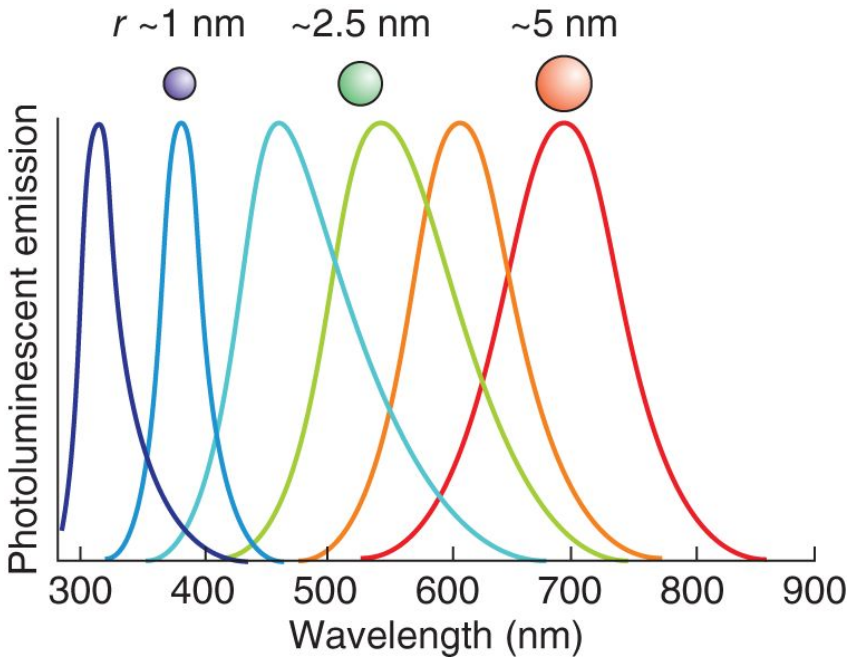


Figure 12.42 Photoluminescent colours emitted by CdS quantum dots.

The colour variation comes about because the conduction and valence bands of the bulk semiconductor are transformed from the continuous energy bands of the bulk into a closely lying set of typically atomic discrete energy levels as the dimensions of the particle approach the atomic scale. Moreover, the energy gap between the highest energy level in the valence band group drops in energy while the lowest energy level of the conduction band group increases in energy, so that the effective band gap appears to increase steadily as the dot size falls (Figure 12.43a). To produce light, electrons are excited from the lower band to the upper band with ultraviolet radiation. The electrons in higher energy levels subsequently loose energy by non-radiative transitions to end in the lowest energy level of the upper set. A photon is them emitted as the electron drops to the topmost energy level of the lower set (Figure 12.43b).

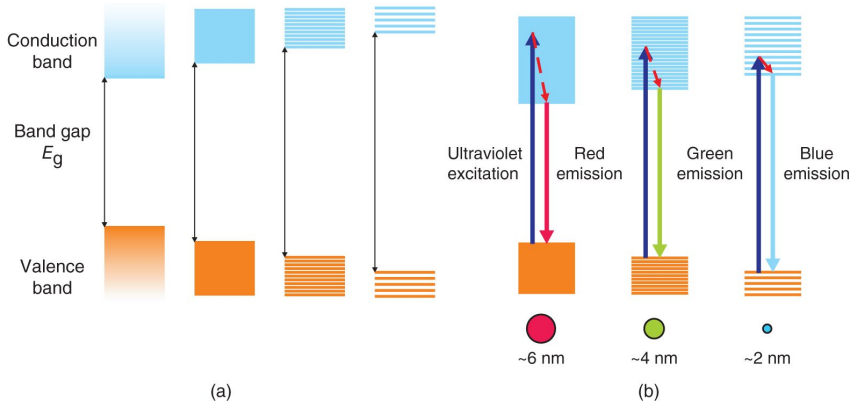


Figure 12.43 Quantum dot colours: (a) the change in energy level structure of a quantum dot as the size decreases; (b) emission colours of different diameter dots (schematic).

There are many potential applications for photoluminescent quantum dots, because they constitute minute but very bright lamps that can be activated at will by an ultraviolet or blue light probe. Moreover, the colour output is pure in the sense that the emission spectrum is narrow. Applications include the biological imaging of processes in living cells and production of quantum dot lasers and white LEDs.

Simple quantum dots have a number of shortcomings. The relatively large surface area of the dots reduces the light generating efficiency considerably. This is in part due to the fact that many of the bonds on the surface atoms are not complete. These *dangling bonds* serve to trap electrons and holes so that the excited dot loses energy other than by emission of photons. The surface can thus be considered as a defect-rich region that interferes with the mechanism of luminescence. Additionally, for biological imaging of processes in living cells, not only is the luminous efficiency important, but also the quantum dots must be treated so that they are water soluble and non-toxic.

The commonest approach to overcoming both of these difficulties is to coat the quantum dot with a thin covering of another material to make *core–shell* structures. For example, CdSe nanoparticles surrounded by a thin shell of ZnS have modified light emitting properties, while CdS nanoparticles coated with silica or organic surfactants are water soluble and less toxic.

A curious feature of core–shell quantum dots is that they sometimes cease to emit light for a period before returning to normal, referred to as *blinking*. In quantum dots the cause is believed to be due to the shell acting as a hole or electron trap. When the core of the dot is excited by ultraviolet radiation, an electron–hole pair is generated. It appears that if the electron or hole is able to migrate into the shell and become trapped at a surface site photon emission is suppressed. In effect the core becomes positively charged and although the core can still absorb photons, internal energy loss takes over, and the excess is lost as heat. Eventually the trapped electron re-unites with the core, after which normal photoluminescence is restored. This on–off sequence is blinking.

12.9.3 Nanoparticle Arrays

One of the earliest ways of exploiting arrays of light emitting quantum dots was in displays, typically television screens. The term quantum dot light emitting device (QLED) is used for these devices, or, more explicitly, an electroluminescent quantum dot light emitting device (EL QD LED). The quantum dots are arranged as a planar film. Electrons are injected into quantum dots via a cathode and an electron transporting layer and holes via an anode and a hole transporting layer (Figure 12.44). Radiative combination of electrons and holes occurs in the dot and light is emitted through a transparent electrode. As each size of quantum dot gives out a light of a specific colour an array of dots of different sizes will directly produce white light when viewed from a distance. The colours can be manipulated by varying the sizes of the quantum dots in the emitting layer.

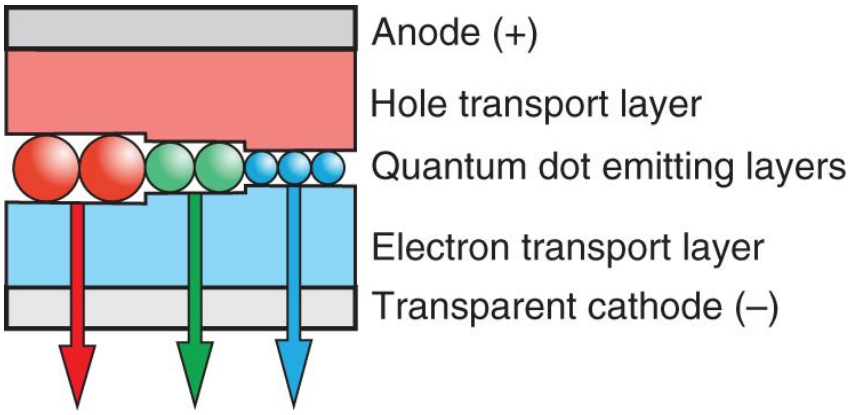


Figure 12.44 A QLED screen (EL QD LED screen) schematic.

Early versions of these light emitting devices used CdSe quantum dots, but blinking caused problems for continuous operation and more complex core/shell dots are now employed. Similarly, the electron and hole transporting layers were once inorganic layers, but now organic films are often preferred. As with conventional LEDs, refractive index matching and design architecture are being modified to improve light emission and contrast.

The optical behaviour of three-dimensional crystal-like arrays of nanoparticles can be treated by assuming that they behave like photonic crystals, although the diffraction conditions become more complex when the spacing falls below the wavelength of light. Similar, but two-dimensional arrays of nanoparticles can also be used to successfully manipulate light and modify the optical properties of the

surfaces they are deposited on. This idea was evolved in night-flying insects some millions of years ago. The eye facets of most day-flying insects such as bees and dragonflies are smooth, but certain night-flying moths and a few butterflies, which have to optimise light gathering, have facets that are covered with an array of nano-sized bumps that mimic an array of nanoparticles, to form an antireflection layer (Figure 12.45). In order for such surfaces to act as in this way the protuberances must be small enough not to be resolved by light, which means in practice that the bumps must be separated by less than half the wavelength of the light. If this is not so, the array will act as a diffraction grating. The surface architecture must also not be confused with surface roughness, which will increase diffuse reflection compared to specular reflection, but not decrease the total amount of light returned towards the source. As these ordered patterns of nano-projections were first recorded on the surface of the eyes of some moths, these types of surface antireflection coatings are called *moth-eye antireflection coatings*. The more formal name for moth-eye antireflection coating is *ultrahigh spatial-frequency surface relief grating*.

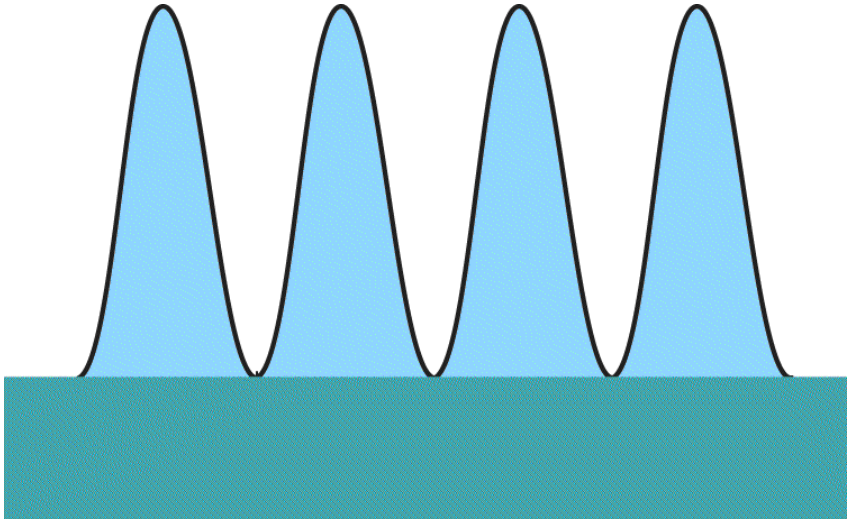


Figure 12.45 A 'moth-eye' surface structure.

The antireflective properties of such surfaces can be determined assuming that the refractive index varies smoothly from the surroundings (air) to that of the inner surface to form a graded refractive index (GRIN) coating. This is carried out by (theoretically) dividing the nanostructural surface array into slices parallel to the underlying substrate, estimating the average refractive index and reflectivity of each layer and then summing over the slices to determine the reflectivity of the whole structure. The optical properties of these surface features can also be considered in terms of a two-dimensional diffraction grating. In essence a moth-eye surface grating is a reflection grating. To optimise night vision, it is important to *minimise* all reflected light. This means that the diffracted orders must be suppressed so that the surface is completely anti-reflecting. Moreover, in order to determine the intensity of this zero-order reflection for a moth-eye structure, it is necessary to calculate the intensity as a function of the surface profile, which may be square, sinusoidal, or irregular, and also take into account the depth of the grooves, the angle of incidence and polarisation of the light, so that numerical estimates rather than analytic solutions are derived.

These structures can be copied in the laboratory by placing an array of suitable nanoparticles on a smooth surface, adjusting both the spacing and the refractive index of the nanoparticles to tailor the observed and desired optical properties.

Further Reading

General:

Fox, M. (2001). *Optical Properties of Solids*. Oxford: Oxford University Press.

Much of the Material in this Chapter is Covered in Greater Detail in:

Tilley, R.J.D. (2020). *Colour and the Optical Properties of Materials*, 3rd edition. Chichester: Wiley.

The Properties of Light with Respect to Colour are Found in:

Arcand, K. and Watzke, M. (2015). *Light: The Visible Spectrum and Beyond*. New York: Hachette Book Group.

Nassau, K. (2001). *The Physics and Chemistry of Colour*, 2nde. New York: John Wiley & Sons.

Watson, B. (2016). *Light: A Radiant History from Creation to the Quantum*. New York: Bloomsbury.

The Engineering Aspects of Optical Fibres are Described by:

Hecht, J. (1999). *Understanding Fiber Optics*, 3rde. Upper Saddle River, NJ: Prentice Hall.

Perovskite Solar Cells are Described in:

Chang, X., Yang, S., Cao, B. et al. (2020). Single crystal perovskite solar cells: development and perspectives. *Adv. Funct. Mater.* 30: 1905021.

Kim, J.Y., Lee, J.-W., Jung, H.S. et al. (2020). High-efficiency perovskite solar cells. *Chem. Rev.* 120: 7867.

Ono, L.K., Liu, S., and Qi, Y. (2020). Reducing detrimental defects for high performance metal halide perovskite solar cells. *Angew. Chem. Int. Ed.* 59: 6676.

For Nanostructures and Surfaces See the Following Review Articles and References Therein:

Kahmann, S., Shulga, A., and Loi, M.A. (2020). Quantum dot light emitting transistors. *Adv. Funct. Mater.* 30: 1904174.

Lassaline, N., Brechbühler, R., Vonk, S.J.W. et al. (2020). Optical Fourier surfaces. *Nature* 582: 506.

Ren, W., Lin, G., Clarke, C. et al. (2020). Optical nanomaterials and enabling technologies for high-security-level anticounterfeiting. *Adv. Mater.* 32: 1901430.

Selopal, G.S., Zhao, H., Wang, Z.M., and Rosei, F. (2020). Core/shell quantum dots solar cells. *Adv. Funct. Mater.* 30: 1908762.

Wang, W. and Qi, L. (2019). Light management with patterned micro- and nanostructured arrays for photocatalysis, photovoltaics, and optoelectronic and optical devices. *Adv. Funct. Mater.* 29: 1807275.

Problems and Exercises

Quick Quiz

12.1 The long wavelength part of the electromagnetic spectrum is associated with:

a. Radiowaves.

b. X-rays.

c. Infrared radiation.

12.2 The short wavelength region of the visible spectrum is associated with the colour:

a. Red.

b. Violet.

c. Green.

12.3 A beam of light is said to be coherent when:

a. All of the waves have the same wavelength.

b. All of the waves are in step.

c. All of the waves travel at the same speed.

12.4 The emission of light by a solid at high temperatures is called:

a. Luminescence.

b. Incandescence.

c. Phosphorescence.

12.5 The light produced by spontaneous emission is:

a. Coherent.

b. Incoherent.

c. Partly coherent.

12.6 The centres that produce the laser action in a ruby crystal are:

a. Cr^{3+} ions.

b. Al^{3+} ions.

c. O^{2-} ions.

12.7 Electron transitions involved in colour (optical transitions) are:

a. Spin allowed transitions.

b. Phonon assisted transitions.

c. Radiationless transitions.

12.8 The four-level neodymium laser utilises:

a. p electron transitions.

b. d electron transitions.

c. f electron transitions.

12.9 Mixing the three additive primary colours gives:

a. Black.

b. White.

c. Colourless.

12.10 Mixing the three subtractive primary colours gives:

a. Black.

b. White.

c. Colourless.

12.11 Which of the following subtractive primary colours absorbs *green*?

a. Cyan.

b. Yellow.

c. Magenta.

12.12 A ray of light passes from a material A, with a refractive index 1.33 to material B, with refractive index 1.5. The wavelength of the light is:

a. Longer in A than B.

b. Longer in B than A.

c. The same in both materials.

12.13 The optical thickness of a film of transparent material is:

a. Less than the physical thickness.

b. Greater than the physical thickness.

c. The same as the physical thickness.

12.14 The refractive index of a transparent material is mainly due to:

a. Ions in the material.

b. Dipoles in the material.

c. Electrons in the material.

12.15 The refractive index of a transparent material:

a. Increases as the wavelength of the light increases.

b. Decreases as the wavelength of the light increases.

c. Does not change as the wavelength of the light changes.

12.16 A simple glass lens is used to form an image of a white circular object. The edges of the image appear coloured. Which colour is outermost?

a. White.

b. Violet.

c. Red.

12.17 Rayleigh scattering applies to particles with a diameter less than:

a. 1/2 the wavelength of the radiation.

b. 1/3 the wavelength of the radiation.

c. 1/10 the wavelength of the radiation.

12.18 Rayleigh scattering obscures a distant object. It will be clearer if imaged in:

a. Ultraviolet light.

b. Infrared light.

c. No difference between the two.

12.19 Diffraction is a form of scattering that occurs when the light has a wavelength:

a. Comparable in size to the object.

b. Much larger than the object.

c. Much smaller than the object.

12.20 A beam of white light is diffracted during passage through a small circular aperture. The colour of the diffracted rings will have:

a. Red on the inside.

b. Violet on the inside.

c. White on the inside.

12.21 An opal strongly diffracts green light of wavelength 550 nm. The approximate diameter of the silica spheres in the opal is:

a. 110 nm.

b. 220 nm.

c. 550 nm.

12.22 Play-back of a hologram with a beam of light results in:

a. One output beam.

b. Two output beams.

c. Three output beams.

12.23 The cladding of an optical fibre consists of:

a. Air.

b. Plastic.

c. Glass.

12.24 Intrinsic attenuation in an optical fibre can be caused by:

a. Minute air bubbles in the glass.

b. Impurities in the glass.

c. Density fluctuations in the glass.

12.25 The 'spreading out' of a monochromatic pulse in an optical communications fibre is called:

a. Pulse dispersion.

b. Modal dispersion.

c. Wavelength dispersion.

12.26 Amplification of signals in optical communications fibres uses:

a. Yttrium ions doped into the glass.

b. Ytterbium ions doped into the glass.

c. Erbium ions doped into the glass.

12.27 The energy of a photon emitted from a quantum well is:

a. Less than the band gap of the well.

b. Equal to the band gap of the well.

c. Greater than the band gap of the well.

12.28 As the size of a quantum dot increases, the photoluminescent wavelength emitted tends to:

a. Increase.

b. Remain the same.

c. Decrease.

Calculations and Questions

12.1 Calculate the frequency and energy of photons associated with wavelengths 425, 575, and 630 nm. What colours are attributed to these wavelengths?

12.2 Light of wavelength 400 nm is shone through a gas of absorbing molecules. Calculate the energy absorbed by 1 mol of gas if each molecule absorbs one photon.

12.3 Carbon dioxide, CO_2, is a greenhouse gas that absorbs infrared radiation escaping from the Earth. What is the energy per mole absorbed by the gas if each CO_2 molecule absorbs one photon of wavelength 15 μm?

12.4 The energy required to break the bond linking the two oxygen atoms in a molecule of O_2 is 495 kJ mol^{-1}. What is the longest wavelength light that could cause this decomposition to occur?

12.5 The energy required to dissociate ozone, O_3, into O_2 plus O is 142.7 kJ mol^{-1}. What is the longest wavelength light that will dissociate ozone in the upper atmosphere?

12.6 Calculate the wavelength at which the rates of spontaneous and stimulated emission become equal at 300 K.

12.7 7 The optical transitions in ruby are: $^4T_{2g} \leftarrow {}^4A_{2g}$, 556 nm; $^4T_{1g} \leftarrow {}^4A_{2g}$, 407 nm.

a. What is the energy gap of each of these transitions?

b. What are the populations of the upper states relative to the ground state at 300 K?

12.8 The laser light from a ruby laser is at 694.3 nm.

a. What is the energy of the ^2E_g lasing state above the ground state, $^4A_{2g}$?

b. Estimate the fraction of Cr^{3+} ions in this upper state due to thermal equilibrium alone at 300 K.

12.9 What is the separation of the energy levels in neodymium ions that give rise to laser lines at:

a. 0.914 μm?

b. 1.06 μm?

12.10 Gallium arsenide has a band gap of 1.35 eV, and aluminium arsenide has a band gap of 2.16 eV.

a. What is the wavelength and colour of photons emitted by these solids?

b. In order to make an orange LED with an emission at a wavelength of 600 nm, it is proposed to make a solid solution $Ga_xAl_{1-x}As$.

c. Taking the variation in band gap with composition as linear, what value of x is required?

12.11 Gallium nitride has a band gap of 3.34 eV, and indium nitride has a band gap of 2.0 eV.

a. What is the wavelength and colour of photons emitted by these solids?

b. In order to make a green LED with an emission at a wavelength of 525 nm, it is proposed to make a solid solution $Ga_xIn_{1-x}N$.

c. Taking the variation in band gap with composition as linear, what value of x is required?

12.12 A solid solution of indium phosphide and aluminium phosphide, $In_xAl_{1-x}P$ is made up.

a. At what value of x will the light emitted by an LED made from this compound just be visible?

b. What colour will it be?

c. The band gap of InP is 1.27 eV and that of AlP is 2.45 eV.

12.13 A solution is quoted as having a 22% transmittance. What is the absorbance?

12.14 The linear absorption coefficient of zinc metal for X-rays from a nickel target is 5.187 m^{-1}, and for cadmium metal is 18.418 m^{-1}.

a. What thickness of plates of these metals in needed to reduce the irradiance of the radiation passing through the plate to 0.1 of the incident irradiance.

b. What will be the transmittance and absorbance of the plates?

12.15 A plate of a cadmium–zinc alloy 21.7 cm thick is used to reduce the X-radiation from a nickel target to 0.05 of its incident value.

a. What is the transmittance and absorbance of the plate?

b. Assuming that the absorption coefficients, given in the previous question, can be added, what is the composition of the alloy, in atom %?

12.16 A lead glass fibre has a refractive index of 1.682. What is the critical angle for total internal reflection at the interface with an acrylic coating with refractive index 1.498?

12.17 A ray of light passing through water (refractive index, n = 1.33), in a glass tank (refractive index 1.58), hits the water/glass surface at an angle of 23°.

a. What angle does it make with the surface as it continues through the glass?

b. What is the critical angle for light passing through the water striking the water/glass interface?

c. What is the critical angle for light passing through the glass striking the glass/water interface?

12.18 Using the Gladstone–Dale formula, estimate the refractive indices of the ceramics:

a. Barium titanate, $BaTiO_3$.

b. Lead titanate, $PbTiO_3$.

c. Densities: $BaTiO_3$, 6017 kg m^{-3}; $PbTiO_3$, 8230 kg m^{-3}.

12.19 19 Using the Gladstone–Dale formula, estimate the refractive index of the minerals:

a. Spinel, $MgAl_2O_4$.

b. Akermanite, $Ca_2MgSi_2O_7$.

c. Densities: spinel, 3600 kg m^{-3}; akermanite, 2940 kg m^{-3}.

12.20 20 Using the Gladstone–Dale formula, estimate the refractive index of the minerals:

a. beryl, $Be_3Al_2(SiO_3)_6$.

b. garnet, $Mg_3Al_2Si_3O_{12}$.

c. Densities: beryl, 2640 kg m^{-3}; garnet, 3560 kg m^{-3}.

12.21 Estimate the refractive coefficient of Al_2O_3 using the information that the mineral andalusite, Al_2SiO_5, has a density of 3150 kg m^{-3} and a refractive index of 1.639. The refractive coefficient of SiO_2 is 0.21.

12.22 Calculate the reflectivity of the surfaces of the following transparent materials in air;

a. Na_3AlF_6, n = 1.35;

b. glass, n = 1.537;

c. Al_2O_3, n(average) = 1.63;

d. ZrO_2, n = 2.05;

e. Ta_2O_5, n(average) = 2.15.

12.23 Calculate the reflectivity of the surfaces of the following metals in air. The optical constants are for a wavelength of 550 nm:

a. aluminium, $n = 0.82$, $k = 5.99$;

b. silver, $n = 0.255$, $k = 3.32$;

c. gold, $n = 0.33$, $k = 2.32$;

d. chromium, $n = 2.51$, $k = 2.66$;

e. nickel, $n = 1.85$, $k = 3.27$.

12.24 A thin film on a substrate viewed in air has an optical thickness of $\lambda/4$. Will the film be reflecting or not (a), if the substrate has a lower refractive index than the film, and (b), if the substrate has a higher refractive index than the film.

12.25 25 A thin film on a substrate viewed in air has an optical thickness of $\lambda/2$. Will the film be reflecting or not

a. if the substrate has a lower refractive index than the film, and

b. if the substrate has a higher refractive index than the film.

12.26 Describe the differences between the reflectivity of a soap film, thickness $\lambda/4$, in air, compared to the same film on an oil surface. Assume that the refractive index of the soap film is equal to that of water, 1.33, and that of the oil is 1.44.

12.27

a. What is the minimum physical thickness of a film of titanium dioxide in air that will give rise to constructive interference of green light, $\lambda = 550$ nm? The refractive index of TiO_2 is 2.875 (average). Estimate the colour that the film would appear when viewed in white light:

b. by reflection;

c. by transmission.

12.28 Derive the relationship

$$n_f = \sqrt{n_s}$$

for an antireflection coating, refractive index n_f, on a substrate, refractive index n_s, in air.

12.29 Determine the reflectivity of a $\lambda/4$ film of:

a. silicon oxide, $n = 2.0$;

b. titanium dioxide, $n = 2.775$; on glass, $n = 1.504$, in air.

12.30

a. Determine the reflectivity of a $\lambda/4$ and a $\lambda/2$ film of magnesium fluoride, $n = 1.384$, on glass, $n = 1.504$, in air.

b. What changes would occur if the substrate was titanium dioxide, $n = 2.775$ (average)?

12.31

a. Plot a graph of reflectivity vs. the number of pairs of layers for a quarter-wave stack on a glass substrate, $n = 1.495$, in air, using alternating layers of magnesium fluoride, $n = 1.384$, and titanium dioxide, $n = 2.775$ (average).

b. How many pairs are needed to achieve a reflectivity of at least 99.9%?

12.32 A quarter wave stack on a glass substrate, $n = 1.545$, in air, is required to reflect light of 650 nm from a laser. The materials chosen have refractive indices of 2.15 (Ta_2O_5) and 1.35 (Na_3AlF_6). What is the physical thickness of each layer?

12.33 The irradiance of a light beam traversing a solution placed into a cell of 10 cm path length drops to 80.3% of the incident irradiance. Calculate the linear absorption coefficient of the

solution.

12.34 The visibility of the atmosphere is reported as being 10 km. Assuming that at this distance an object can just be perceived, with 1% of the initial light falling on the object reaching the observer, determine the linear absorption coefficient of the atmosphere.

12.35 Determine the relative amount of light scattered by dust particles with a refractive index of 1.45 relative to that of water droplets with a refractive index 1.33, if Rayleigh scattering occurs.

12.36 Will air visibility be improved if water droplets responsible for scattering are replaced by similar sized limestone dust particles? The refractive index of limestone is 1.53.

12.37 The linear scattering coefficient of limestone dust in the air is 0.000 219 4 m^{-1}. Over what distance will the irradiance of a light beam diminish to 10% of its initial value?

12.38 What is the relative amount of light scattered by Rayleigh scattering:

a. in the incident direction?

b. at 45° to the incident direction?

c. perpendicular to the incident direction?

d. in the reverse direction?

12.39

a. What slit width is needed in a transmission diffraction grating to cause red light, λ = 700 nm, to be deviated by 7.5° to the normal?

b. What will the deviation of the violet light, λ = 400 nm?

12.40 The first photonic crystal made was formed by drilling a block of material with a refractive index of 3.6 so as to form a face-centred cubic array of holes with a unit cell parameter of 2 mm. What wavelength radiation will not pass in the [100] direction?

12.41 A photonic crystal is made by laying down close packed layers of polystyrene spheres of refractive index 1.595, to form a hexagonal closest packed array. The sphere diameter is 250 nm, and the layer separation normal to the layers can be taken as 0.8 of the sphere diameter.

a. What is the wavelength that will not be transmitted in light normal to the close-packed layers?

b. What colour does this correspond to?

c. If the ordering is imperfect and composed of random hexagonal (AB) and cubic (ABC) packing close-packed layers how will the result change?

12.42 Estimate the attenuation of an optical fibre in which 0.5% of the initial power is lost in a distance of 1 km.

12.43 The attenuation of ordinary window glass is of the order of 10^6 dB km^{-1}. What thickness of glass would cause the incident light intensity to fall to 50%?

12.44 An instantaneous pulse of white light is introduced into a silica optical fibre. What will the pulse spread be, in km, after 1 second? The refractive indices of silica are n(400 nm) 1.47000; n(average), 1.462 65; and n(700 nm) 1.455 30.

12.45 A thin film of silver oxide forms on a silver surface. The band gap for silver oxide is 2.25 eV.

a. What colour would be emitted by bulk silver oxide?

b. Treating the thin film as a quantum well, what film thickness is needed to obtain an emission at a wavelength of 413 nm? The effective mass of both electrons and holes in Ag_2O is 0.3 m_e.

12.46 A film of zinc sulfide 4 nm in thickness forms on metallic zinc. Treating the film as a quantum well, what is the wavelength of the transition Δn = 2. The band gap of ZnS is 3.54 eV and the effective mass of both electrons and holes is 0.4 m_e.

Notes

1 The momentum of an electron is equal to $kh/2\pi$, and the **k** axis is often referred to as the momentum axis.

2 Note that in spectroscopic convention, the higher energy state is written first and the lower energy state second. Absorption of radiation is then described with a right–left pointing arrow and emission of radiation by a left–right pointing arrow.

13
Thermal Properties of Solids

- Why are metals poor thermal insulators?
- What materials show thermal contraction not thermal expansion?
- How does a thermocouple work?

In a normal solid the component atoms are in constant motion and the vibrations constitute the thermal energy of the material. The way in which materials respond to changes in thermal energy forms the subject of thermodynamics. Although an account of the formal structure of thermodynamics is beyond the scope of this chapter, a number of more general physical properties allied to changes in thermal energy are described.

13.1 Heat Capacity

13.1.1 The Heat Capacity of a Solid

The *heat capacity* of a solid quantifies the relationship between the temperature of a body and the energy supplied to it. If a large amount of heat supplied to a body produces only a small temperature increase, the solid has a large heat capacity, formally defined by:

$$C = \Delta Q / \Delta T$$

where ΔQ is the amount of heat (energy) needed to produce a temperature change ΔT so that C, the heat capacity, represents the amount of heat required to raise the temperature of the sample by 1 °K. The heat capacity is frequently quoted as the *molar heat capacity*; the amount of heat that increases the temperature of 1 mol of substance by 1°, units J mol^{-1} K^{-1}, or the *specific heat capacity*, the amount of heat that increases the temperature of 1 g of a substance by 1°, units J g^{-1} K^{-1}.

The heat capacity is found to depend upon whether the measurement is made at constant volume, C_v, or at constant pressure, C_p. For solids, the difference between C_p and C_v is very small at room temperature and below. Most experimental studies on materials are made at atmospheric pressure and C_p values are of most relevance (Table 13.1). The heat capacity of a solid generally increases with temperature. Empirically C_p can be fitted to an equation of the form:

$$C_p = a + bT + cT^{-2} + dT^2 + eT^{-1/2}$$

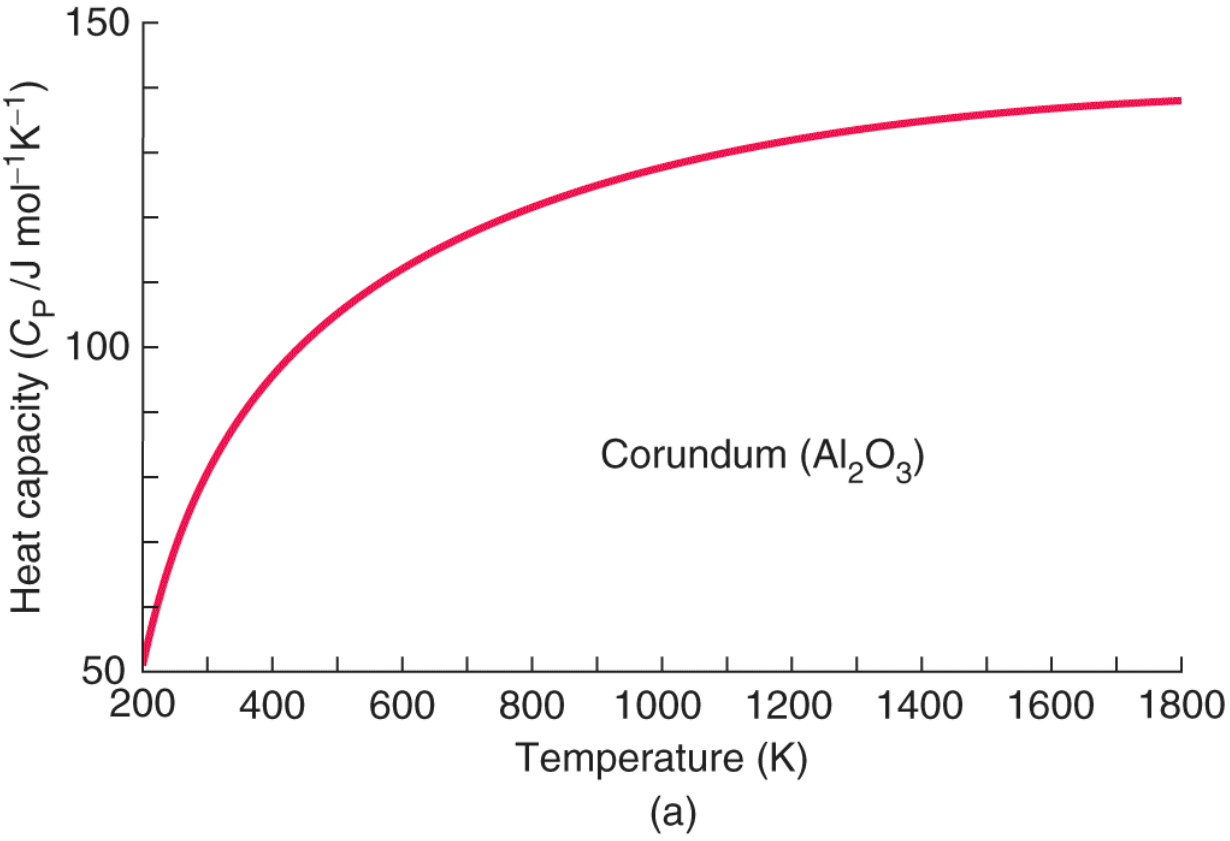
where a, b, c, d, and e are constants. For example, the heat capacity of Al_2O_3 can be approximated by:

$$C_p = 106.6 + 0.0178\,T - 2.85 \times 10^6 / T^2$$

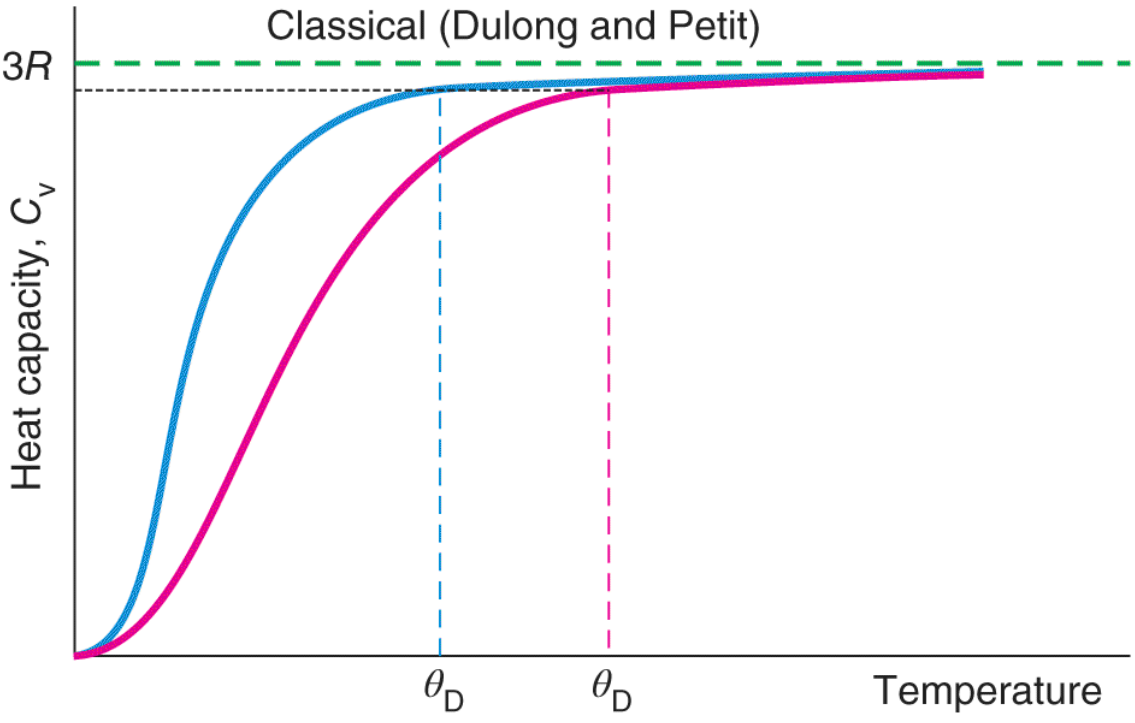
in the temperature range from 300 to 1800 K (Figure 13.1a).

Table 13.1 The Debye temperature and room temperature heat capacity of some elements.

Element	**Debye temperature θ_D (K)**	**Heat capacity C_p (J mol^{-1} K^{-1})**
Li	345	24.9
Na	158	28.2
K	91	29.6
Rb	56	31.1
Cs	38	32.2
C (diamond)	2340	13.0
Si	645	19.8
Ge	375	23.2
Sn (white = β)	230	27.1
Pb	105	26.7


150
100
50
Heat capacity (C_P /J mol^{-1}K^{-1})
Corundum (Al_2O_3)
200
400
600
800
1000
1200
1400
1600
1800
Temperature (K)

(a)


Classical (Dulong and Petit)
3R
Heat capacity, C_v
θ_D
θ_D
Temperature

(b)

Figure 13.1 The variation of heat capacity of a solid with temperature: (a) experimental C_p data for corundum, Al_2O_3; (b) C_v for two elements with differing Debye temperatures (schematic).

13.1.2 Theories of Heat Capacity

The heat capacity of most metals was found to be approximately 25 J mol^{-1} K^{-1} in 1819 by Dulong and Petit. This was explained by Boltzmann, who analysed the statistics of vibrating particles assuming that there is no constraint upon either the allowed energy values or the number of particles that can have any energy value, that is, each atom vibrated quite independently of the others. The resulting particle statistics is called *Classical*, *Boltzmann*, or *Maxwell–Boltzmann* statistics. Central to the analysis is that each atom could possess six degrees of freedom associated with different trajectories of motion in the three dimensions of space. Each of these degrees of freedom had a total energy of ½ $k_B T$ and as there are N_A atoms per mole, the molar heat capacity is:

$$C_v(\text{classical}) = 6N_A \, ½ \, k_B T$$
$$= 3R \approx 25 \text{ J mol}^{-1} \text{ K}^{-1}$$

where k_B is Boltzmann's constant and R ($=N_A k_B$) is the gas constant. In this model the heat capacity was independent of temperature.

The classical heat capacity value is reasonable for high temperatures, but completely incorrect at low temperatures, and the heat capacity of a solid is not at all independent of temperature. The discrepancy was resolved by realising that the energy of the vibrating atoms, E, was quantised in the following way:

$$E = (n + ½) h\nu$$

where n is an integer quantum number, h is Planck's constant, and ν is the vibration frequency. The quantum of vibrational energy is called a *phonon* and at normal temperatures a phonon energy distribution, the *phonon spectrum*, exists. Any number of phonons is allowed to occupy each energy level so as the temperature falls, more and more occupy the lower energy levels. At the lowest temperatures

all phonons can be in the lowest energy state. However, the constraint imposed by quantization shows that the lowest energy possible is not zero, as it is for a classical system, but is equal to ½ $h\nu$ when n is zero. This energy is called *zero-point energy*. Any solid will possess this amount of energy even at 0 K.

To calculate the total energy of a system, it is necessary to allocate the phonons to the available energy levels, using quantum rather than classical statistics. The appropriate tool for this is *Bose–Einstein statistics*, which also applies to photons and gases at very low temperatures. The particles that obey these statistics are called *bosons* to distinguish them from electrons and other *fermions* that obey Fermi–Dirac statistics (Section 1.4.2). The resulting energy calculation, by Einstein, which assumed that each atom vibrated independently of the others, was a good overall fit to the data, but was not in perfect accord with the experimental values at low temperatures. Debye amended the calculation by including the fact that the phonons throughout a crystal are coupled together, via chemical bonding. The phonons are then likened to waves throughout the whole of the solid body, which must fit into the dimensions of the solid with the wavelength, frequency, and energy quantised. In Debye's model the heat capacity is:

$$C_v = 9R\left(\frac{T}{\theta_D}\right)^3 \int_0^{\theta_D/T} \frac{x^4 e^x}{(e^x - 1)^2} dx$$

$$x = h\nu/k_B T; \quad \theta_D = h\nu_D/k_B$$

where θ_D is the *Debye temperature*, ν_D is the *Debye frequency*, R is the gas constant, and T the temperature (Figure 13.1b)

For temperatures well *above* the Debye temperature, the value of C_v is found to be $3R$, and so all solids would be expected to have a high temperature heat capacity of about 25 J mol^{-1} K^{-1}, equal to the classical value. (This works quite well for many metallic elements, but does not hold for most compounds.) The Debye temperature controls the rate at which the C_v curve approaches the value of $3R$, and broadly speaking the value of C_v at the Debye temperature is

approximately 23.9 J mol^{-1} K^{-1} (Figure 13.1b). Note that the Debye temperature drops systematically on moving down a periodic table group (Table 13.1).

At temperatures well *below* the Debye temperature, the value of C_V is given by:

$$C_v = \frac{12\pi^4}{5} R \left(\frac{T}{\theta_D} \right)^3$$

That is,

$$C_v = A\, T^3$$

where A is a constant.

13.1.3 Heat Capacity at Phase Transitions

Although the phonon contribution to the heat capacity is the most important, others also occur. Significant heat capacity changes accompany phase changes, order–disorder changes, when a ferromagnetic solid becomes paramagnetic and so on. In all of these cases the curve of heat capacity vs. temperature will show a break at the transformation temperature.

First-order transitions are characterised by the absorption or evolution of the latent heat of the transformation. This means that, at the transition itself, heat energy supplied to the material is used to propagate the transformation and not increase the temperature. The heat capacity at this point is nominally infinite as ΔQ is finite and ΔT is zero. The heat capacities of the phases either side of the transition are generally different and so the heat capacity vs. temperature curve will show a discontinuity (Figure 13.2a).

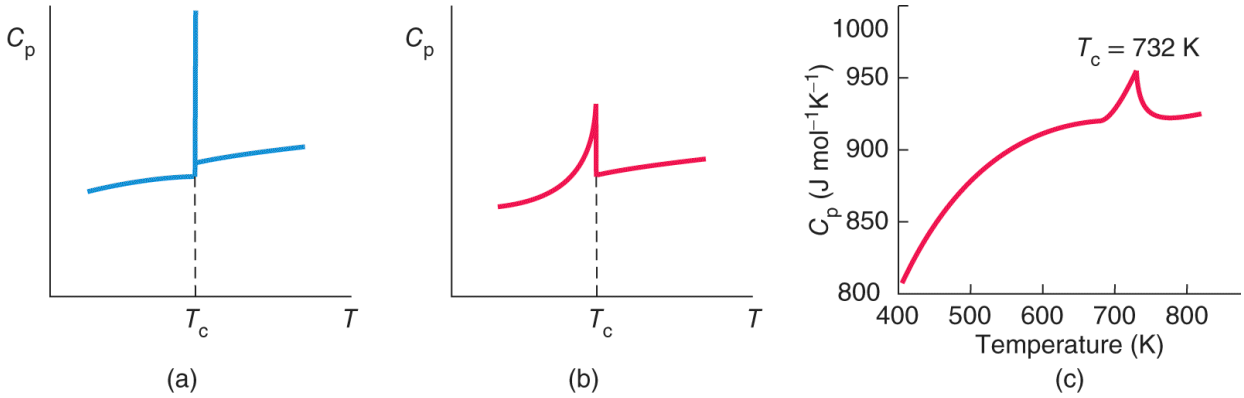


Figure 13.2 Variation of the heat capacity of a solid at a phase transition: (a) first-order; (b) second-order; (c) the second-order ferrimagnetic to paramagnetic transition in $SrFe_{12}O_{19}$.

Source: Modified from Rakshit et al. (2007).

A second-order transformation is similar, but in these cases no latent heat of transformation appears. The heat capacity curve will show a discontinuity, but it will not be infinite (Figure 13.2b). These breaks are often used to characterise second-order transitions such as those involved in ferroelectric and ferromagnetic transformations (Figure 13.2c).

13.2 Thermal Conductivity

13.2.1 Heat Transfer

When a solid is heated or cooled, heat is transferred through the structure, leading to the concept of *thermal conductivity* as a parallel to electrical conductivity – these ideas arising when both heat and electricity were supposed to be material fluids. Thermal conductivity is defined as the amount of heat passing through a material over a unit length per unit area for a unit temperature gradient, with units $W\,m^{-1}\,K^{-1}$. The thermal conductivity of a solid is an important parameter: ultrahigh thermal conductivity is of importance in stabilising thermal management in heat producing systems, while very low thermal conductivity materials find use as thermal barriers.

The equations describing heat transfer, initially formulated by Fourier, predate and are of identical mathematical form to Fick's laws of diffusion (Sections). In the case of steady-state heat transfer, the one-dimensional heat transfer equation is:

$$J_Q = -K\,\mathrm{d}Q_c/\mathrm{d}x \tag{13.1}$$

where J_Q is the heat flux, $\mathrm{d}Q_c/\mathrm{d}x$ is the heat concentration gradient along the direction of heat flow, x, and K is the *thermal diffusivity*, defined by:

$$K = \kappa/\rho c_p$$

where κ is the thermal conductivity, ρ is the density, and c_p is the specific heat capacity at constant pressure. All of these thermal quantities are analogous to their diffusion counterparts. The solutions of Eq. (13.1) are identical in form to those given in Section 5.3 for steady-state diffusion and can be used to determine the heat flow under steady-state conditions.

In the case of non-steady-state heat transfer, the equation analogous to the diffusion equation is the *heat equation*, also known as the *Biot–Fourier* equation:

$$dT/dt = K\, d^2T/dx^2$$

where dT/dt is the change of temperature with time at a point x in the solid. The solutions to this equation are identical in form to those given for the non-steady-state diffusion equation (Section 5.2).

Thermal conductivity is generally determined under steady-state conditions using Eq. (13.1). When the two ends of a solid are held at different temperatures, heat flows across the body from the hot to the cold side and a uniform temperature gradient forms, which is a measure of the thermal conductivity of the material (Figure 13.3a). (This figure is the analogue of Figure 5.6.) The amount of heat transferred per unit time, Q_t, depends upon the cross-sectional area over which the heat is conducted, A, the temperature difference between the hot and cold regions ($T_H - T_C$), and the separation ($x_H - x_C$):

$$Q_t = \kappa A\,(T_H - T_C)/(x_H - x_C)$$

or

$$Q_t = \kappa A\, dT/dx$$

where κ is the thermal conductivity and dT/dx is the temperature gradient. When the heat transfer is across a number of materials, the contribution to each slice is summed:

$$Q_t = A \sum(\kappa_i\, \Delta T_i/\Delta x_i)$$

where ΔT_i is the temperature drop across a slice of thickness Δx_i and thermal conductivity κ_i. The temperature drop across a series of different materials will vary in steps, depending on the thermal conductivity of each slice (Figure 13.3b).

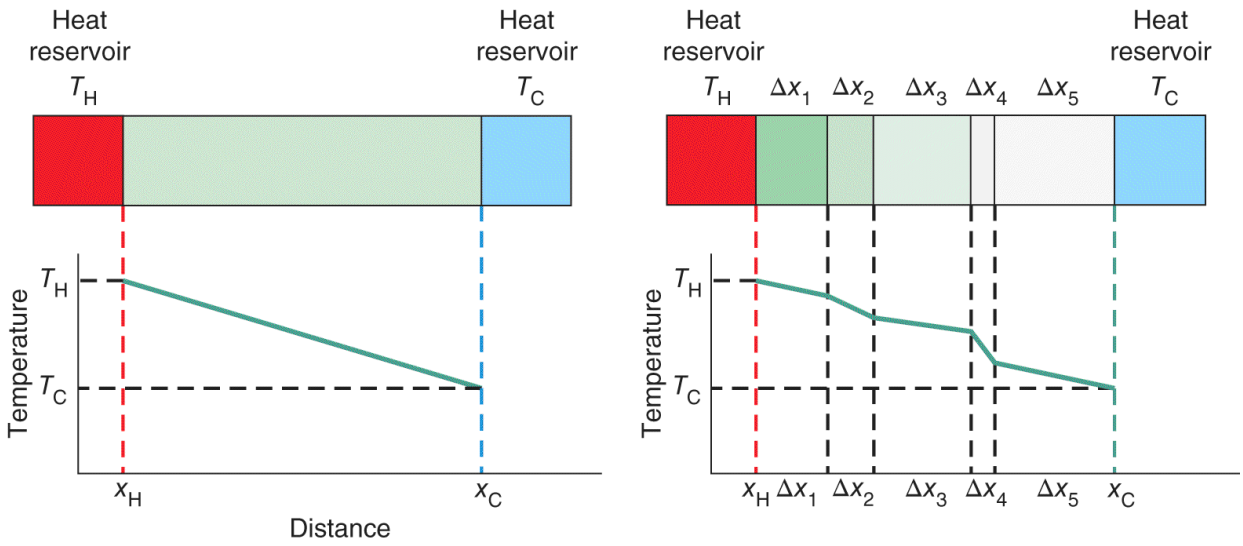


Figure 13.3 Thermal conductivity: (a) the temperature gradient across a homogeneous solid; (b) The temperature gradient across a series of different solids.

The thermal conductivity of solids varies considerably (Table 13.2, Figure 13.4). Metals have a high thermal conductivity, with silver having the highest room temperature thermal conductivity, 430 W m^{-1} K^{-1}. Alloys have lower thermal conductivities than pure metals. Ceramics are even lower, especially porous porcelains or fired clay products. The lowest thermal conductivities are shown by plastic foams such as foamed polystyrene. As would be expected, the thermal conductivity of crystals varies with direction. For example, the thermal conductivity of the metal cadmium (hexagonal structure) is 83 W m^{-1} K^{-1} parallel to the **c**-axis and 104 W m^{-1} K^{-1} parallel to the **a**-axis. At 25 °C, hexagonal quartz has a thermal conductivity parallel to the **c**-axis of 11 and 6.5 W m^{-1} K^{-1} parallel to the **a**-axis.

Table 13.2 Coefficients of thermal conductivity and thermal expansion.

Material	**Thermal conductivity (W m^{-1} K^{-1})**	**Thermal expansion (10^{-6} K^{-1})**
Metals		
Silver	428	18.9
Copper	403	16.5
Gold	319	14.2
Iron	83.5	11.8
Nickel	94	13.4
Titanium	22	8.6 (mean)
Alloys		
Brass	106	17.5
Bronze	53	17.3
Carbon steel	~50	~10.7
Monel (67 wt% Ni, 29 wt% Cu, 4 wt% Fe)	21	~14
Lead-tin solder	~50	~24
Refractories		
Alumina	38	5.5
Magnesia	40	9.5
Silica	1.6	0.49
Porcelain	~2	~4.5
Polymers		
Nylon 6,6	0.25	80
Polyethylene	~0.4	~200
Polystyrene foam	~0.04	—
High thermal conductivity materials		
Diamond	2000	1

Material	**Thermal conductivity (W m^{-1} K^{-1})**	**Thermal expansion (10^{-6} K^{-1})**
Graphite ⊥ **c**-axis	2000	0.6
cubic BN	1600	—
SiC	490	3.3
BeO	370	—
BP	360	—
AlN	320	—
BeS	300	—
BAs	210	—
Si_3N_4	200	2.5
GaN	170	—
Si	160	2.6
AlP	130	—
GaP	100	4.7

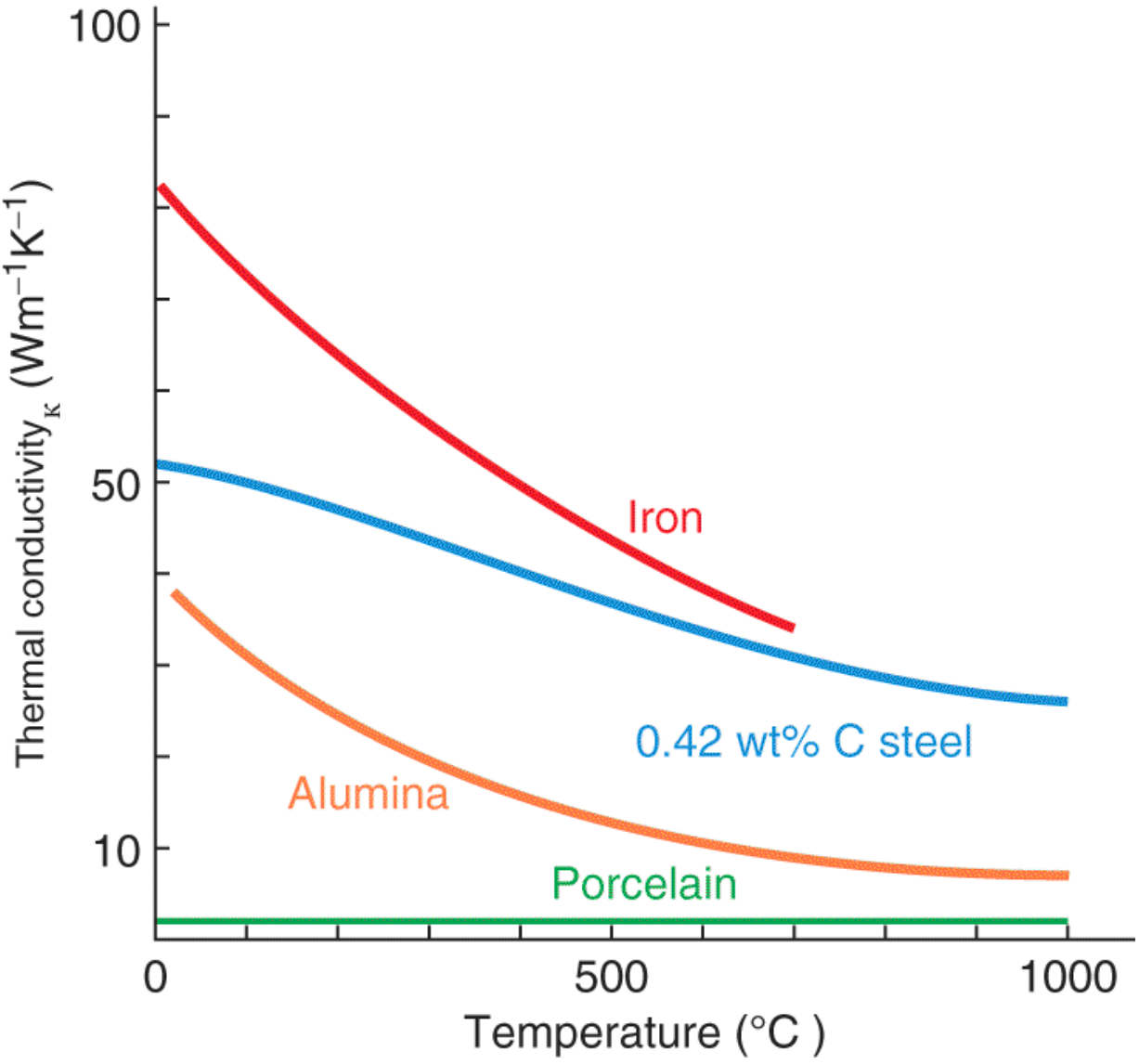


Figure 13.4 The variation of thermal conductivity with temperature of some common materials.

A number of non-metallic materials are called *high thermal conductivity materials*. The most notable of these is diamond, with a thermal conductivity of 2000 W m^{-1} K^{-1}. All of the others have a diamond-like structure, and include cubic boron nitride and aluminium nitride (Table 13.2).

13.2.2 Thermal Conductivity and Microstructure

Thermal conductivity in metals and semiconductors is attributed mainly to the movement of mobile electrons present in the solid and

lattice vibrations. Just as for light waves/photons, it is sometimes more convenient to consider lattice vibrations as waves and at other times as quantized 'particles' termed phonons. This is the case when treating thermal conductivity. Hence, we can write:

$$\kappa = \kappa\,(\text{electrons}) + \kappa\,(\text{phonons})$$

In metals mobile electrons make the greatest contribution, and the phonons can be set to one side. As electrons are responsible for both electronic and thermal conductivity, it would be expected that there would be a correlation between high electronic conductivity and thermal conductivity, and indeed this is so, with, for example, Cu and Ag having very high electronic and thermal conductivities. At the simplest level, the electrons in the crystal can be imagined as a free electron gas (Section 1.4.2), moving with a velocity that is higher at the hot end of the solid than the cold end. The kinetic energy is gradually transferred to the cold end by collisions between the electrons themselves and with the atoms in the structure. The thermal conductivity increases as the number of free electrons increases. The model is successful in many ways. For example, it predicts that thermal conductivity will be proportional to electrical conductivity, σ and temperature T:

$$\kappa = L_0\,\sigma\,T$$

where L_0 is the *Lorentz coefficient* (2.5×10^{-8} W S^{-1} K^{-2}). However, it does not explain the differences between one metal and another, and for this, knowledge of the dynamics of the electrons at the Fermi surface of an individual metal is needed. A similar approximation can be used with metallic alloys, but as this depends upon the concentration and chemical nature of the alloying elements, the equation is usually written:

$$\kappa = L_0\,\sigma\,T + C \qquad (13.2)$$

where L_0 and C are considered to be empirical constants. For example, the thermal conductivity of many aluminium alloys (with Al as the principal component) is:

$$\kappa = 2.22 \times 10^8 \, \sigma \, T + 10.5$$

In addition, the thermal conductivity of metals and alloys will depend upon the defects present, which are the result of the chemical, thermal, and mechanical history of the alloying process. In this case Eq. (13.2) can also be used to approximate the behaviour.

Thermal conductivity in crystalline semiconductors and insulators, including ceramics and polymers, is mainly attributed to the phonon component. The solid is imagined to contain a phonon gas similar to the electron gas in a metal. At the hot end of a solid the kinetic energy of the phonons is greater than at the cold end. This energy is gradually transferred from hot to cold by phonon–phonon interactions and by interactions between the phonons and the solid structure. The thermal conductivity depends upon the mean free path of the phonons, which is the distance that phonons traverse in the structure before colliding. A short mean free path correlates with a low thermal conductivity.

As with metals, defects in a semiconducting or insulating crystal structure shorten the mean free path and reduce thermal conductivity significantly. Point defects can drastically lower the thermal conductivity of a pure material and is an important consideration when attempts are made to synthesise high thermal conductivity ceramics. For example, SiO_2 is readily formed on silicon nitride, Si_3N_4, by oxidation in air. This can react with the bulk nitride to introduce substitutional defects and vacancies in the following way:

$$2SiO_2(s) \rightarrow 2Si_{Si} + 4O_N^{\bullet} + V_{Si}^{4\prime}$$

where Si_{Si} represents silicon atoms on normal silicon sites, $O_N^{\bullet}$ represents oxygen atoms on nitrogen atom sites, and $V_{Si}^{4\prime}$ represents a vacancy on a silicon site (see Section 3.6.5). Thus, two SiO_2 units produce five defects in the silicon nitride, all of which degrade the thermal conductivity. Purification therefore forms an important step in the manufacture of high thermal conductivity ceramic materials.

In many materials thermal conductivity is hampered by the presence interfaces and successful heat transfer across interfaces is often crucial to the performance of electronic devices, especially as the size

of these reduces towards the nanoscale. By analogy with thermal transfer through the bulk (Eq. (13.1)), it is possible to define an interfacial thermal conductivity γ by:

$$Q_t = \gamma A \, \mathrm{d}T/\mathrm{d}x$$

At present, calculation of γ for interfaces and a full understanding of heat transfer across such boundaries is incomplete. Empirical studies show that surface roughness lowers γ appreciably and both molecular dynamics simulations and experimental studies indicate that suitable chemical bonding in the interfacial region can improve γ.

The presence of defects is a particular problem with polycrystalline ceramic materials that are produced by sintering. This process naturally leads to the formation of grain boundaries, pores, and voids. Because of this, the thermal conductivity of sintered bodies is usually much lower than the intrinsic thermal conductivity. Fired clay ceramics are among the poorest ceramic thermal conductors as they have very high porosity. At present, the best sintered ceramic solids have about 75% of the intrinsic thermal conductivity of the parent phase.

The thermal conductivity of amorphous solids cannot be understood in terms of the electron/phonon model that is applicable to crystalline materials. The heat transferring phonons are not present as such, but, in the main, are replaced by three other quantized vibrational pseudoparticles: *propagons*, *diffusons*, and *locons*. The propagons resemble the traditional phonon thermal transfer agents, but operate only over a low frequency range. Diffusons also contribute to heat transfer by a diffuse form of energy transport rather than via particle–particle interactions as with phonons. Locons correspond to quantized high-frequency vibrations generally considered to be localised on atom centres and so do not contribute strongly to thermal conductivity. The theory of the thermal behaviour of amorphous materials is still in the process of development.

As this suggests, the thermal conductivity of polymers is found to depend upon the degree of crystallinity of the material as the crystalline portions of the structure have a higher thermal conductivity than the amorphous regions. Materials with high porosity, such as plastic foams, have particularly low thermal

conductivities as the voids totally inhibit phonon transfer. Foamed plastics such as polystyrene are widely used as insulating materials.

13.3 Expansion and Contraction

13.3.1 Thermal Expansion

Most materials increase in volume as the temperature is increased, a feature called *thermal expansion*. Thermal expansion is an important materials property. It is used in most everyday thermometers. The shattering of ordinary glass on being cooled rapidly is due to the rapid contraction of the outer layers compared with the slower contraction rate of the inner regions. This can be avoided if the glass has a low thermal expansion, as do Pyrex glass or fused silica. Differences in thermal expansion of materials in building construction can lead to grave difficulties, and it is the coincidence of the thermal expansion of steel and concrete at normal temperatures that allows the use of steel-reinforced concrete in buildings. The relative thermal expansion of components in precision instruments and electronic devices has always been important. Invar, an alloy of composition $Fe_{0.64}Ni_{0.36}$, has an almost zero coefficient of expansion, and for this reason it has been a vital component of precision instruments since its discovery in 1896.

The thermal expansion of a solid usually increases with temperature. The *mean coefficient of linear thermal expansion*, α_m, of a material is the increase in length per unit length valid for a temperature interval ΔT from an initial temperature T_i to a final temperature T_f:

$$\alpha_m = (l_f - l_i)/(T_f - T_i)\, l_i = \Delta l/\Delta T\, l_i$$

where l_f is the final length, l_i is the initial length, and Δl is the length increment (Figure 13.5).

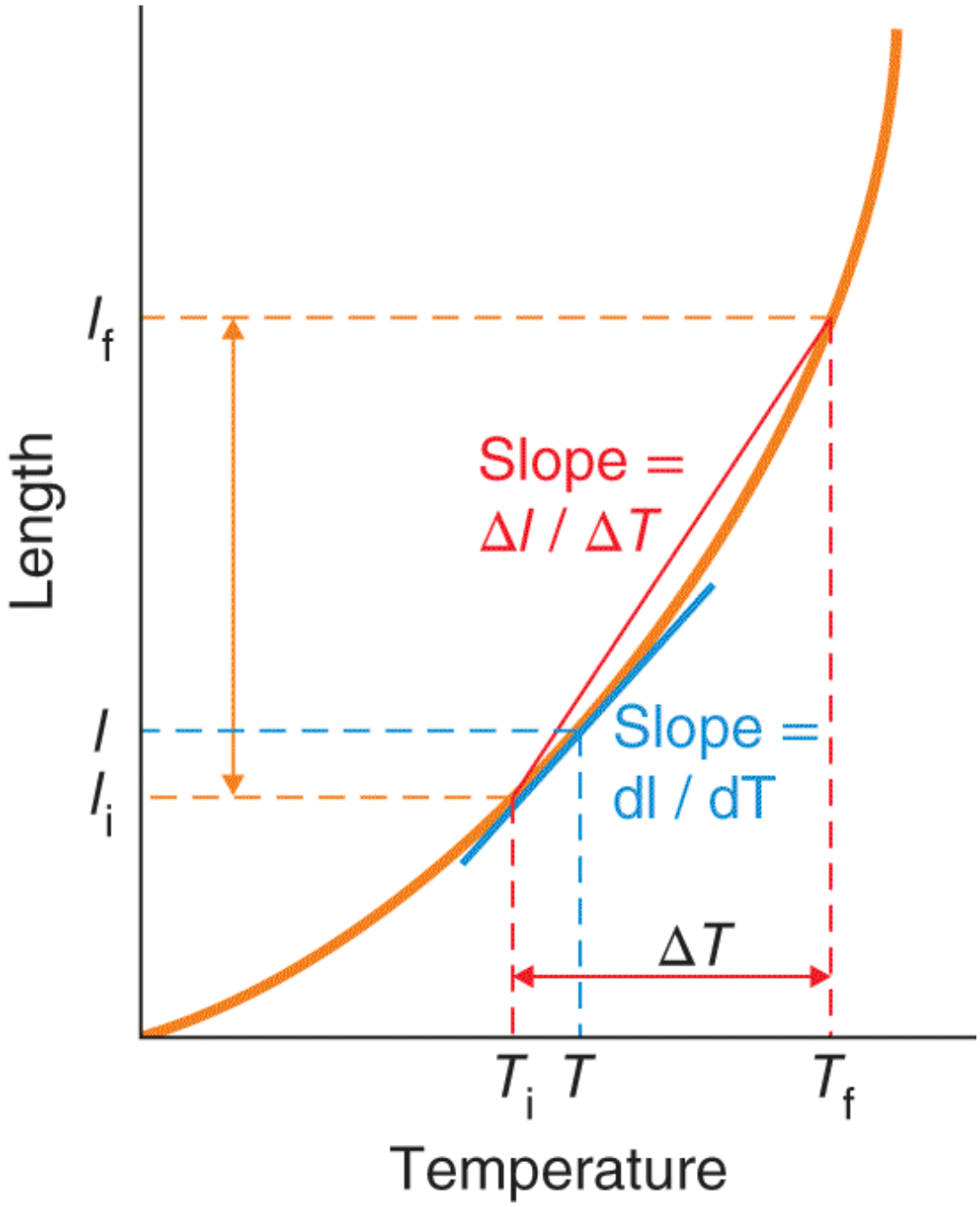


Figure 13.5 Thermal expansion: the mean coefficient of thermal expansion, [$\alpha_m = \Delta l/\Delta T\, l_i$] and expansivity, [$\alpha = dl/dT\, l$].

A similar expression can be written for the *mean coefficient of volume expansion*, also called the *cubical expansion coefficient*, β_m, valid over a temperature interval ΔT:

$$\beta_m = \Delta V/\Delta T\, V_i$$

where ΔV is the volume change and V_i is the original volume at T_i. A reasonable approximation is:

$$3\alpha_m \approx \beta_m$$

In the case of a solid with different mean thermal expansion coefficients along *x*-, *y*-, and *z*-axes,

$$(\alpha_x + \alpha_y + \alpha_z) \approx \beta_m$$

The *linear expansivity of a solid* α is defined as the increase in length per unit length at a given temperature:

$$\alpha = (1/l)\,(\mathrm{d}l/\mathrm{d}T)$$

The expansivity is the *slope* of the length vs. temperature curve at a length l and temperature T (Figure 13.5), and is usually different in value to α_m. The value of α_m tends to the expansivity as Δl and ΔT become small. The expansivity of many solids tends to increase as the temperature increases (Figure 13.6).

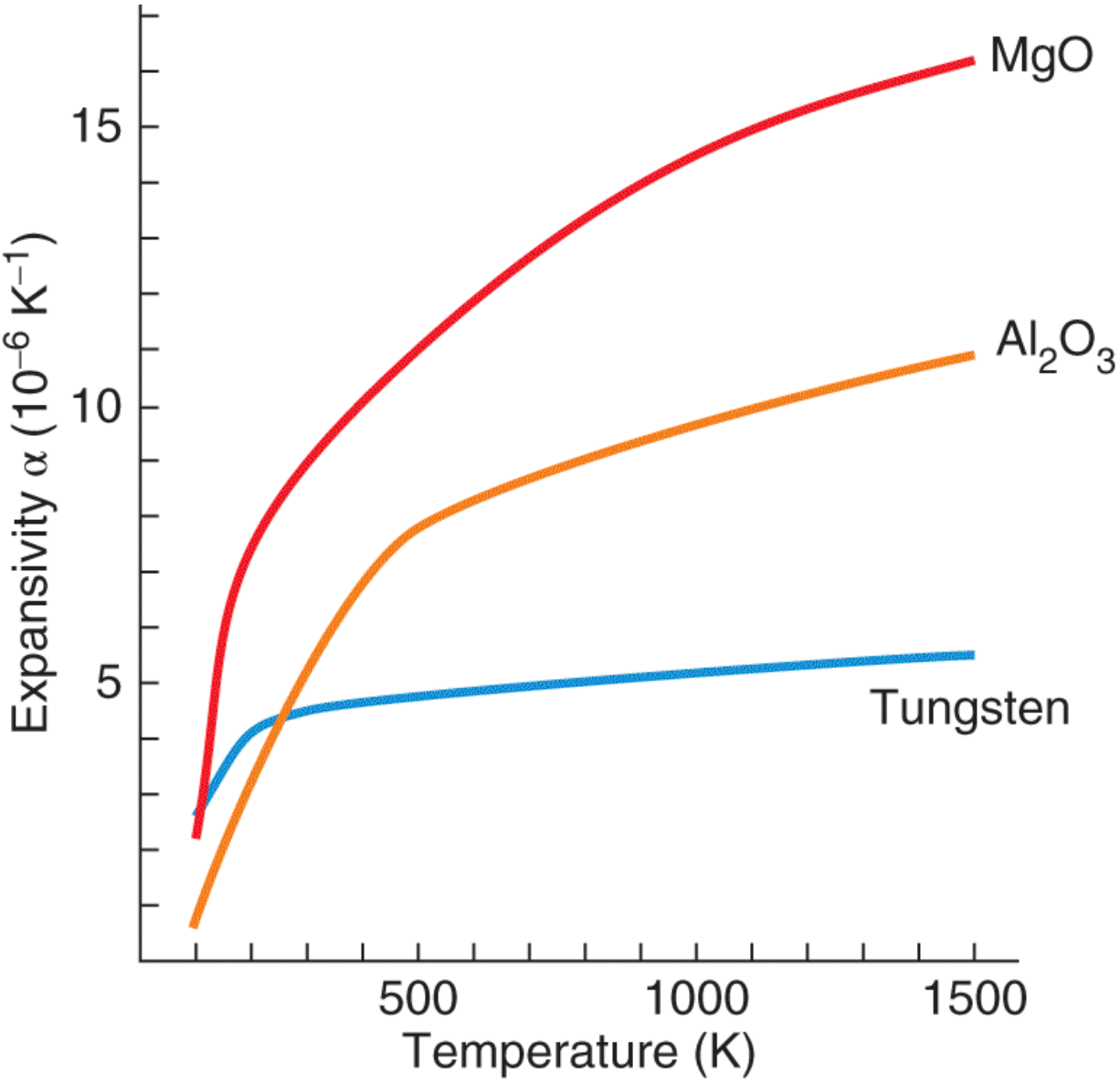


Figure 13.6 The variation of expansivity with temperature for tungsten, magnesium oxide, MgO, and alumina, Al_2O_3.

The thermal expansion of a multiphase solid depends upon the expansivity of the individual components and the ratios present. Thus, the thermal expansion of alloys, glasses and glass ceramics can be tailored by changing the bulk composition of the material. For many applications, a very small coefficient of expansion is desirable, and in cooking ware, for example, glass ceramics with negligible thermal expansion over the temperature ranges normally encountered are widely available.

13.3.2 Thermal Expansion and Interatomic Potentials

An idea of the origin of thermal expansion can be obtained from a consideration of the potential energy of a pair of atoms as a function of their spacing (Figure 13.7a). The extent of the vibrational potential energy V_0, at a low temperature, T_0, leads to a mean separation of the atoms, r_0. As the temperature increases, the vibrational potential energy increases to V_1, V_2, ... and the mean separation increases to r_1, r_2, and so on. Because of the asymmetrical nature of the potential energy curve, this results in a gradual increase in r_1 and r_2. Further temperature increases magnify this off-centre displacement, and the net result is an expansion.

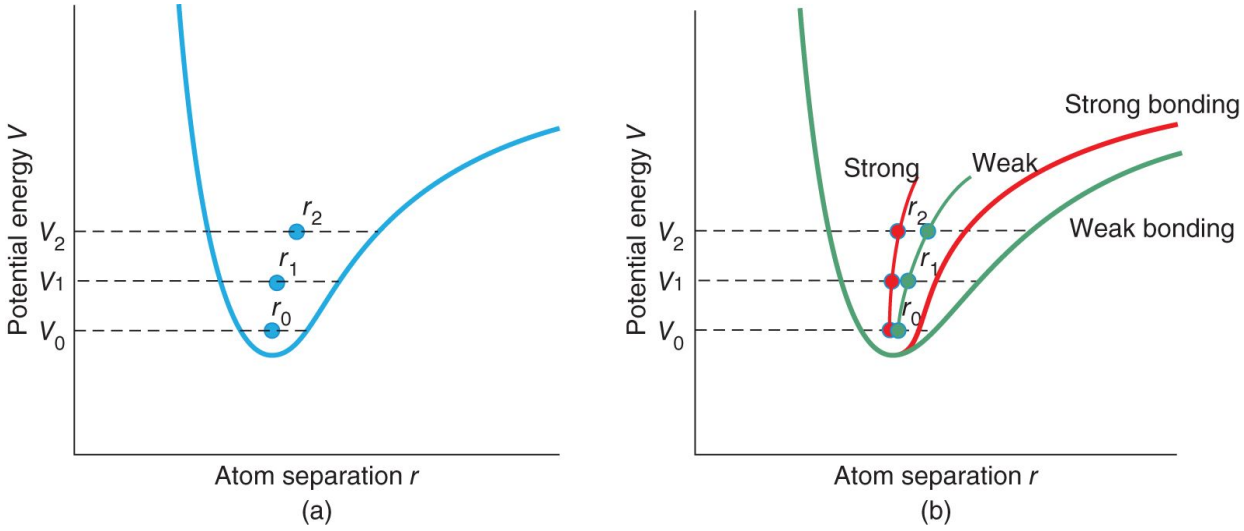


Figure 13.7 Interatomic potentials, schematic: (a) variation of potential with interatomic spacing; (b) curves for a pair of atoms linked by weak or strong bonds.

The shape of the interatomic energy curve is related to the chemical bond strength between the atoms. Strong bonds result in a steep potential energy curve that is reasonably symmetrical close to the minimum. Weak bonding results in a flatter curve that is very unsymmetrical. In this case, the off-centre displacement corresponding to weak bonding will be considerably greater than that for strong bonding (Figure 13.7b). This suggests that strongly bonded solids, such as silica and other ceramics, would have low expansivity,

while polymers, in which the chains are linked by weak chemical bonds, would have high expansivity. This is the case (Table 13.2).

In a real solid, this simple model is insufficient to explain all observations. Obviously, an account has to be taken of all of the atoms in the unit cell, and the interatomic potentials between each pair of atoms have to be evaluated, so as to obtain the mean change of expansion of the unit cell as a whole. Crystals expand less along directions corresponding to strong bonds and more along directions corresponding to weak bonds. The chain silicates, for example, have a higher coefficient of expansion perpendicular to the chains than parallel to them. In general, crystals of lower than cubic symmetry have different expansivities along the different crystal axes. Atomistic simulations can successfully compute thermal expansivity and extend values into temperature and pressure regions that are difficult to study experimentally.

13.3.3 Thermal Contraction

The thermal expansion of a material depends upon the overall balance between all of the interatomic and intermolecular forces present. In some cases, this can result in solids that contract as the temperature increases, giving rise to *negative thermal expansion* (NTE) materials as opposed to normal (positive) thermal expansion (PTE) materials.

The best-known material that behaves in this anomalous fashion is water, between 0 and 3.98 °C (Figure 13.8). This arises from hydrogen bonding between the molecules. At lowest temperatures, the hydrogen bonds pull the water molecules closer together as the thermal vibrations of the fluid decrease. However, the angular structure prevents them packing closely together and they maintain an open structure that is similar to the structure of ice, which is also open and has a lower density than water. Above a temperature of 3.98 ° C, the thermal vibrations begin to dominate, molecular rotation in the liquid increases, and the molecules become effectively spherical. As the temperature then increases, normal expansion is observed.

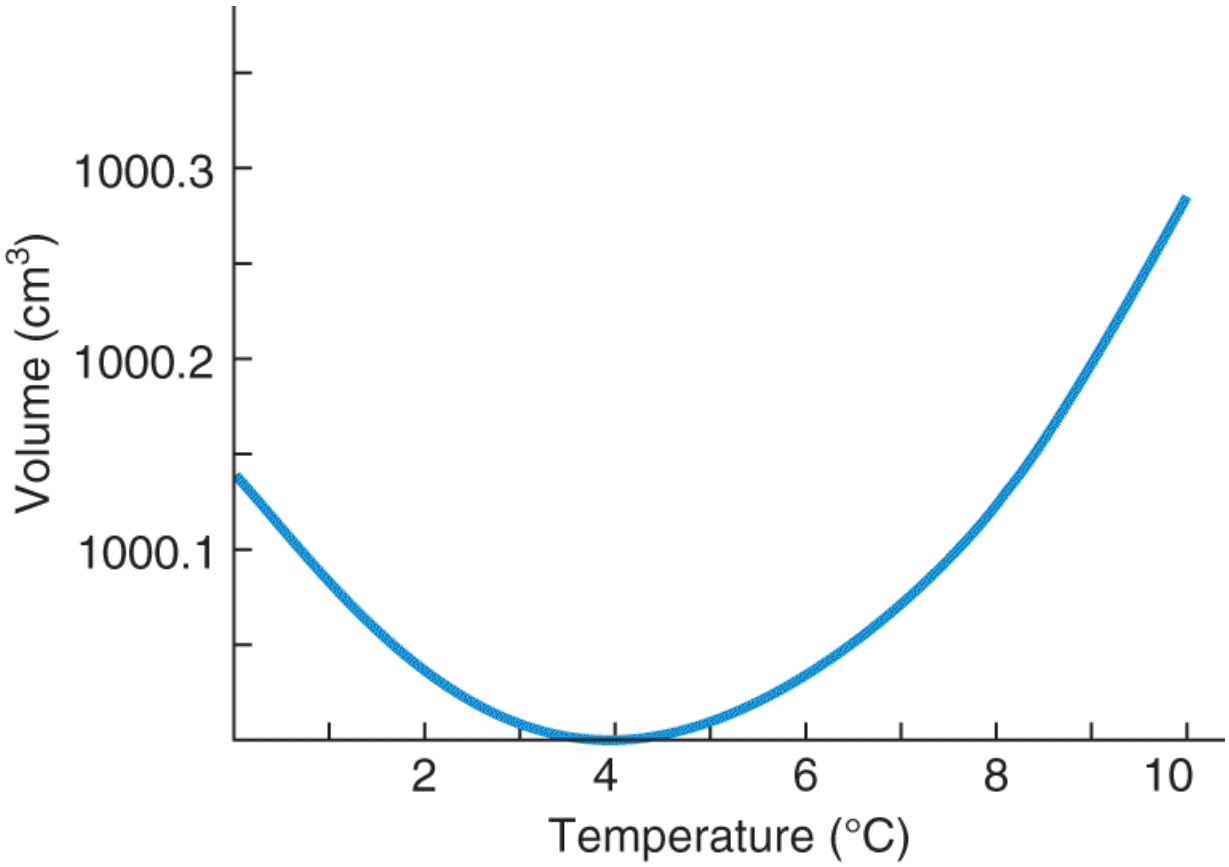


Figure 13.8 The thermal expansion/contraction of water close to 0 °C.

Solids that contract upon heating are still being reported, and include ZrW_2O_8 ($\alpha \sim -7 \times 10^{-6}$ K^{-1}), cordierite, $Mg_2Al_4Si_5O_{12}$, β-eucryptite, $LiAlSiO_4$, β-spodumene, $LiAlSi_2O_6$, ScF_3, $NaZr_2P_3O_{12}$, analogues of the compound Prussian Blue, $KFe^{3+}Fe^{2+}(CN)_6$, $Y_{0.22}Sm_{0.78}S$ ($\alpha \sim -65 \times 10^{-6}$ K^{-1}, and $Ca_2RuO_{3.74}$ ($\alpha \sim -115 \times 10^{-6}$ K^{-1}). Many of these solids show a variation in the thermal contraction over a wide composition range. For example, β-spodumene can take compositions $Li_2Al_2Si_nO_{4+2n}$, in which n can take values from 4 to 9. Similarly, $NaZr_2P_3O_{12}$ can form solid solutions in which phosphorus is replaced by silicon ($Na_{1+x}Zr_2P_{3-x}Si_xO_{12}$) and sodium by calcium and strontium ($Ca_{1-x}Sr_xZr_4P_6O_{24}$). This ability allows the thermal expansion and contraction to be carefully tailored, and materials with almost zero thermal expansion (ZTE) can be fabricated. The details of the contraction observed and the temperature range over which this

contraction is found, varies widely. Cubic ZrW_2O_8 and a number of others show contraction along all axes as the temperature rises. Many others are anisotropic and show a contraction along one or two crystallographic axes as the temperature increases even though other directions may show normal thermal expansion, exemplified by $CaZr_4P_6O_{24}$ (Figure 13.9).

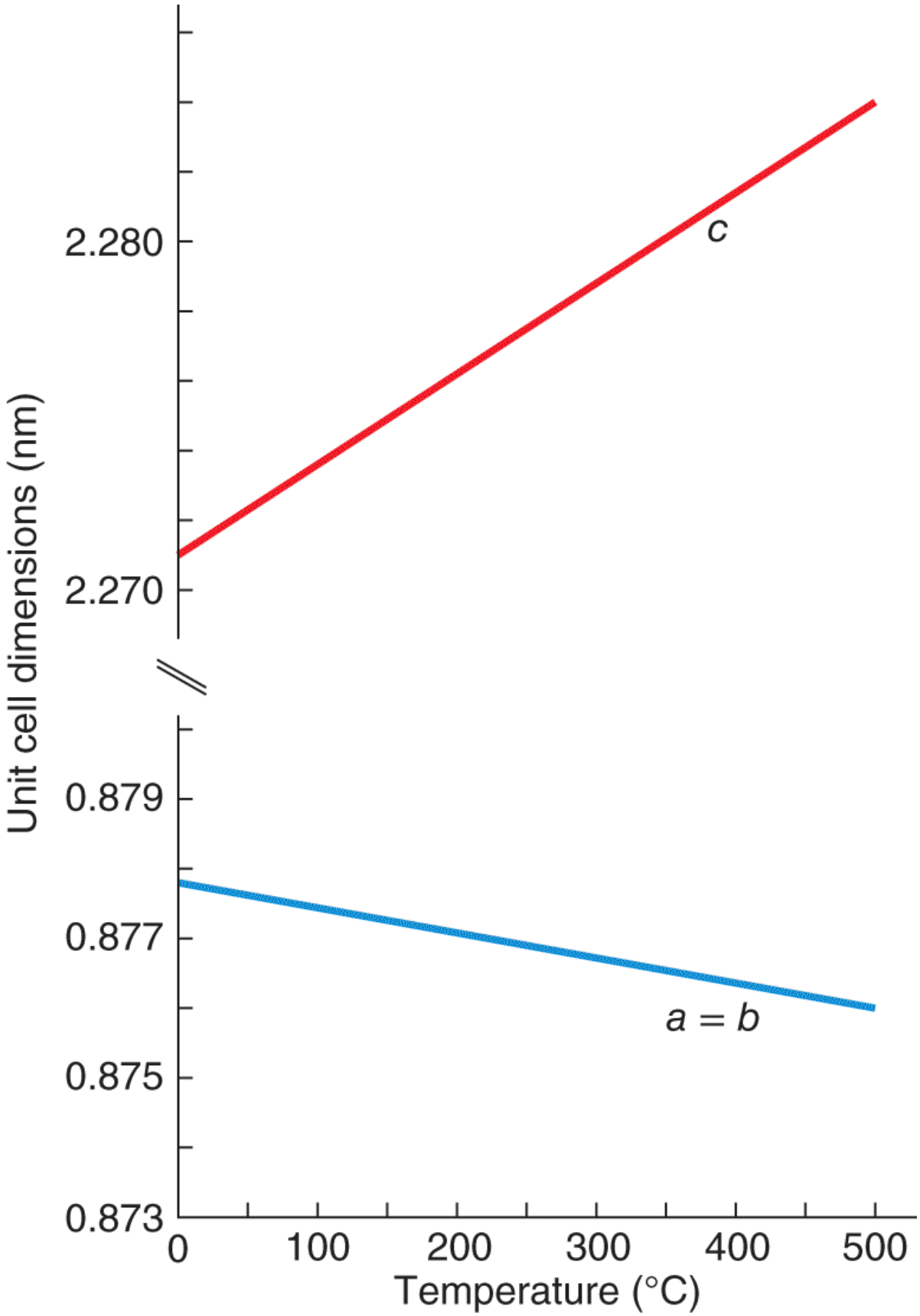

Unit cell dimensions (nm)
2.280
2.270
c
0.879
0.877
0.875
0.873
a = b
0
100
200
300
400
500
Temperature (°C)

Figure 13.9 The thermal expansion of the **c**-axis and contraction of the **a**- and **b**-axes for $CaZr_4P_6O_{24}$.

Source: Data from Agrawal (1994).

Thermal contraction is not the result of a single mechanism. In a number of structures that can be considered to be built of corner-shared polyhedra, the contraction can be the result of changing polyhedral distortion and linkage. For example, in many ABO_3 perovskite and perovskite-related structures at lower temperatures, the metal–oxygen BO_6 octahedra have one B□O□B bond longer than the other two, leading to an elongated octahedron. As the temperature increases, this elongation tends to decrease because of changes in vibrational energy and a decrease in anion–anion repulsion. This change can lead to a relative contraction in unit cell dimensions along this direction as the temperature increases. At the same time the undistorted O–B–O diagonals expand as normal. When the contraction outweighs the normal expansion of the lattice, the material shows overall thermal contraction (Figure 13.10). This behaviour exhibited by $PbTiO_3$.

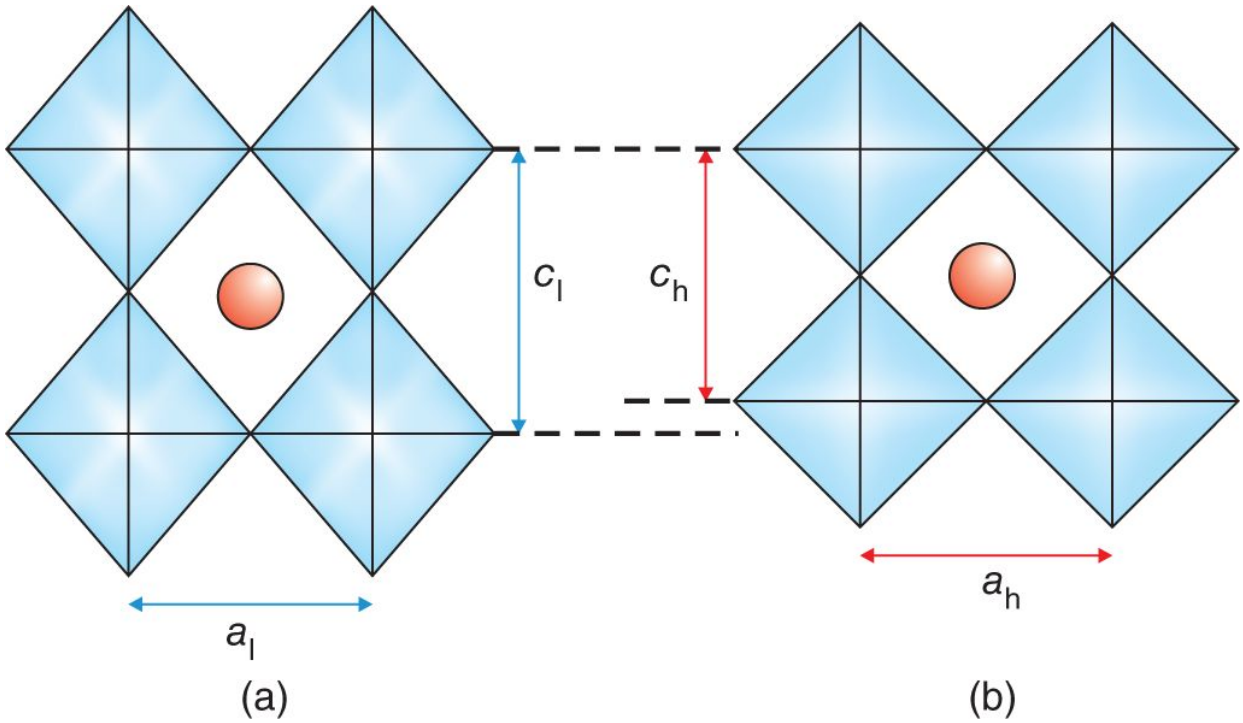


Figure 13.10 Thermal contraction due to the removal of octahedral distortion in the perovskite structure. (a) Low temperature and (b) high temperature.

Contraction can also be achieved by 'rocking' or cooperative rotation of the polyhedra, found in cubic Zr_2WO_8, $NbOPO_4$, and $ThO(PO_4)_2$. These solids have relatively open structures with a skeleton composed of corner-linked AO_6 octahedra and BO_4 tetrahedra. In this group of materials, the polyhedra remain the same shape, but as the temperature rises, increased vibration allows the polyhedra to rotate slightly, producing a contraction of the unit cell (Figure 13.11). This type of mechanism has been suggested for thermal contraction in SrF_3, which has a structure of corner-linked SrF_6 octahedra, similar to that of the TiO_6 corner-linked octahedra framework of the perovskite $SrTiO_3$ (Figure 3.11).

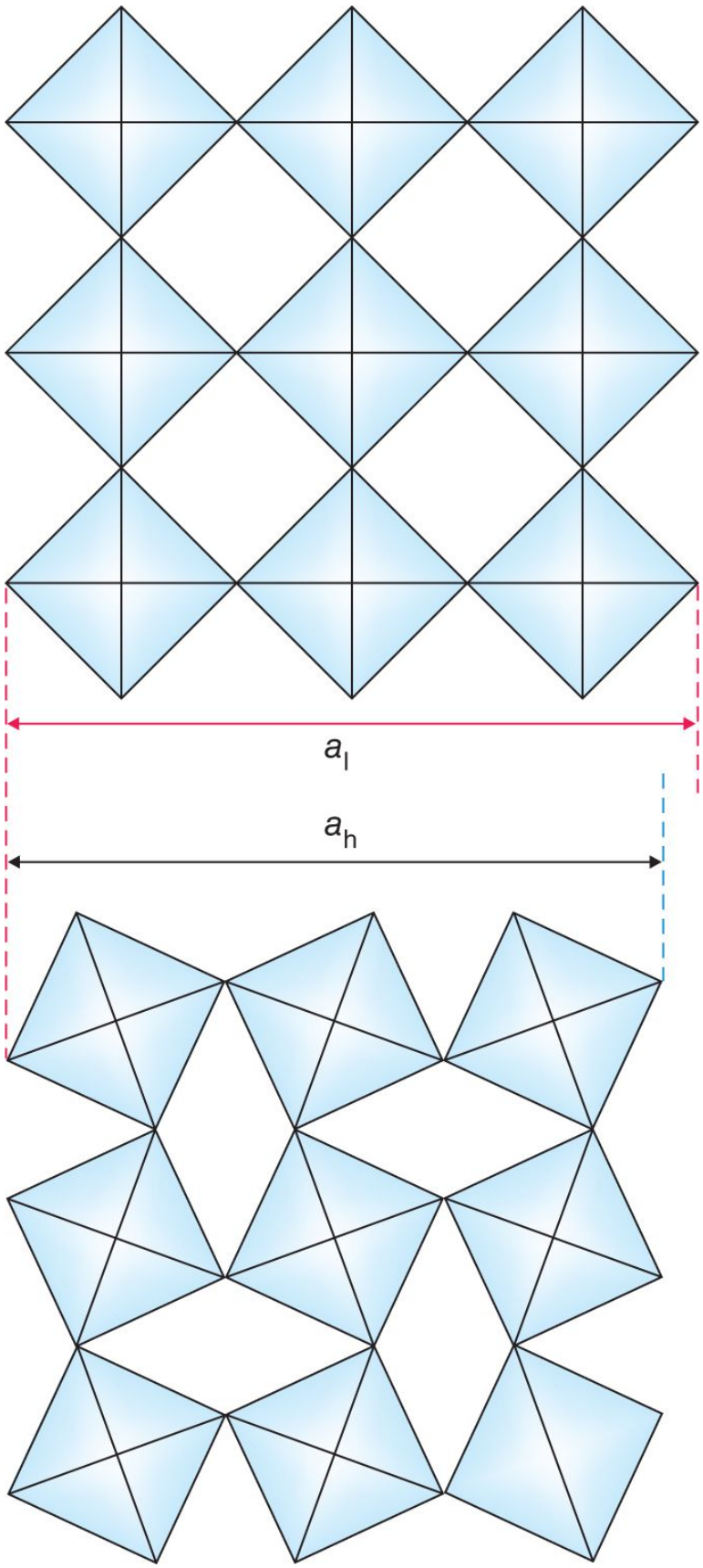

a_l
a_h

Figure 13.11 Thermal contraction due to 'rocking' (rotation) of octahedra.

In the families of cordierite, β-eucryptite, β-cordierite, and $NaZr_2P_3O_{12}$, a mechanism similar to that giving rise to auxetic (negative Poisson's ratio) materials seems to occur (Section 8.5.2). These structures are built from inflexible layers, similar to those found in clay minerals linked by Si□O□Si and O□Si□O bonds (Figure 13.12). As the temperature rises the layers expand, mostly laterally. The Si□O or P□O bonds linking the layers are strong and do not break to relieve the stress generated. Instead, the bond angles change, and the groups act as hinges that pull the layers closer, giving rise to thermal contraction in a direction normal to the layers.

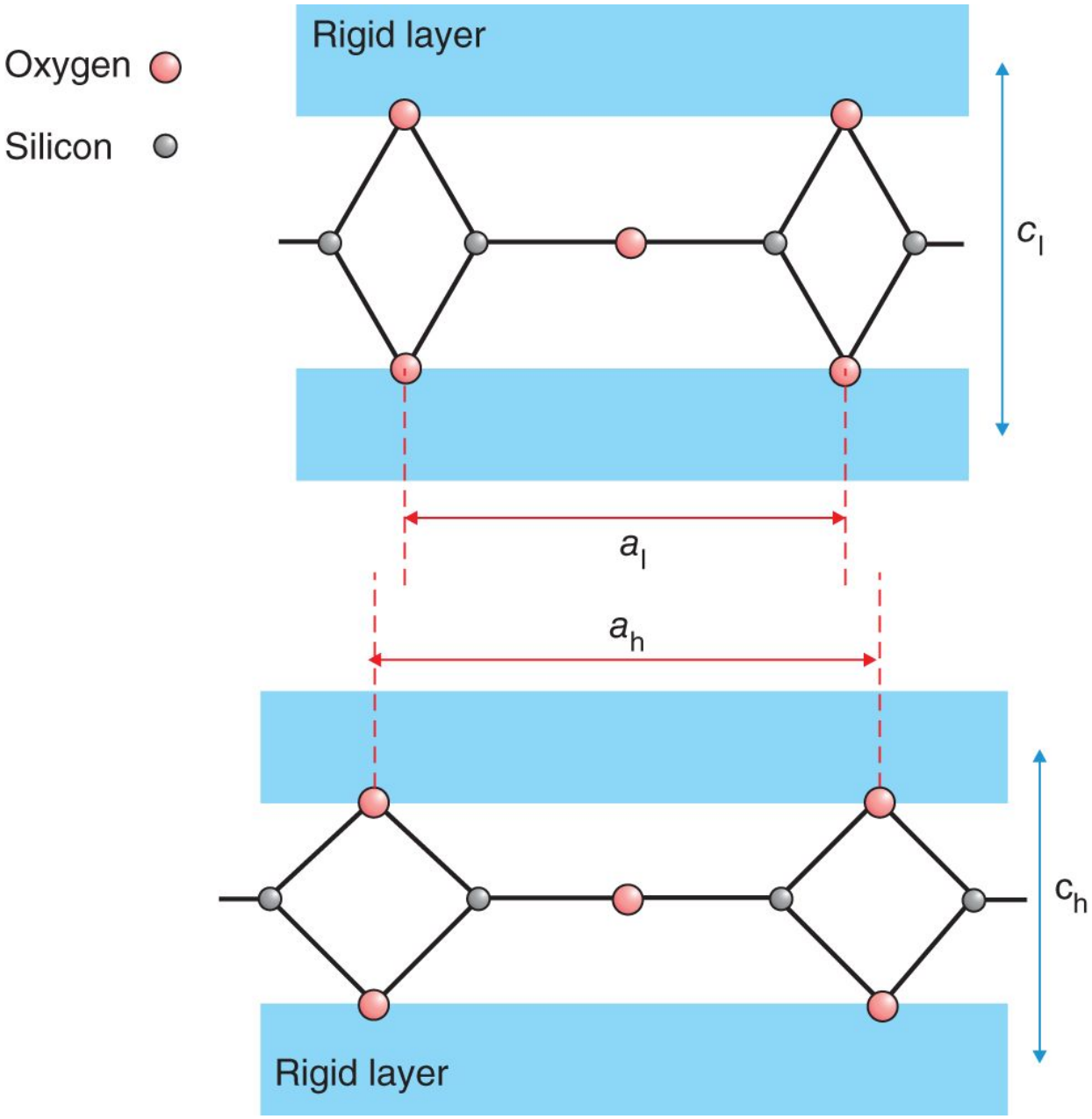


Figure 13.12 Thermal contraction brought about by silicon-oxygen 'hinges' connecting rigid sheets in some silicates.

13.3.4 Zero Thermal Contraction Materials

Clearly materials that do not expand at all as the temperature rises would be of value for many purposes, especially in microelectronic devices that may become warm due to power consumption. For this reason, materials that show zero thermal expansion (ZTE) are being actively sought. In the past a zero thermal expansion solid could be

constructed by using a composite of two materials, one showing negative thermal expansion and one showing positive thermal expansion. However, a number of solids have now been fabricated that show virtually no expansion at all as the temperature rises. Notable among these are the alloy YbGaGe, compounds related to Prussian blue, including $Fe[Co(CN)_6]\cdot nH_2O$ and $Ni(CH_3)_4CuZn(CN)_6\cdot nH_2O$ and a number of nitride phases $(Cu,Sn)NMn_3$, $(Zn,Sn)NMn_3$, and $(Cu,Ge)NMn_3$, such as $Ge_{0.5}Cu_{0.5}NMn_3$. These latter solids have an antiperovskite structure in which the Cu and Ge atoms (or their equivalents) occupy the A positions, the N atoms take the octahedrally coordinated B positions and the Mn atoms occupy the positions of oxygen in the ABO_3 perovskite structure. In them the zero thermal expansion is due to a combination of a normal thermal expansion of the structure on the one hand and a shrinkage of the unit cell due to magnetic ordering between the Mn atoms on the other. This means that the degree of expansion or contraction can be controlled by manipulating the Mn contribution via the introduction of defects on the Mn sites in the structure.

13.4 Thermoelectric Effects

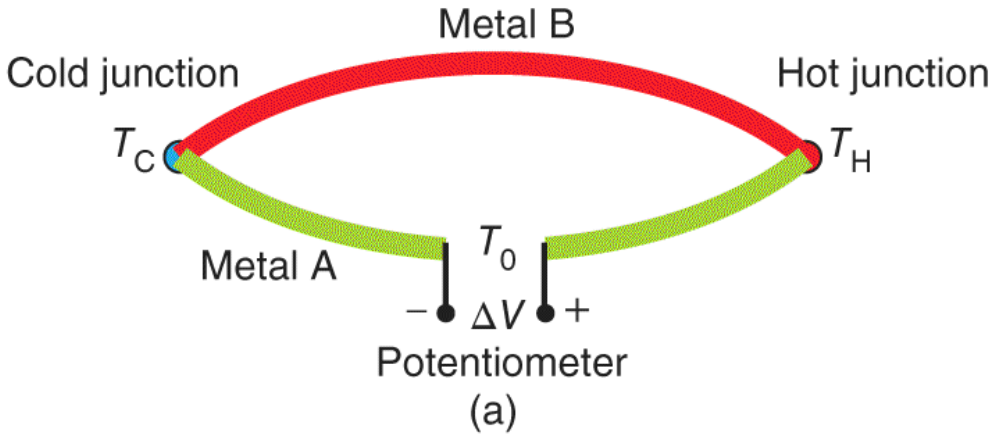
13.4.1 Thermoelectric Coefficients

The conduction of electricity and the conduction of heat are closely linked. These have been treated separately in Chapter 11 and this chapter (Chapter 13) because it has been implicit that only one effect was important at any time. However, the conduction of heat affects electrical conduction and vice versa, and forms the subject of *thermoelectricity*. The topic is of current importance for the conversion of 'waste' heat into electrical energy.

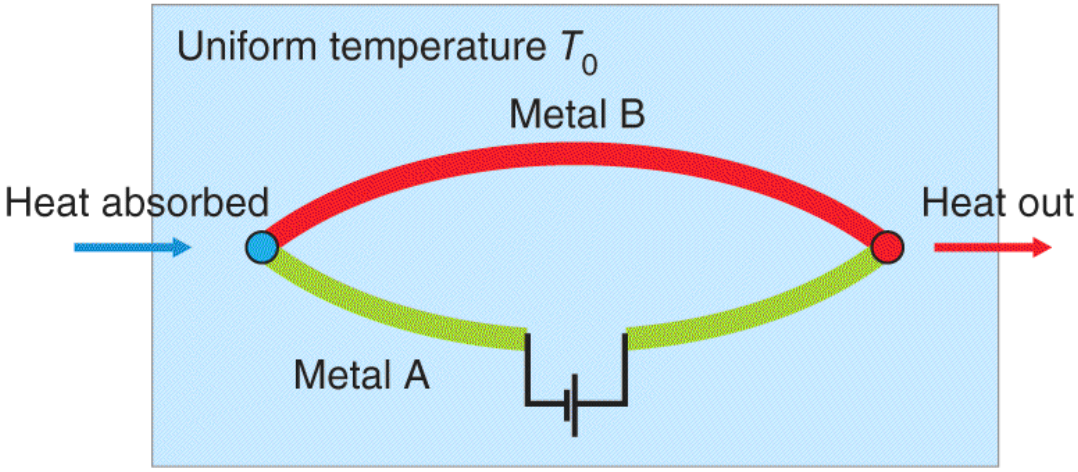
The first thermoelectric effect to be discovered was the *Seebeck effect*, reported in 1821. It is most often described in terms of a circuit made up of two dissimilar metals. In this arrangement, a current flow is induced in a circuit made of different conductors A and B, when the junctions between the materials are held at different temperatures. The effect is generally observed by breaking the circuit and observing the voltage generated with a potentiometer (Figure 13.13a). This is given by:

$$\Delta V_{AB} = \Sigma_{AB} (T_H - T_C) = \Sigma_{AB} \Delta T$$

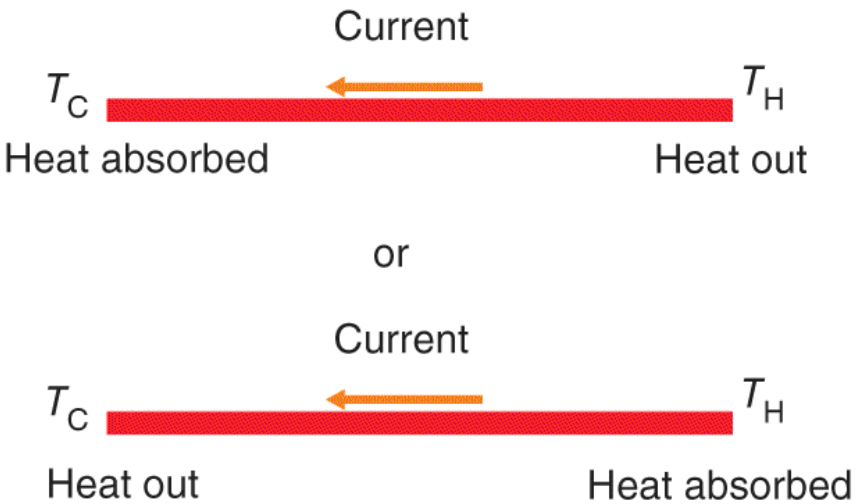
where Σ_{AB} is called the *thermoelectric power* or *Seebeck coefficient* (often written as S), ΔV_{AB} is the voltage measured, and ΔT is the temperature difference between the hot junction, T_H, and the cold junction, T_C. The voltage depends only upon the two materials chosen and the temperature difference between the junctions. The Seebeck coefficient, which varies significantly with temperature, is of the order of 10 μV K^{-1} for metals and 200 μV K^{-1} for semiconductors.


Metal B
Cold junction
Hot junction
T_C
T_H
Metal A
T_0
$-$
ΔV
$+$
Potentiometer

(a)


Uniform temperature T_0
Metal B
Heat absorbed
Heat out
Metal A

(b)


Current
T_C
T_H
Heat absorbed
Heat out
or
Current
T_C
T_H
Heat out
Heat absorbed

(c)

Figure 13.13 Thermoelectric effects: (a) the Seebeck effect; (b) the Peltier effect; (c) the Thompson effect.

The complementary effect, discovered in 1834, is the *Peltier effect*. Again, this is most often described in terms of a circuit consisting of two different conductors. In such a circuit a current flow induces a temperature difference between the two junctions (Figure 13.13b). Heat is liberated at one junction and absorbed at the other. If the direction of the current is reversed, the heat output and input occur at the opposite junctions. The amount of heat produced or absorbed, ΔQ, is given by:

$$\Delta Q = \Pi_{AB} I t$$

where Π_{AB} is the *Peltier coefficient*, I is the current flowing in the circuit, and t is the time.

The Seebeck and Peltier effects were shown to be connected by Thomson (later Lord Kelvin) in 1854. The relationship is:

$$\Pi = \Sigma T$$

where T is the temperature (K). Thomson also predicted the existence of a third thermoelectric effect now known as the *Thomson effect*, in which a reversible heating or cooling is observed when a current flows along a (single) conductor that has one end at a different temperature than the other (Figure 13.13c). The amount of heat energy absorbed or given out, ΔQ, is given by:

$$\Delta Q = \tau I t \Delta T$$

where τ is the Thompson coefficient, I is the current flowing for a time t, and ΔT is the temperature difference between the points of measurement. (Note that a current passing through a conductor will heat it. The effect is called *Joule heating* and is used in, for example, electric fires (Section 11.1.3). Peltier or Thomson heat production are quite different from Joule heating.)

Although usually described in terms of junctions in circuits, the Seebeck and Peltier coefficients are *not* caused by the junctions themselves. All materials that contain mobile charge carriers show

thermoelectric effects when heated. That is, thermal gradients produce electrical effects and electrical effects produce thermal gradients. Thermoelectric effects are properties of pure materials and a material is characterised by an *absolute Seebeck coefficient*, α_S, and an *absolute Peltier coefficient* π. (The Thomson coefficient, τ, only refers to a single material.) For example, a single-phase sample with one end maintained at a high temperature and the other at a low temperature will develop a potential difference across its length due to the Seebeck effect (Figure 13.14). As would be expected, these coefficients vary with direction in non-cubic crystals.

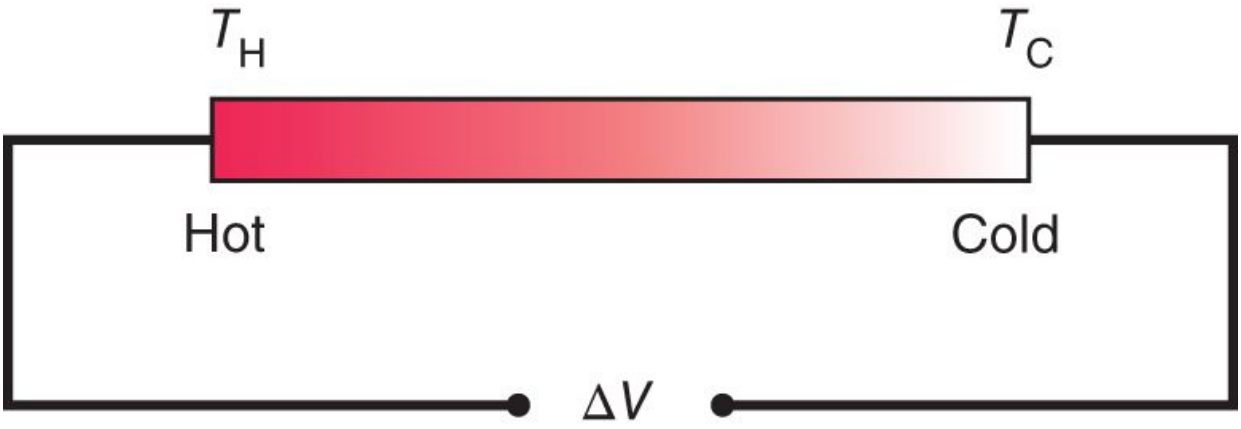


Figure 13.14 The Seebeck effect in a single phase.

The Seebeck coefficient, Σ, and the Peltier coefficient, Π, apparent in circuits made of two different electronically conducting materials are *relative* coefficients, that is, the difference between the absolute coefficients of the two materials. For the arrangement in which the positive terminal of the metal A is connected to the hot junction (Figure 13.13a), the Seebeck coefficient of two the materials Σ_{AB}, at a temperature T, is given by the difference between the absolute Seebeck coefficients of the components:

$$\Sigma_{AB} = \alpha_S(A) - \alpha_S(B)$$

13.4.2 Thermoelectric Effects and Charge Carriers

Thermoelectric effects can be explained by considering the electron, hole, and phonon distributions in a material. It is apparent that the charge carriers near to the Fermi surface in the hot region of a single

material will have a higher kinetic energy, and hence a higher velocity, than those in the cold region. This means that the net velocity of the charge carriers at the hot end moving towards the cold end will be higher than the net velocity of the charge carriers at the cold end moving towards the hot end. In this situation, more carriers will flow from the hot end towards the cold end than vice versa. This will cause a voltage to build up between the hot and cold ends of the sample. Eventually equilibrium will be established and a potential will be set up. The same is true for the phonons. As phonons interact strongly with electrons and holes, they will drag these along with them, to create an additional potential. The magnitudes of the measured thermoelectric coefficients are a complex function of both of these features and they vary considerably with temperature.

Despite the intricate nature of the relationship between phonon and electron transport, the preceding ideas can be used to estimate the magnitude of the thermoelectric coefficients of a material theoretically, provided simplifying assumptions are made. In the case of classical doped semiconductors such as n- and p-type silicon, the Seebeck coefficient for electron transport Σ_e can be approximated by:

$$\Sigma_e = -(k_B/e)\,[5/2 - \ln(N_c/n) + a_e]$$

and for hole transport Σ_h is given by:

$$\Sigma_h = +(k_B/e)\,[5/2 - \ln(N_v/p) + a_h]$$

where N_c is the density of states in the conduction band, N_v is the density of states in the valence band, k_B is the Boltzmann constant, n is the concentration of mobile electrons, p the concentration of mobile holes, and a_e and a_h are correction terms that take into account thermal collisions and related factors. Writing:

$$n = N_c \exp\,[(E_F - E_c)/k_B T]$$

$$p = N_v \exp\,[(E_v - E_F)/k_B T]$$

(Section 11.2.1) and ignoring corrections we find:

$$\Sigma_e = -(k_B/e)\,[5/2 - (E_F - E_c)/k_B T]$$

$$\Sigma_h = +(k_B/e)\,[5/2 - (E_v - E_F)/k_B T]$$

A simple test of the nature of the charge carriers in a semiconductor can thus be made by measuring the sign of the voltage developed when one end of the semiconductor is hotter than the other. In the case of semiconductors that have mobile electrons, i.e. *n*-type semiconductors, the colder end of the rod will be *negative* with respect to the hotter end and the sign of the Seebeck coefficient is negative. In the case where the mobile charge carriers are positive holes, i.e. *p*-type semiconductors, the colder end of the rod will be *positive* with respect to the hotter end, making the Seebeck coefficient positive.

The Peltier coefficient for electron transport, Π_e, can be obtained by similar reasoning and is given by:

$$\Pi_e = [E_c - E_F + (¾)\,k_B T]/-e$$

and for hole transport, Π_h is given by:

$$\Pi_h = [E_F - E_v + (¾)\,k_B T]/+e$$

where E_c is the energy at the bottom of the conduction band, E_v is the energy at the top of the valence band, E_F is the Fermi energy, and e is the electron charge, equal to the hole charge.

13.4.3 The Seebeck Coefficient of Solids Containing Point Defect Populations

A population of point defects in a nominally insulating crystal, such as simple oxides NiO, CoO, and FeO, or more complex materials such as perovskites, ABO_3, spinels, and AB_2O_4 among many others, tends to generate a population of mobile charge carriers. The sign and magnitude of the Seebeck coefficient can provide a measure of the concentration of charge carriers, the nature of the charge carriers, and, with some simple assumptions, the number of defects present in

these materials. As in the case of classical semiconductors, the colder end of a bar of material will be *negative* with respect to the hotter end if the material contains mobile electrons, making the Seebeck coefficient negative. Materials containing mobile holes will show the converse and the colder end of the rod will be *positive* with respect to the hotter end, making the Seebeck coefficient positive.

This can be illustrated with reference to the perovskite structure cobaltite $LaCoO_3$, which is generally slightly oxygen rich when prepared in air. Any additional oxygen ions, O^{2-}, will form a negatively charged point defect population and in order to maintain charge neutrality some compensating positively charged electronic defect must be generated at the same time. The La ions have a fixed valence hence the charge balance must be preserved by the Co^{3+} ions, which have an easily variable valence. In this case each added O^{2-} ion will generate *two* Co^{4+} ions to maintain neutrality. Now a Co^{4+} ion can be regarded as Co^{3+} together with a hole, but this might be strongly bound to the Co^{3+}, that is, the Co^{4+} state is very stable, or it could be weakly bound and jump from one Co^{3+} to another under the influence of temperature or an electric field. It is found that the compound usually shows a large positive Seebeck coefficient, of the order of +700 μV K^{-1}, indicating that weakly bound holes are present.

The relationship between the number of defects and the Seebeck coefficient is obtained by estimating the configurational entropy of the defect-containing material. A number of forms for this estimate are found, each depending upon slightly different approximations. The simplest is:

$$\alpha = \pm(k_B/e)\,[\ln(n_0/n_d)] \tag{13.3}$$

where n_o is the number of sites in the sublattice containing defects and n_d is the number of defects giving rise to mobile electrons or holes. The positive version applies to hole conducting materials and the negative expression to electron conducting materials. Note that n_0/n_d increases as the number of defects falls and so the value of α is expected to be greatest for lowest defect populations, precisely when

other methods of analysing defect populations give least precision. Equation (13.3) is formally equivalent to the *Heikes equation*:

$$\alpha = \pm(k/e)\ \ln[(1-c)/c]$$

where c is the fraction of defects (or mobile charge carriers) present.[1] Using Eq. (13.3), and the value of +700 μV K^{-1} for the value of the Seebeck coefficient for $LaCoO_3$ given earlier, the value of n_d, the number of holes (or Co^{4+} ions) present, is $3 \times 10^{-4}\ n_0$ (the number of Co^{3+} ions present). Each Co^{4+} ion is generated by ½ O^{2-} added to the overall composition of the phase so that the material has a formula $LaCoO_{3.00015}$. This population of Co^{4+} ions will also change the magnetic properties of the phase, which can now be estimated from the known composition.

13.4.4 Thermocouples, Power Generation, and Refrigeration

Thermocouples are a widely used application of the Seebeck effect, and they are the main means of temperature monitoring and regulation above about 150 °C. They simply consist of a loop of two different metals, one junction of which is placed in contact with the hot object whose temperature is to be measured whilst the other is maintained at 0 °C (Figure 13.15). The voltage generated by a thermocouple is related to the temperature to be measured by a polynomial function:

$$T_H = a_0 + a_1 V + a_2 V^2 + \cdots + a_n V^n$$

where T_H is the temperature of the hot junction, V is the measured voltage of the thermocouple, and a_0, a_1, and so on are empirical constants. These coefficients depend upon the reference junction temperature and the materials used in the device. The relationship between temperature and voltage is usually found by reference to 'thermocouple tables' supplied by manufacturers or located in handbooks.

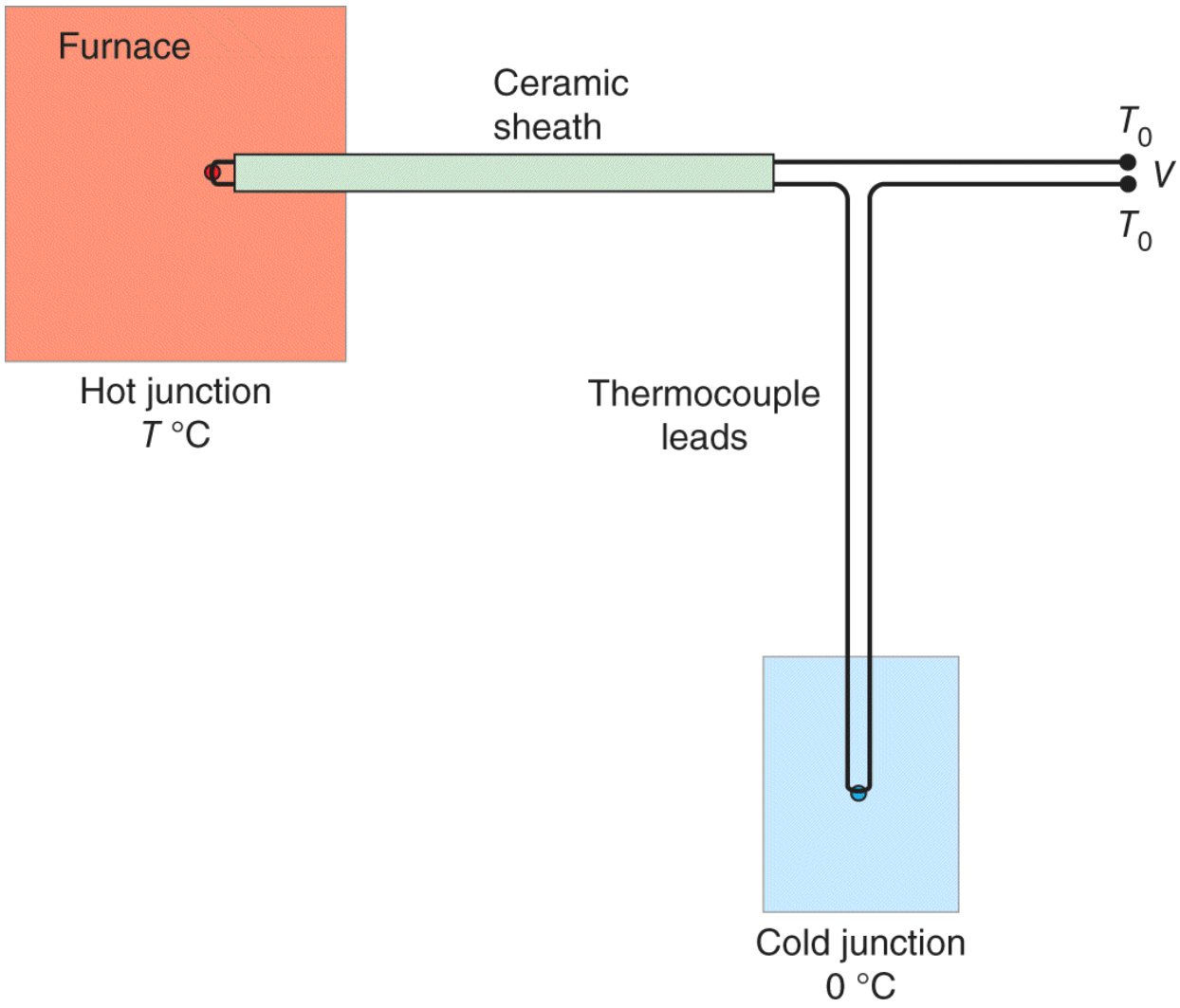


Figure 13.15 Thermocouple temperature measurement.

A series of thermocouples linked in series forms a *thermopile* (Figure 13.16). This has increased sensitivity when used to measure temperatures compared with a single thermocouple. The same arrangement can be used as a power generator. For this purpose, the low temperature junctions, T_C, are maintained at a fixed temperature by connecting them to a heat sink, and the high temperature junctions are in contact with a heat source, such as a furnace or radioactive sample. In this mode they can make use of 'waste heat' and as such are of considerable value.

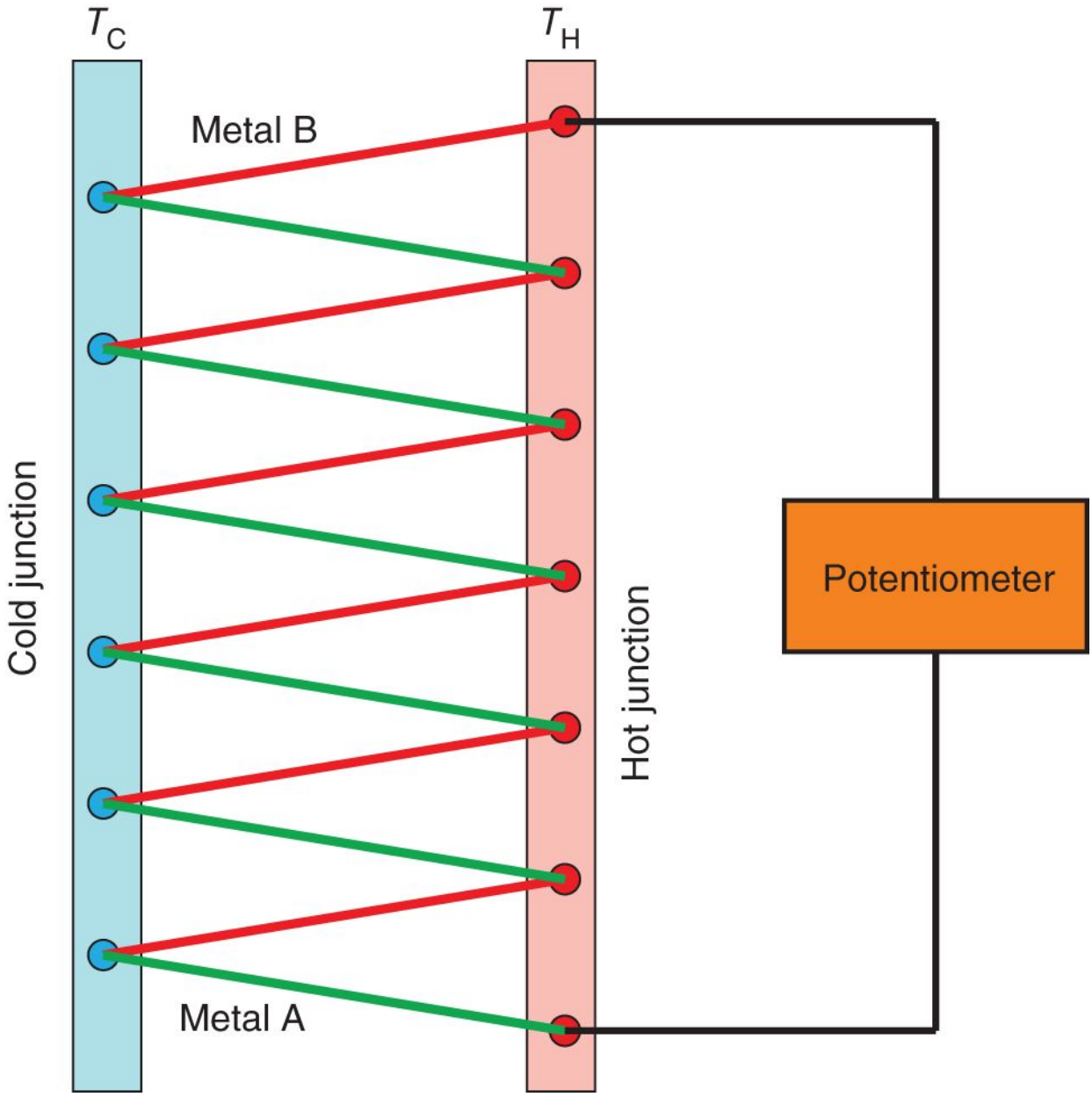


Figure 13.16 A thermopile: thermocouples connected in series.

Thermoelectric materials are also used for heating, refrigeration, and electricity generation utilising the Peltier effect. In both cases two thermoelectric materials, n-type and p-type, are coupled by metal plates, which act as the junctions. A current passed through the circuit in one direction will heat one plate and cool the other (Figure 13.17a). If the temperature of the hot junction is constant, maintained by connection to a heat sink, continuous cooling will occur at the cold junction. Alternatively, if the temperature of the cold

junction is fixed, continuous heating will occur at the hot junction. A reversal of the current will change the hot plate and cold plate. Such devices, called *heat pumps*, are widely used in food and drinks coolers powered by car batteries. Similarly, if one plate is continuously maintained hotter than the other, a current will flow in the circuit, and power is generated (Figure 13.17b). In this format, these devices are used in space probes that operate too far from the sun for photoelectricity to be used for power supplies. In such cases, heat is generated by the slow decay of radioactive isotopes.

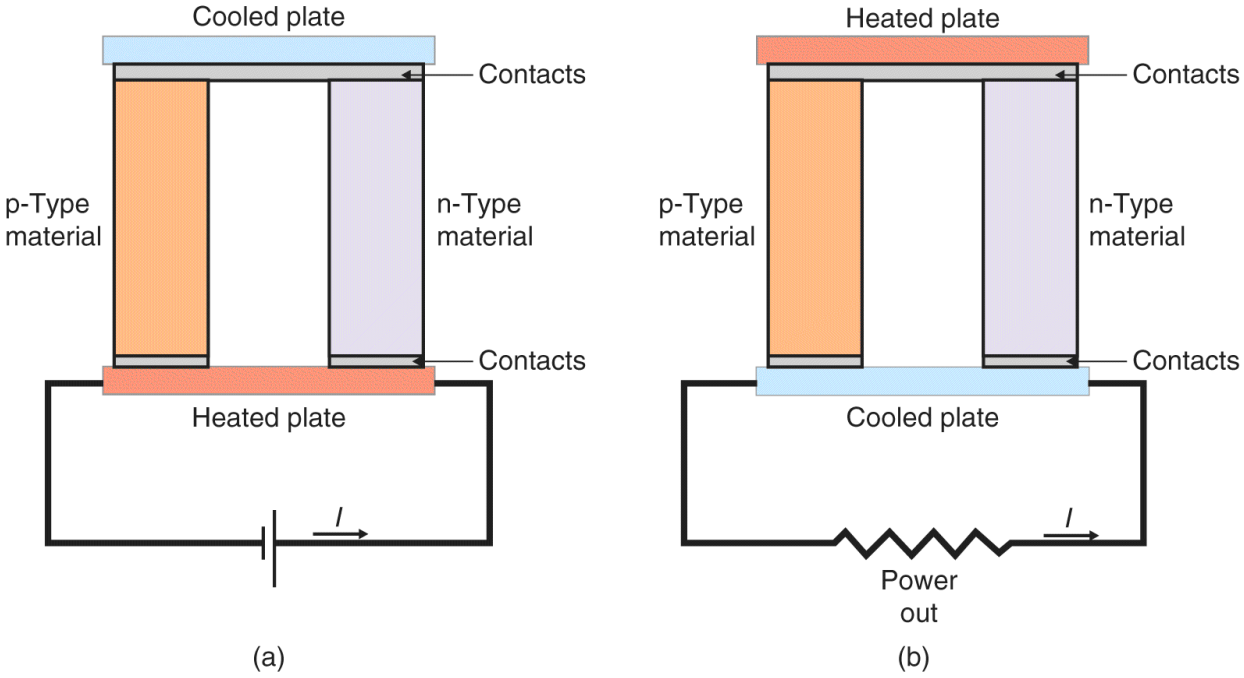


Figure 13.17 Use of the Peltier effect for (a) thermoelectric heating; (b) thermoelectric power generation.

The effectiveness of devices using thermoelectric effects depends upon the magnitude of the relative Peltier coefficient, Π, or its equivalent, the relative Seebeck coefficient, Σ. However, these are not the only materials parameters of importance. As an example, consider the operation of a heat pump. The amount of heat produced or absorbed, ΔQ, is:

$$\Delta Q = \Pi_{AB} I = \Sigma T I$$

where I is the current flowing. The requirement for a large relative Seebeck coefficient acts to rule out metals as components, as metals

have very low Seebeck coefficients. However, a low electrical resistivity is also needed, to cut down on Joule heating, thus favouring metals. Additionally, the thermal conductivity of the thermoelectric elements must be low to reduce the flow of heat from the hot to the cold region. This suggests an insulator, but these have high values of electrical resistivity. All of these conflicting factors are taken into account by using a *figure of merit*, *ZT*, for the material, given by:

$$ZT = \Sigma^2 \sigma T/\kappa$$

where Σ is the relative Seebeck coefficient of the thermoelectric elements, σ the electrical conductivity, and κ is the thermal conductivity, made up of with electron and phonon contributions. The best compromise is given by the material with the highest figure of merit. The figure of merit varies considerably with temperature, and although the best materials have a figure of merit of about 1.0, when combined in a device, an overall energy conversion efficiency of only a few percent is realised at present. For small portable coolers, solid solutions of the semiconductors bismuth telluride and antimony telluride, $Bi_xSb_{1-x}Te_3$, doped p- and n-type, are used. Space vehicles use more expensive silicon-germanium alloys, Si_xGe_{1-x}.

13.5 The Magnetocaloric Effect

13.5.1 The Magnetocaloric Effect and Adiabatic Cooling

Magnetic and thermal properties are linked via the *magnetocaloric effect*, although the observed changes are usually much weaker than those involved in thermoelectricity. The magnetocaloric effect describes the fact that ordinary magnetic materials heat up when placed in a magnetic field and cool down when the field is removed. The inverse magnetocaloric effect applies to materials that cool when placed in a magnetic field and warm up when the field is removed. The cause of the magnetocaloric effect is an order–disorder transformation in the magnetic lattice of the solid brought about by the application of a magnetic field. The first application of the magnetocaloric effect was *adiabatic cooling*, used to reach temperatures well below 1 K in the laboratory.

Cooling is achieved by the following steps:

(i) The magnetocaloric phase (or *working substance*), typically a paramagnetic salt or an alloy, is placed in a carefully insulated enclosure that prohibits heat transfer in or out of the system, at a temperature T without any magnetic field present, $\mathbf{H} = 0$ (Figure 13.18).

(ii) The magnetic field around the working substance is increased and as a consequence the magnetic dipoles align. This means that the entropy of the system decreases. Now as the process is adiabatic, that is, no heat passes in or out of the system, the entropy decrease is balanced by a temperature rise. The working material is now at a temperature $T + \Delta T$ in a magnetic field $\mathbf{H}$.

(iii) The temperature is reduced to T by removing a small amount of heat ΔQ by contact with a cooling medium. In low temperature work this is liquid or gaseous helium. The magnetic field is kept on to ensure that the dipoles stay aligned, so that the working substance reaches temperature T in a magnetic field $\mathbf{H}$.

(iv) The magnetic field is slowly decreased to zero while maintaining the adiabatic insulation. The magnetic dipoles

gradually disorder again and the energy for this transformation, coming from the thermal energy of the solid, causes it to cool to $T - \Delta T$.

(v) The cool working substance is brought into contact with the material to be cooled. This passes a small amount of heat ΔQ to the working substance and is itself cooled in the exchange, so that the condition of the working substance is now at a temperature T without any magnetic field present, identical to the initial state.

(vi) The whole cycle is now repeated until the material to be cooled reaches the desired temperature.

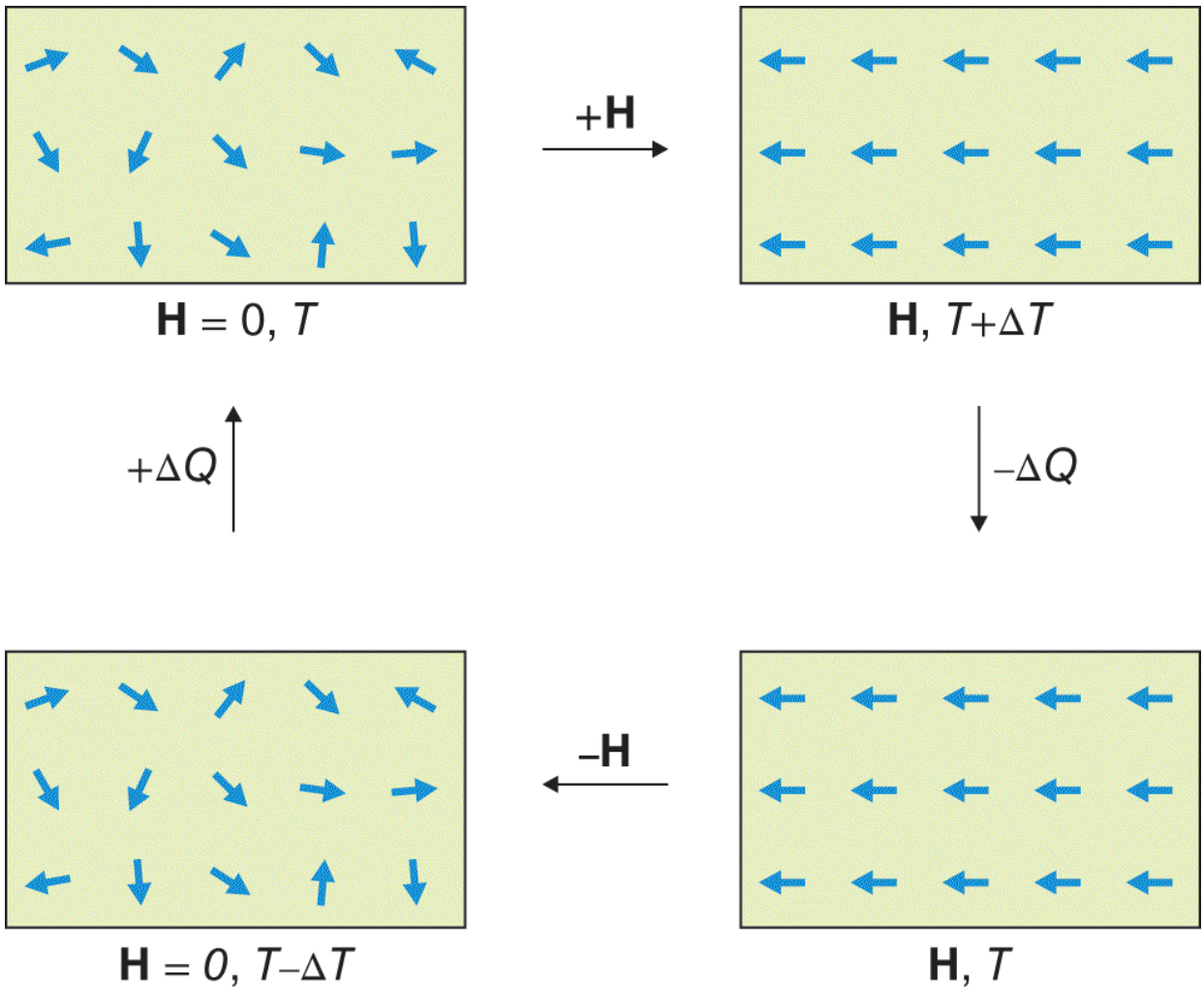


Figure 13.18 Adiabatic cooling: the initial state is regenerated by the cooling cycle which can then be repeated until the desired temperature is attained.

The technique only works well at low temperatures. This is because the change in temperature, ΔT, is inversely proportional to heat capacity, C_p. The heat capacity of all solids decreases rapidly at very low temperatures hence to obtain a sizable temperature change it is necessary that the working substance is maintained in a low temperature environment. The original materials used for adiabatic cooling were paramagnetic salts such as cerium magnesium nitrate, but these suffer from low thermal conductivity and have mostly been replaced by materials with higher thermal conductivity, including intermetallic alloys such as $PrNi_5$ and oxides including yttrium iron garnet and gadolinium iron garnet. Because of the significant

constraints in the technique, it has remained a laboratory method for investigating extremely low temperatures rather than a potential method of refrigeration.

13.5.2 The Giant Magnetocaloric Effect

The discovery of so-called *giant magnetocaloric effects* at room temperature has stimulated research into the feasibility of making both commercial and domestic refrigerators using magnetic cooling. This is because such units would be more energy efficient and environmentally benign than current refrigerators, which use gas compression to achieve cooling.

The first compounds to show the giant magnetocaloric effect were the intermetallic phases $Gd_5Si_{4-x}Ge_x$, $LaFe_{13-x}Si_xH$, and MnFeP(As, Ge), but more recent interest has centred on a group of *Heusler alloys* containing Ni and Mn, which are also shape-memory materials (Section 6.5.4). Heusler alloys are ferromagnetic intermetallic phases, typified by $AlMnCu_2$. The ferromagnetic ordering comes about by way of double exchange between the paramagnetic ions, notably Mn.

The giant magnetocaloric effect can be illustrated by reference to the changes in the ferromagnetic Heusler alloys $(Mn_{0.38}In_{0.62})MnNi_2$, $(In_{15.8}Mn_{34.8}Ni_{50.4})$, and $(Mn_{0.4}In_{0.6})MnNi_2$ $(In_{15.2}Mn_{35}Ni_{49.8})$. The transformation that gives rise to the magnetocaloric effect is a double conversion; a magnetically induced change from paramagnetic to ferromagnetic together with a simultaneous martensite to austenite crystallographic rearrangement of the sort that occurs in shape-memory alloys. The initial step is identical to that described in the previous section. The working substance, in the paramagnetic heavily twinned martensitic state, is held in adiabatic conditions in the absence of a magnetic field. The imposition of a magnetic field, **H**, results in a conventional magnetocaloric effect that generates a small temperature increase from T to $T + \Delta T_{mag}$ under adiabatic conditions. Concurrently the field also forces a crystallographic transition to take place from the martensite phase to a ferromagnetic high symmetry austenite phase (Figure 13.19). In the alloys mentioned, the austenite start temperature, A_s, is close to 317 K and the austenite finish temperature, A_f, is close to 327 K, both near room

temperature. Now this is a first-order transformation and involves an enthalpy change, the latent heat of transition. Under adiabatic conditions (in which no heat enters the system), this results in a large cooling $-\Delta T_{cryst}$. The net change is given by the sum of these two thermal effects:

$$\Delta T = +\Delta T_{mag} + (-\Delta T_{cryst})$$

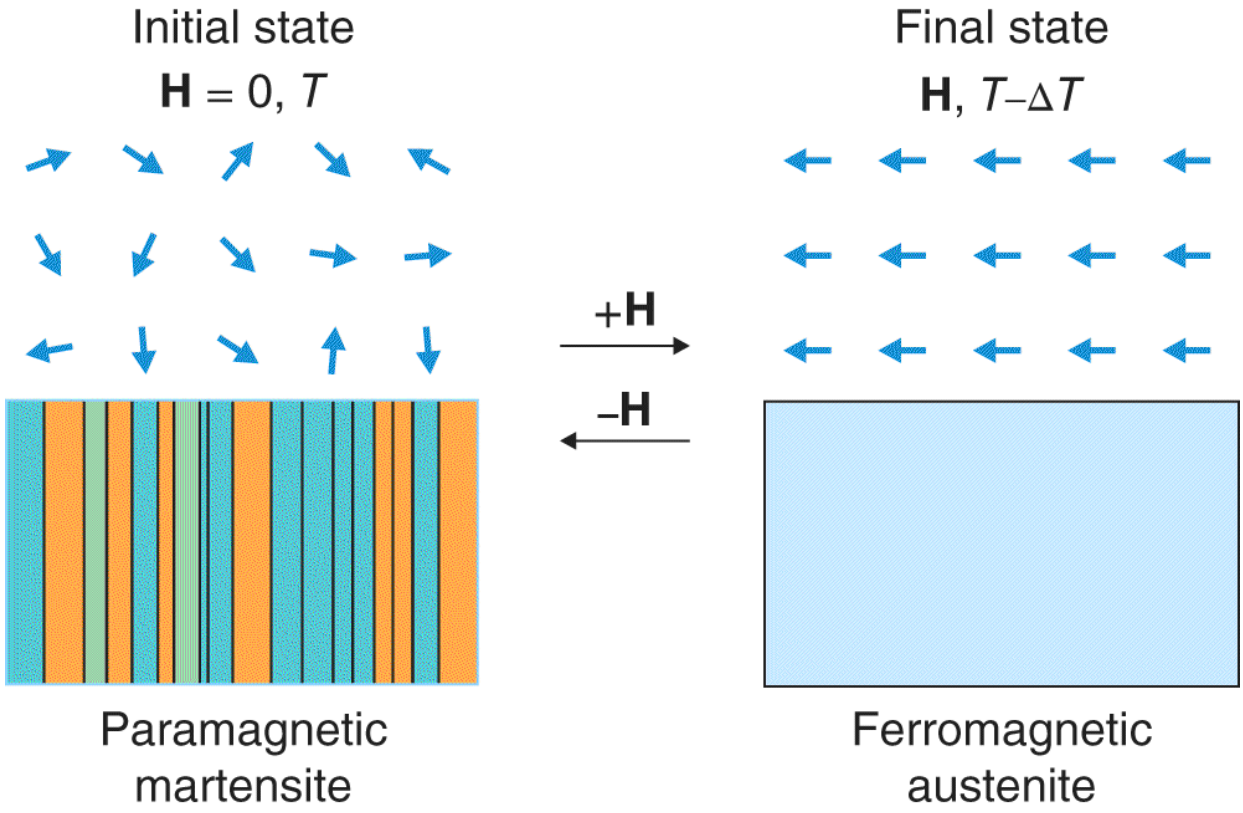


Figure 13.19 The giant magnetocaloric effect: the initial paramagnetic multiply twinned martensite is transformed to a ferromagnetic untwinned austenite accompanied by a drop in temperature.

The crystallographic term far outweighs the magnetocaloric term, resulting in a net cooling that, in the alloys cited, amounts to approximately −6 K. The ferromagnetic austenite working substance is now brought into contact with the material to be cooled and the magnetic field diminished. Heat exchange allows the initial martensitic phase at the initial temperature T to be regenerated. The cycle can then be repeated.

13.6 Thermochromic Effects

The term *thermochromic* refers to a change of colour of a material caused by a change in temperature. Within this group, there is no single mechanism of colour change, and every system has to be treated separately. Here, two examples are given to illustrate the phenomenon, one involving structural organisation and one involving a phase change.

13.6.1 Liquid Crystal Display Thermometers

Liquid crystal thermometers show a bright temperature reading against a dark background. As the temperature changes, different figures light up as part of a series (Figure 13.20). The cause of the colour is a diffraction effect. The display consists of sheets of asymmetric organic molecules, typically cholesteryl benzoate, acetate, and chloride. In each layer, these molecules are more or less parallel and the average orientation is described by a *director*. As the molecules in each layer are asymmetric, the interactions between the layers is are minimised if molecules in one layer rotate slightly compared to those in the layer below and, as the interactions are the same from one layer to another the same twist occurs between each layer. In this way a uniform helical structure results, formally called a cholesteric *chiral nematic* liquid crystal structure. The defining scale in these arrays is the distance over which the director rotates by 360°, which is called the *pitch* (*p*) of the helix (Figure 13.21).



Figure 13.20 Liquid crystal thermometer display.

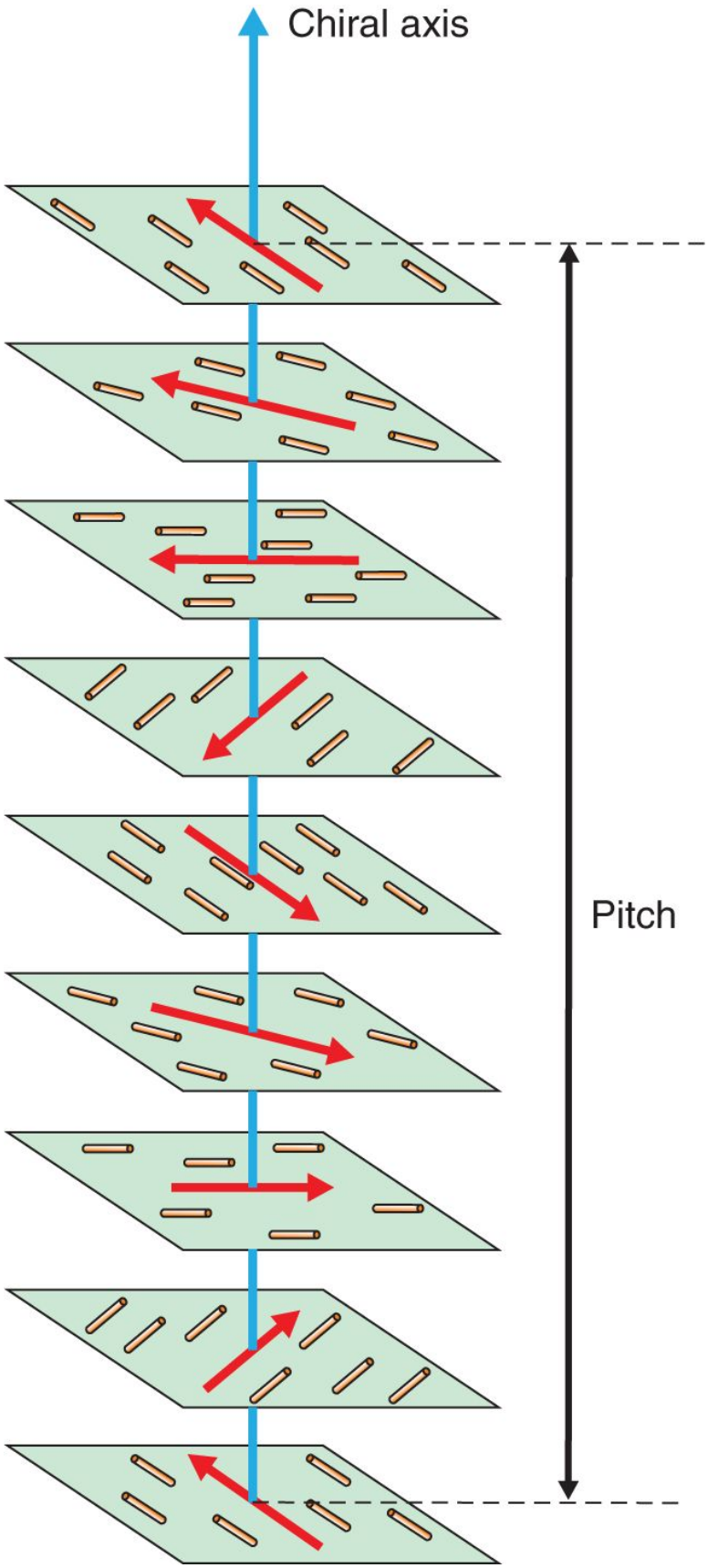

Chiral axis
Pitch

Figure 13.21 The cholesteric chiral nematic liquid crystal structure, schematic. The average orientation of the molecules (the director) is indicated by an arrow.

The optics of the array is similar that photonic crystals (Section 12.6.3). When the repeat distance along the helix is similar to the wavelength of light, diffraction will occur. For light normally incident on the film the central wavelength reflected, λ_0 is given by:

$$\lambda_0 = 2n_m p$$

where p is the pitch of the helix, and n_m is the average refractive index of each layer. When illuminated with white light, a wavelength satisfying this relationship will be diffracted back towards the observer and so will appear bright.

The thermometers operate in the following way. A sequence of different cholesteric structures, chosen so that the pitch of the molecular helix in each structure will diffract visible light at a precisely defined temperature, is arranged in a row. An increment of the temperature, δT, of the order of 1 °C is engineered between each successive material by additives or by slightly modifying the cholesteric molecule. Within the design temperature range, each spot will diffract light only when the phase is at the correct temperature and not do so otherwise. This causes the appropriate temperature value to 'light up'.

13.6.2 Vanadium Dioxide

In efforts to reduce the energy consumption of buildings, much effort has been put into studying thin-film coatings on windows in order to reduce heat loss in colder seasons and prevent heat entry in hot weather. One of the solids most investigated for this purpose is vanadium dioxide, VO_2. This phase has the rutile (TiO_2) structure (Figure 13.22) above 68 °C containing Ti^{4+} cation centred oxygen octahedra running in chains parallel to the **c**-axis. The room temperature phase, which exists below 68 °C, is a distorted form of the high temperature structure in which the chains of V^{4+} cation that run parallel to the **c**-axis form pairs and the surrounding oxygen

octahedra distort. The symmetry alters from tetragonal to monoclinic and, more importantly, the physical properties change from metallic to semiconducting. A thin film of the room temperature form applied to a window will effectively behave as a transparent thin-film coating (Section 12.4.2), allowing almost all of the incident visible and infra-red (heat) wavelengths to pass through (Figure 12.23a). When the temperature passes 68 °C, the film becomes metallic and reflects some visible wavelengths and especially most of the incident infra-red (heat) wavelengths (Figure 13.23b). The film now functions as a heat barrier, thus making an ideal coating for preserving the temperature inside the building.

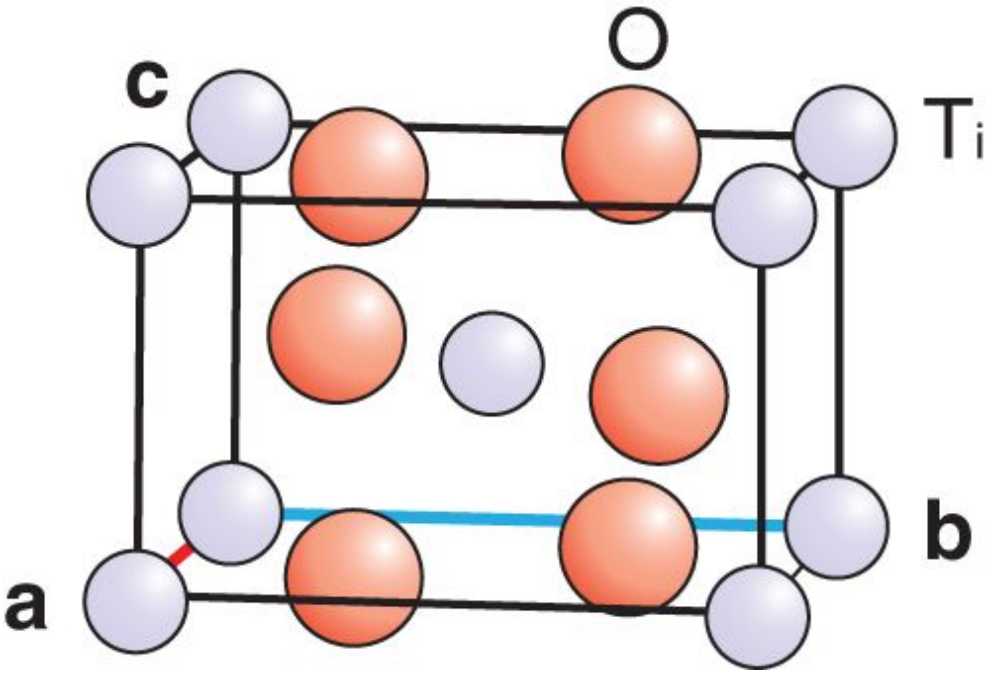


Figure 13.22 The rutile structure of metallic VO_2 above 68 °C.

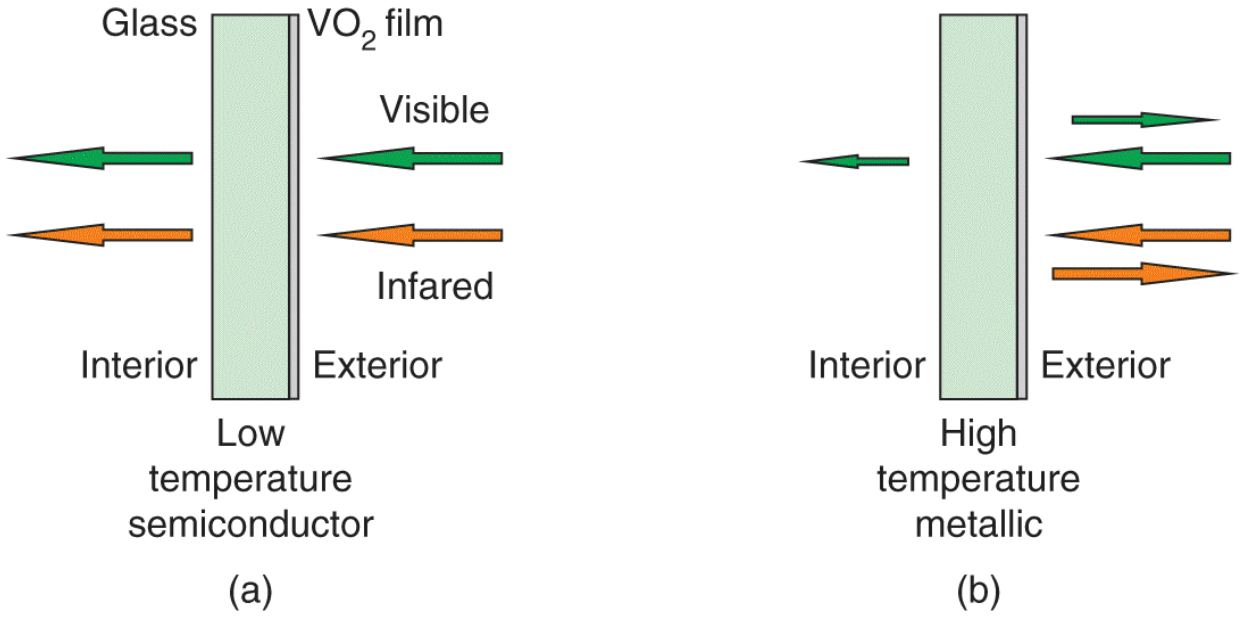


Figure 13.23 VO_2 thermochromic coating, schematic: (a) below the transition temperature; (b) above the transition temperature.

Unfortunately, the transition temperature of pure VO_2 is too high to make this a practical solution to heat transfer problems. A great number of studies have been carried out to remedy this drawback. In particular doping the V^{4+} site with other cations, especially W^{6+}, Mo^{6+}, Nb^{5+}, and doping the O^{2-} sites with F^- all reduce the transition temperature significantly. The system $W_xV_{1-x}O_2$ is illustrative. The addition of W^{6+} causes some of the V^{4+} ions to become V^{3+} ions to maintain charge neutrality in the crystal. For this, each W^{6+} must be balanced by two V^{3+} ions to give an overall formula $W_x{}^{6+}V_{1-3x}{}^{4+}V_{2x}{}^{3+}O_2$. The effect of the added dopant is to disrupt the $V^{4+} - V^{4+}$ pairs and hence to strengthen the d band. The transition temperature drops by approximately 20 °C per at.% added W and a metallic state forms at approximately 24 °C for $W_{0.025}V_{0.975}O_2$. This is now at a suitable temperature to make a passive window coating that can help with temperature control.

References

Agrawal, D.K. (1994). *J. Mater. Edu.***16**: 139.

Rakshit, S.K., Parida, S.C., Dash, S. et al. (2007). Thermodynamic studies on $SrFe_{12}O_{19}$(s), $SrFe_2O_4$(s), $Sr_2Fe_2O_5$(s) and $Sr_3Fe_2O_6$(s). *J. Solid State Chem.***180**: 523.

Further Reading

General:

Rowe, D.M. (ed.) (1995). *CRC Handbook of Thermoelectrics*. Boca Raton, FL: CRC Press.

An Interactive Demonstration of the Debye Formula for the Heat Capacity of Solids Is:

Heat capacity of solids in the Debye approximation: http://demonstrations.wolfram.com/HeatCapacityOfSolidsinTheDebyeApproximation.

Thermal Conductivity:

Materials Research Society Bulletin, 26 June (2001) contains a series of articles on thermal conductivity including G. P. Srivastava, Theory of thermal conduction in nonmetals, p 445.

Chen, K., Song, B., Ravichandran, N.K. et al. (2020). Ultrahigh thermal conductivity in isotope-enriched cubic boron nitride. *Science***367**: 555.

Losego, M.D., Grady, M.E., Sottos, N.R. et al. (2012). Effects of chemical bonding on heat transport across interfaces. *Nat. Mater.***11**: 502–506. and references therein.

Zhou, W.-X., Cheng, Y., Chen, K.-Q. et al. (2020). Thermal conductivity of amorphous materials. *Adv. Funct. Mater.***30**: 1903829.

Negative and Zero Thermal Expansion:

Adak, S., Daemen, L.L., Hartl, M. et al. (2011). Thermal expansion in 3D-metal Prussian Blue analogues – a survey study. *J. Solid State Chem.* 184: 2854–2861.

Sleight, A.W. (1998). Compounds that contract on heating. *Inorg. Chem.***37**: 2854–2860.

Song, X. et al. (2011). Adjusable zero thermal expansion in antiperovskite manganese nitride. *Adv. Mater.***23**: 690–4694.

The Magnetocaloric Effect in Alloys:

Liu, J., Gottschall, T., Skokov, K.P. et al. (2012). Giant magnetocaloric effect driven by structural transitions. *Nat. Mater.***11**: 620–626. and references therein.

Thermoelectric Materials:

Balusu, B. and Walker, D.G. (2008). Review of electronic transport models for thermoelectric materials. *Superlattices Microstruct.***44**: 1.

Yu, Y., Cagnoni, M., Cojocaru-Mirédin, O., and Wuttig, M. (2020). Chalcogenide thermoelectrics empowered by an unconventional bonding mechanism. *Adv. Funct. Mater.***30**: 19004862.

Problems and Exercises

Quick Quiz

13.1 At room temperature the heat capacity of a solid at constant volume, C_v:

a. Is greater than the heat capacity at constant pressure, C_p .

b. Is less than the heat capacity at constant pressure, C_p .

c. Is approximately equal to the heat capacity at constant pressure, C_p.

13.2 The low temperature heat capacity of a solid is proportional to:

a. T^4.

b. T^3.

c. T^2.

13.3 The high temperature heat capacity of a solid is approximately:

a. 2.5 J K^{-1} mol^{-1}.

b. 25 J K^{-1} mol^{-1}.

c. 250 J K^{-1} mol^{-1}.

13.4 The thermal conductivity of a solid is mainly due to:

a. Phonons.

b. Defects.

c. Electrons.

13.5 Alloys generally have:

a. A higher thermal conductivity than the parent metals.

b. A lower thermal conductivity than the parent metals.

c. About the average thermal conductivity of the parent metals.

13.6 Compared to a poorly crystalline polymer, a highly crystalline polymer has a:

a. Lower thermal conductivity.

b. Higher thermal conductivity.

c. About the same thermal conductivity.

13.7 The mean volume expansivity of liquid mercury is $18.2 \times 10^{-5}\ K^{-1}$. The mean linear expansivity is approximately:

a. $54.6 \times 10^{-5}\ K^{-1}$.

b. $6.06 \times 10^{-5}\ K^{-1}$.

c. $2.63 \times 10^{-5}\ K^{-1}$.

13.8 Solids linked with strong chemical bonds have:

a. A lower thermal expansivity than weakly bonded solids.

b. A greater thermal expansivity than weakly bonded solids.

c. About the same thermal expansivity as weakly bonded solids.

13.9 A current flow induced in a circuit made of two different conductors by holding the junctions between the materials at different temperatures is called:

a. The Peltier effect.

b. The Thomson effect.

c. The Seebeck effect.

13.10 A temperature difference between the two junctions in a circuit made of two different conductors induced by a current flow is called:

a. The Seebeck effect.

b. The Peltier effect.

c. The Thomson effect.

13.11 A thermocouple makes use of:

a. The Seebeck effect.

b. The Peltier effect.

c. The Thomson effect.

13.12 Thermoelectric refrigerators utilise:

a. The Thomson effect.

b. The Seebeck effect.

c. The Peltier effect.

13.13 The thermochromic behaviour of VO_2 is NOT due to:

a. A metal-insulator transition.

b. A crystal structure change.

c. Point defects.

Calculations and Questions

13.1 How much energy is needed to raise the temperature of 2.5 mol of alumina from 0 to 120 °C, taking the specific heat capacity of alumina, 0.907 J K^{-1} g^{-1}, to be independent of temperature?

13.2 How much energy has to be extracted to lower the temperature of 15 g tungsten metal from 2500 to 1500 K? The molar heat capacity at 2000 K, 32.26 J K^{-1} mol^{-1}, can be considered to apply across the whole of this temperature range.

13.3 The specific heat of silicon at 50 K is 2.162 J K^{-1} mol^{-1}. Estimate the value at the boiling point of neon, 27.07 K.

13.4 Calculate the specific heat of silicon at the boiling point of neon using the fact that the Debye temperature of silicon is 645 K. Compare this to the result in the previous question.

13.5 A Styrofoam box is used to transport 10 kg meat at an average external temperature of 25 °C. The box is 50 × 30 × 20 cm, and the foam thickness is 5 cm. How long will it take the contents to increase in temperature from the initial 0 to 5 °C, assuming that the heat capacity of meat is 4.22 J K^{-1} g^{-1} and the thermal conductivity of Styrofoam is 0.035 W m^{-1} K^{-1}.

13.6 Show that the equation for the heat transfer across a number of slabs of material with the same surface area, A, is:

$$dQ/dt = A\Delta T \Big/ \sum_{i=1}^{n} \Delta x_i/\kappa_i$$

where ΔT is the total temperature drop ($T_1 - T_2$), and Δx_i is the thickness of slab i, of thermal conductivity κ_i.

13.7 A cooking pot of 15 cm diameter and a base of 3 mm copper and 1 mm stainless steel, contains 2 l of water at 20 °C.

a. What is the initial rate of heat transfer if the hot plate is at 150 °C?

b. a solid copper bottom 4 mm thick;

c. a solid stainless steel bottom, 4 mm thick.

13.8 A window of area 2 × 1.30 m is glazed with a single sheet of glass 5 mm thick.

a. What is the heat loss per hour from a room at 25 °C when the outside temperature is 4.5 °C?

b. If the area is double glazed with two such sheets, separated by an air gap of 1 cm, what will the heat loss per hour be?

13.9 A gap is left between rails in a railway so that the rails can expand without causing track buckling at high temperatures. What gap needs to be left between 10 m rail lengths installed at 10 °C if the ground temperature might reach 50 °C? The expansivity of steel is 10.7×10^{-6} K^{-1}.

13.10 A volume of mercury of 10^{-6} m^3 at 20 °C is contained in glass bulb, with expansion taken up by the mercury moving into a capillary 0.5 mm diameter, similar to a mercury thermometer. The aim is to allow the mercury to expand and complete an electrical circuit and activate a cooling device. If the circuit contact is 5 mm above the mercury level at 20 °C, what temperature will activate the device? The mean volume expansivity of liquid mercury is 18.2×10^{-5} K^{-1}.

13.11

a. Estimate the mean coefficient of linear expansion for the **a**- and **c**-axes of $CaZr_4P_6O_{24}$, using the data in Figure 13.9.

b. What is the mean coefficient of volume expansion of this material?

13.12 The voltage generated across a Pt–Au thermocouple when the cold junction is at 0 °C and the hot junction is at 100 °C is +780 μV.

a. Calculate the average value of Σ_{AuPt} over the temperature range.

b. Assuming that the absolute value of the Seebeck coefficient for Pt, σ_S is $-6.95\ \mu V\ K^{-1}$ over the whole of this temperature range, estimate the average value of the absolute Seebeck coefficient for Au.

13.13 The voltage generated across a Pt–Al thermocouple when the cold junction is at 0 °C and the hot junction is at 100 °C is +420 μV.

a. Calculate the average value of Σ_{AlPt} over the temperature range.

b. Assuming that the absolute value of the Seebeck coefficient for Pt, σ_S is $-6.95\ \mu V\ K^{-1}$ over the whole of this temperature range, estimate the average value of the absolute Seebeck coefficient for Al.

13.14 The absolute Seebeck coefficient for lead, σ_{Pb}, at 300 K is $-1.047\ \mu K^{-1}$, and for platinum at 300 K is $\sigma_{Pt} = -5.05\ \mu V\ K^{-1}$. Estimate the voltage generated by a thermocouple made from these metals when the cold junction is at 0 °C and the hot junction at 300 K.

Note

1 Note that the form of the equation for holes is often written:

$\alpha = -(k/e) \ln[c/(1 - c)]$

which is identical to that given as $\ln x = \pm \ln (1/x)$

Part IV
Nuclear Properties of Solids

14
Radioactivity and Nuclear Reactions

- What is radioactivity?
- How does carbon dating work?
- What produces energy in a nuclear power station?

The properties described earlier in this book are a function of the outer electrons on the atoms making up the solid and the atomic nuclei can be regarded as simply providing mass. However, some of the heaviest atoms have unstable nuclei, which spontaneously disintegrate. The elements that show this phenomenon are said to be *radioactive*; a term coined by Marie Curie to describe compounds that constantly emit 'penetrating radiation'. Since the discovery of natural radioactivity, many lighter, normally stable nuclei have also been made radioactive and these find uses in industry and medicine. Solids containing radioactive atoms are important in many areas, varying from medicine to dating of archaeological and geological samples to power generation.

14.1 Radioactivity

14.1.1 Naturally Occurring Radioactive Elements

Towards the end of the nineteenth century, only the two heaviest naturally occurring elements known as that epoch, thorium, and uranium were known to be radioactive. The earliest studies indicated that the 'penetrating radiation' characteristic of these elements was found to be composed of three components:

> *Alpha (α) particles*: subsequently found to be helium nuclei. These have a very low penetrating power and can be stopped by thin card.
>
> *Beta (β) particles*: subsequently found to be electrons. These have a medium penetrating power and can traverse about 30 cm of air.
>
> *Gamma (γ) rays*: subsequently found to be high energy photons. These have a high penetrating power and can pass through 1 m of concrete.

Very soon after the radioactivity of thorium and uranium had been discovered it was found that pure samples of both of these elements were only very weakly radioactive. However, such samples became more and more radioactive with time until they reached a steady level identical to that in the original samples before purification. This suggested that the uranium or thorium atoms were transforming or *decaying* into other radioactive *daughter elements* and that hitherto undiscovered elements might exist. The search for the radioactive products of uranium decay by Marie and Pierre Curie led to the characterisation of two new elements, which were named polonium (Po) and radium (Ra). Both are far more radioactive than uranium and decay so rapidly that no ore deposits are formed. They exist only because they are produced constantly from naturally occurring uranium.

It is now clear that other radioactive nuclei are found in nature. Some of these are the result of the impact of high energy radiation on the nucleus of a non-radioactive atom. Radioactive carbon-14, used in radiocarbon dating, is formed in constantly replenished trace amounts by the interaction of high energy cosmic rays with atoms in the upper atmosphere. Others, such as radon (Rn), a radioactive gas associated with granite, result from the radioactive decay of natural uranium ores.

14.1.2 Isotopes and Nuclides

For the present purposes the nucleus of an atom can be considered to be made up of *protons* and *neutrons*, collectively known as *nucleons*. A proton has a charge of +1 and the number of protons in a nucleus determines the *proton number* (also called the *atomic number*), *Z*, of the atom. The neutron bears no charge but has a similar mass to that of the proton. The number of electrons on a neutral atom is also equal to *Z*. Atoms with the same value of *Z* are *chemically identical*. The total number of nucleons in the nucleus is called the *nucleon number* (also called the *mass number*), *A*. The number of neutrons in a nucleus need not be the same as the number of protons. Atoms with the same value of *Z* but different numbers of neutrons are known as *isotopes* of the element in question. Isotopes are represented by the symbol ${}^{A}_{Z}\mathrm{X}$. For example, the radioactive carbon isotope used in radiocarbon dating has a symbol ${}^{14}_{6}\mathrm{C}$. Because the name of the element or its symbol already contains information about the proton (atomic) number of the atom, this is often omitted. Thus, ${}^{14}_{6}\mathrm{C}$ is often written as ^{14}C or carbon-14. An isotope that is radioactive is called a *radioisotope*. A *nuclide* is any atomic species, whether radioactive or not, that has a specified nucleon number and proton number. Hence ${}^{9}_{4}\mathrm{Be}$ is a nuclide.

Nuclear reactions involve interactions between the relatively massive nuclides and a variety of small particles. For the purposes of balancing nuclear equations, these latter particles are given an atomic number (i.e. a charge) and a mass number (Table 14.1).

Table 14.1 Nuclides and particles in nuclear reactions.

Name	Nuclide	Shorthand
Proton	${}_{1}^{1}\mathrm{H}$	p
Deuteron	${}_{1}^{2}\mathrm{H}$	d
Hydrogen-3 (triton)	${}_{1}^{3}\mathrm{H}$	t
Alpha particle	${}_{2}^{4}\mathrm{He}$	α
Neutron	${}_{0}^{1}\mathrm{n}$	n
Electron (beta particle)	${}_{-1}^{0}\mathrm{e}$	β, β^-, e, e^-
Positron	${}_{+1}^{0}\mathrm{e}$	β^+, e^+
Gamma ray	—	γ
Electron neutrino	—	ν_e
Electron antineutrino		$\bar{\nu}_e$

14.1.3 Nuclear Equations

In order to describe radioactive transformations, nuclear equations are needed. These are very similar to chemical equations, except that the nucleon numbers and proton numbers of each reactant must also be specified, that is, the reactions are written with nuclides. Naturally, as we are dealing with nuclei alone, the ionic charges given in chemical equations are no longer relevant. Thus, an α-particle is correctly written ${}_{2}^{4}\mathrm{He}$ not ${}_{2}^{4}\mathrm{He}^{2+}$ and a uranium nucleus is written ${}_{92}^{238}\mathrm{U}$ not ${}_{92}^{238}\mathrm{U}^{92+}$.

A typical nuclear equation, in this example representing the decay of uranium-238, which is accompanied by the emission of an alpha particle (a process called *alpha decay*), is:

$$ {}_{92}^{238}\mathrm{U} \rightarrow {}_{90}^{234}\mathrm{Th} + {}_{2}^{4}\mathrm{He} $$

The reactant is uranium-238 (${}_{92}^{238}\mathrm{U}$), a naturally occurring radioisotope that emits α-particles, ${}_{2}^{4}\mathrm{He}$. The change in the proton number, from 92 to 90, specifies that a different chemical element has been produced; a fact confirmed by the chemical symbol Th. The other product, the alpha

particle, is specified in a similar way. As it is a helium nucleus the chemical symbol is He. The mass number, 4, and the atomic number, 2, are added to the symbol to complete matters. The equation can also be written in an abbreviated form, as long as this does not cause confusion.

$$^{238}\mathrm{U} \rightarrow {}^{234}\mathrm{Th} + \alpha$$

A similar reaction, but involving the emission of an electron, called *beta decay*, is:

$$^{234}_{90}\mathrm{Th} \rightarrow {}^{234}_{91}\mathrm{Pa} + {}^{0}_{-1}\mathrm{e} + \overline{\upsilon}_{\mathrm{e}}$$

or $^{324}\mathrm{Th} \rightarrow {}^{234}\mathrm{Pa} + \mathrm{e}^{-} + \overline{\nu}_{\mathrm{e}}$ or $^{324}\mathrm{Th} \rightarrow {}^{234}\mathrm{Pa} + \beta^{-} + \overline{\nu}_{\mathrm{e}}$

In this decay, radioactive thorium-234 emits an electron (β-particle). There is no change in the nucleon number, but the proton number has increased by one in the transformation. Note that the electron charge, −, may be specified to avoid confusion with positron emission, described in the following text. In addition, an electron antineutrino, $\overline{\upsilon}_{\mathrm{e}}$, needed to conserve energy and spin, is also emitted during normal β decay. Antineutrinos do not give rise to any chemical or physical effects under normal circumstances and although they are needed for exact nuclear physics accounting, they are generally omitted in equations that do not require such information.

There are a number of rules that must be followed in writing nuclear equations.

1. The sum of the proton numbers (i.e. charges) on the left of the equation must equal the sum of the proton numbers on the right.
2. The sum of the nucleon numbers (i.e. mass) on the left must equal the sum of the nucleon numbers on the right.
3. When a radioactive element emits an alpha particle, the daughter element has a nucleon number four less that the parent and a proton number of two less.
4. When a radioactive element emits a beta particle, the daughter element has a nucleon number the same as the parent and a proton number one greater than that of the parent.
5. Electrons are not present in the nucleus, and are produced (at least schematically) via ‘decomposition’ of neutrons, following the

nuclear reaction:

$$^{1}_{0}\mathrm{n} \rightarrow {}^{1}_{1}\mathrm{H} + {}^{0}_{-1}\mathrm{e}$$

$$\text{or} \quad \mathrm{n} \rightarrow \mathrm{p} + \mathrm{e}$$

As in all β decay transformations an electron antineutrino is also emitted but is not usually included in the equation.

Often nuclear equations involving the interaction of a nucleus and an energetic particle to give a different nucleus together with product particles are written in a shorthand form:

Initial nuclide (incoming particle, outgoing particle) final nuclide

The reaction between uranium-238 and a deuteron, the nucleus of deuterium or heavy hydrogen, which produces neptunium-238 and two neutrons, is:

$$^{238}_{92}\mathrm{U} + {}^{2}_{1}\mathrm{H} \rightarrow {}^{238}_{93}\mathrm{Np} + 2{}^{1}_{0}\mathrm{n}$$

or ^{238}U(d, 2n)^{238}Np

14.1.4 Radioactive Series

The series of transformations that take place as a radioactive element changes into successive daughter elements is called a *radioactive series*. Four different radioactive series of historical and technological significance have been described, three of which occur naturally. These all halt when a stable (non-radioactive) element has been produced. The complete form that these series take is rather complex. Here we present only the main reaction chains, involving α-decay, in which an α particle is ejected from the nucleus, or β-decay, in which a β particle is ejected from the nucleus.

14.1.4.1 The Uranium Series

The parent nuclide is the naturally occurring isotope uranium-238 and the series ends with the stable nuclide lead-206 (Figure 14.1). The parent nuclide is shown at the top right of the figure. An α-decay is represented by a diagonal displacement of two proton number units to the left. A β-

decay is shown as a horizontal displacement of one proton number unit to the right. The nucleon numbers, A, of the members of the series conform to the formula $4x + 2$ where x takes values between 51 for lead-206 and 59 for uranium-238.

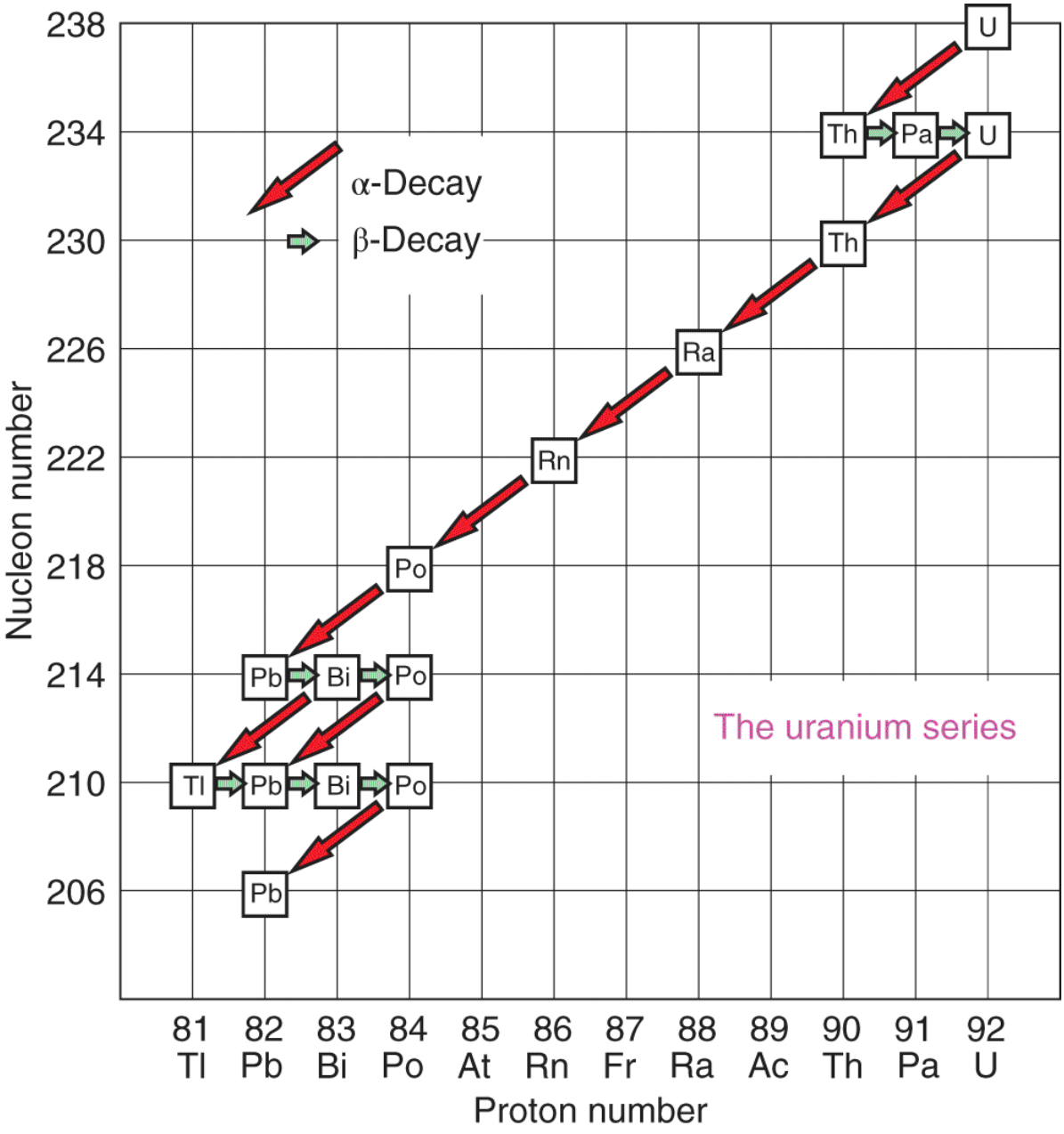


Figure 14.1 The uranium-238 decay series.

Of the atomic species taking part in this cascade, only uranium-238 has a long lifetime. Apart from the stable, non-radioactive end product lead-206, all other species decay rapidly. These include two other radioactive

lead nuclides, Pb-210 and Pb-214, which have a transitory existence. Some nuclides have two alternative ways of decay and so the path occasionally branches. For example, Bi-214 can form Po-214 by a β-decay or Tl-210 by an α-decay. All of the daughter elements in the series are metals with the exception of the noble gas radon, which is a naturally occurring radioactive gas. It is produced by the α-decay of radium-226:

$$^{226}_{88}\text{Ra} \rightarrow {}^{222}_{86}\text{Rn} + {}^{4}_{2}\text{He}$$

The radon isotope formed as part of the uranium series is an α-emitter that decays according to the following reaction:

$$^{222}_{86}\text{Rn} \rightarrow {}^{218}_{84}\text{Po} + {}^{4}_{2}\text{He}$$

Different isotopes of radon form in the other radioactive series described in the following text.

14.1.4.2 The Thorium Series

This series starts with naturally occurring thorium-232 and ends with the stable isotope lead-208 (Figure 14.2). The daughter elements all have mass numbers divisible by four, and the series formula is $4x + 0$. The first reaction in the series is:

$$^{232}_{90}\text{Th} \rightarrow {}^{228}_{88}\text{Ra} + {}^{4}_{2}\text{He}$$

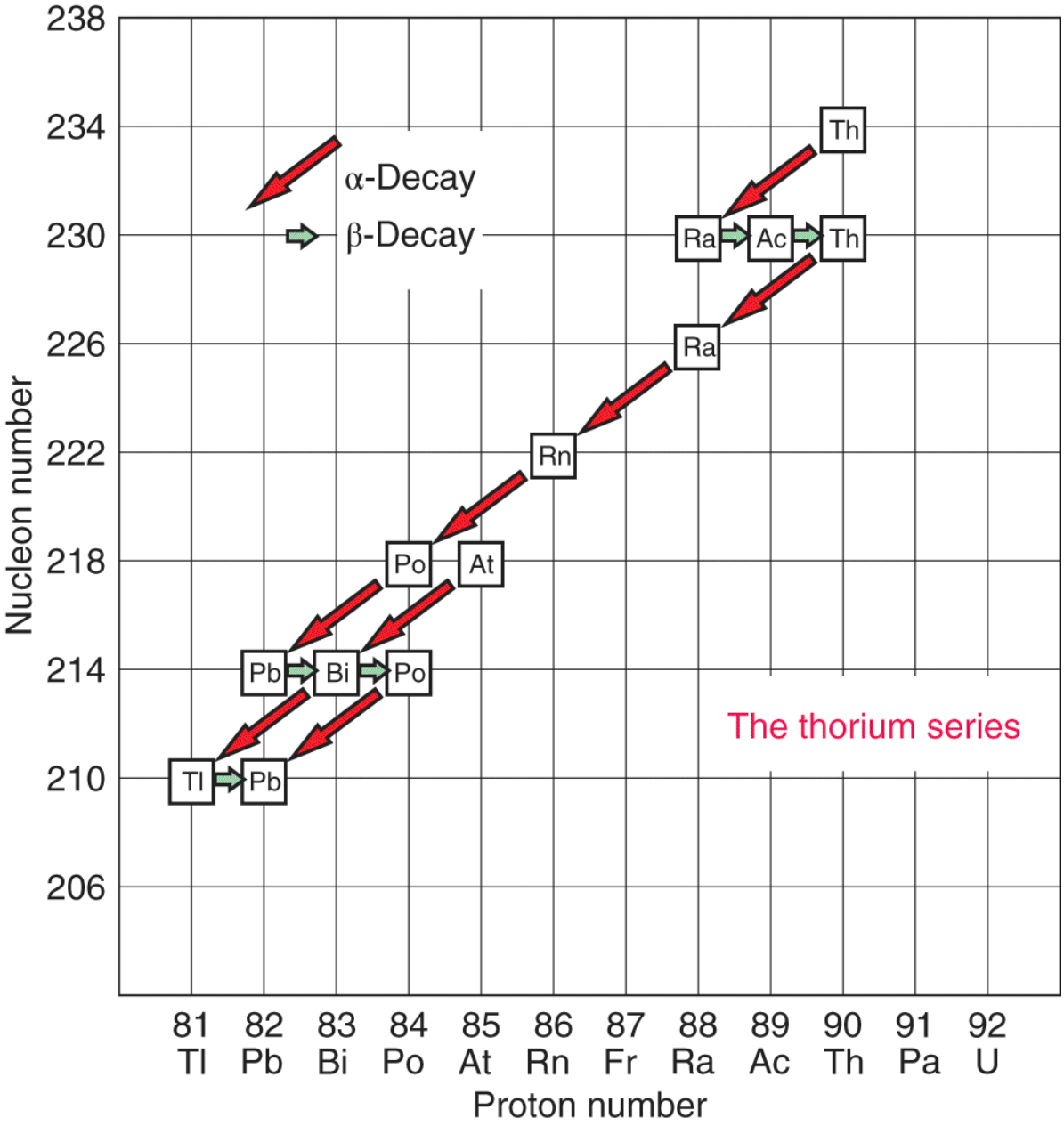


Figure 14.2 The thorium-232 decay series.

14.1.4.3 The Actinium Series

This series was so called because it was originally thought that actinium-227 was the parent element. However, actinium-227 decays too quickly for the series to persist for any length of time, and eventually uranium-235 was proved to be the true parent element (Figure 14.3). Uranium-235 is the less common naturally occurring isotope of uranium. The

series ends with yet another stable lead isotope, lead-207, and the series formula is $4x + 3$.

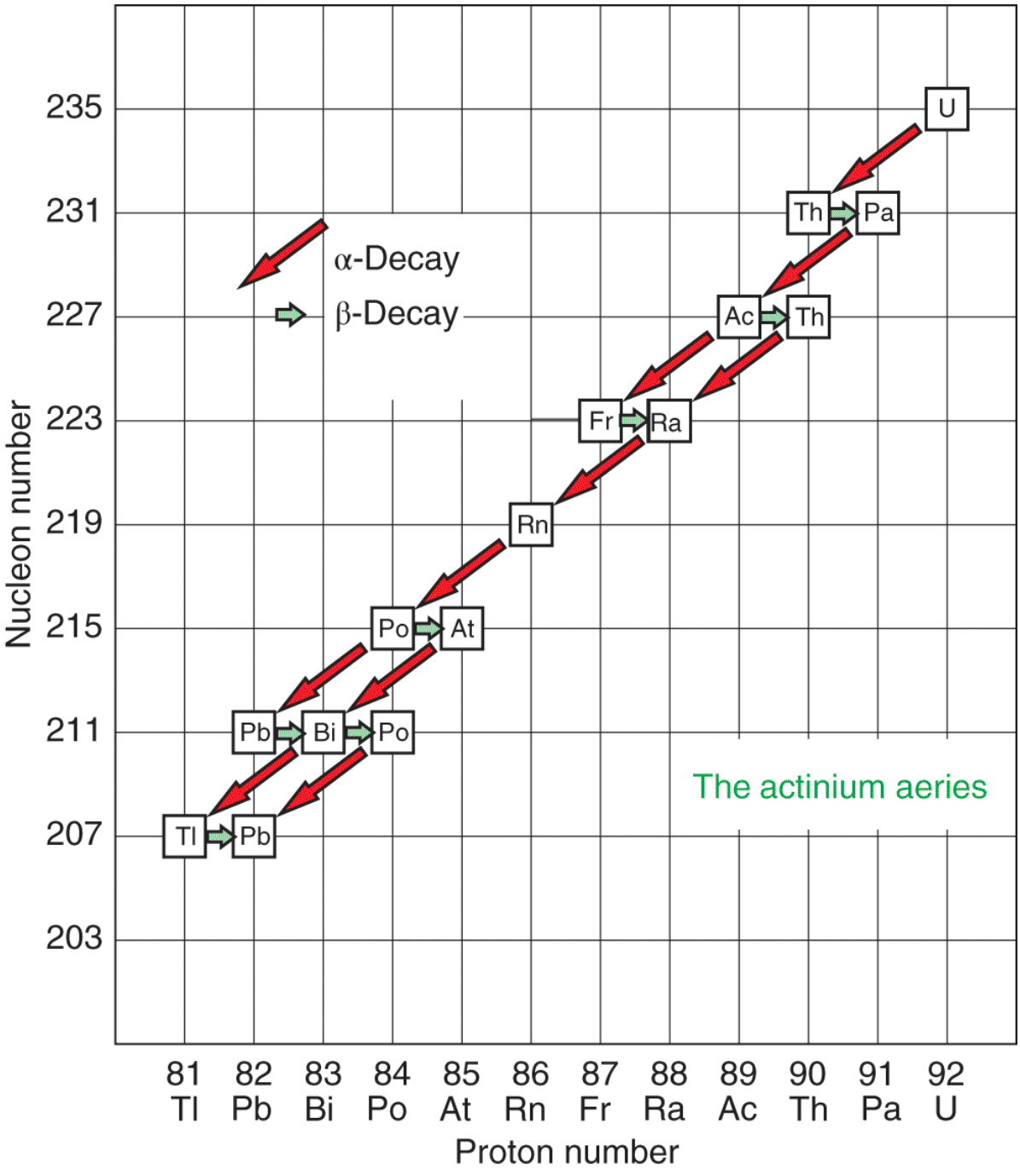


Figure 14.3 The actinium-227 (uranium-235) decay series.

14.1.4.4 The Neptunium/Plutonium Series

The existence of the three naturally occurring series characterised by the formulae $4x$, $4x + 2$ and $4x + 3$, leads to the expectation of a fourth series, with isotopic weights given by $4x + 1$. No isotope with a sufficiently long lifetime is found in nature for this series to exist outside of the laboratory. Eventually the discovery of the transuranic element neptunium enabled much of the series to be constructed (Figure 14.4). As can be seen, the parent nuclide is not neptunium-237, which is the longest-lived radioactive isotope in the series, but plutonium-241. This series differs from the three naturally occurring series by ending, not with an isotope of lead, but with bismuth-209.

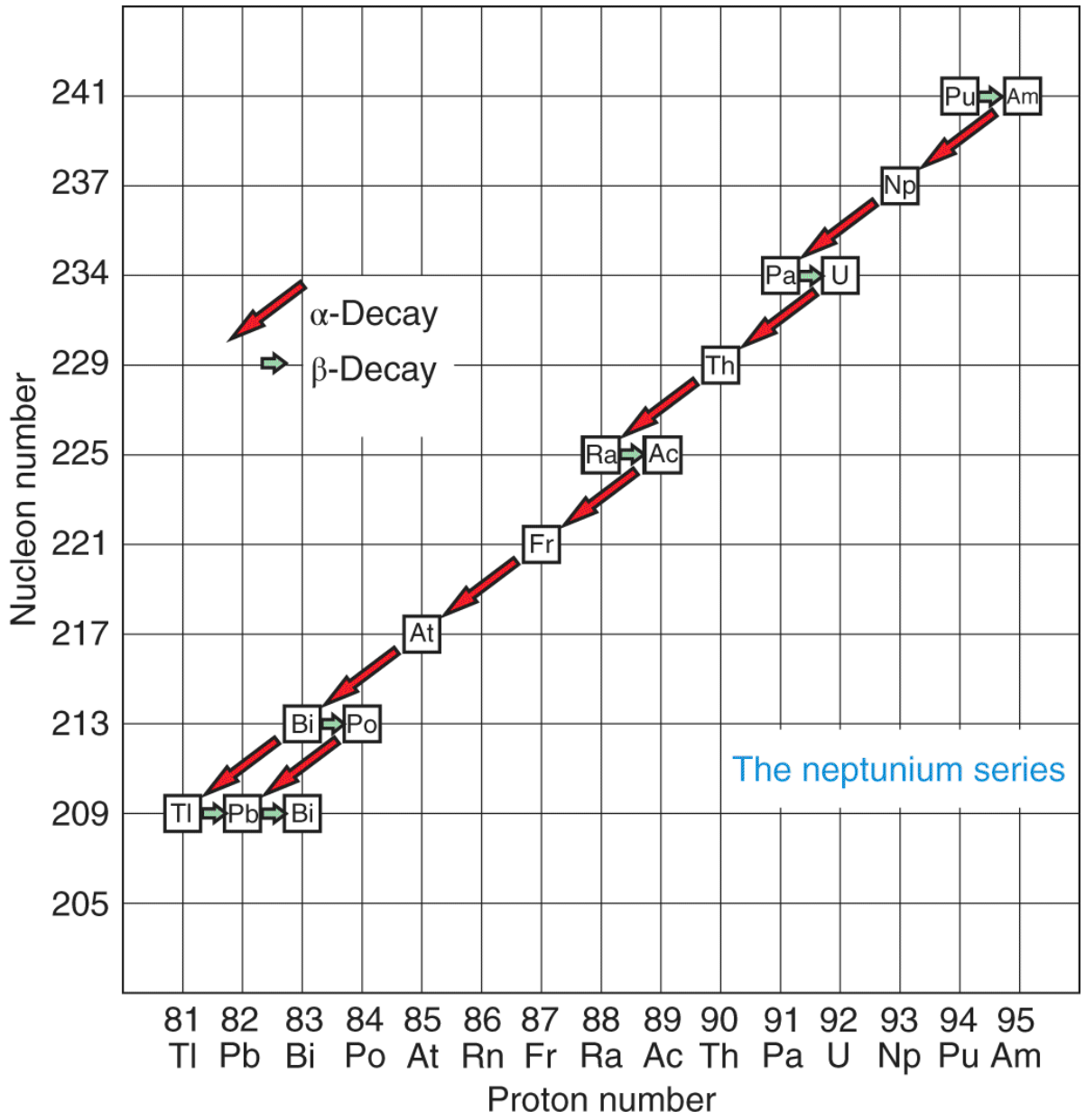


Figure 14.4 The plutonium-241 decay series.

14.1.5 Nuclear Stability

Although many radioactive isotopes have been discovered or prepared, most known nuclei are stable. It is not possible to explain all aspects of nuclear stability theoretically, but empirical observations allow one to guess the likely stability of any particular nuclide. A useful guide is a graph of the proton number, Z, vs. the nucleon number A, or the number of neutrons present, $Z–A$ (Figure 14.5). Stable nuclei all cluster into a narrow strip called the *band of stability*. Nuclei above the band of

stability tend to emit positrons or to incorporate an outer electron into the nucleus (called *electron capture*), so as to move the product nucleus nearer to the band of stability. Nuclei below the band of stability tend to emit β-particles for the same reason. Thus carbon-14 has a proton to neutron ratio of 6/8, i.e. 0.75. This is well below the band of stability, and carbon-14 would be expected to decay by β-emission, which is in accord with experimental evidence. All nuclei with more than 82 protons are radioactive.

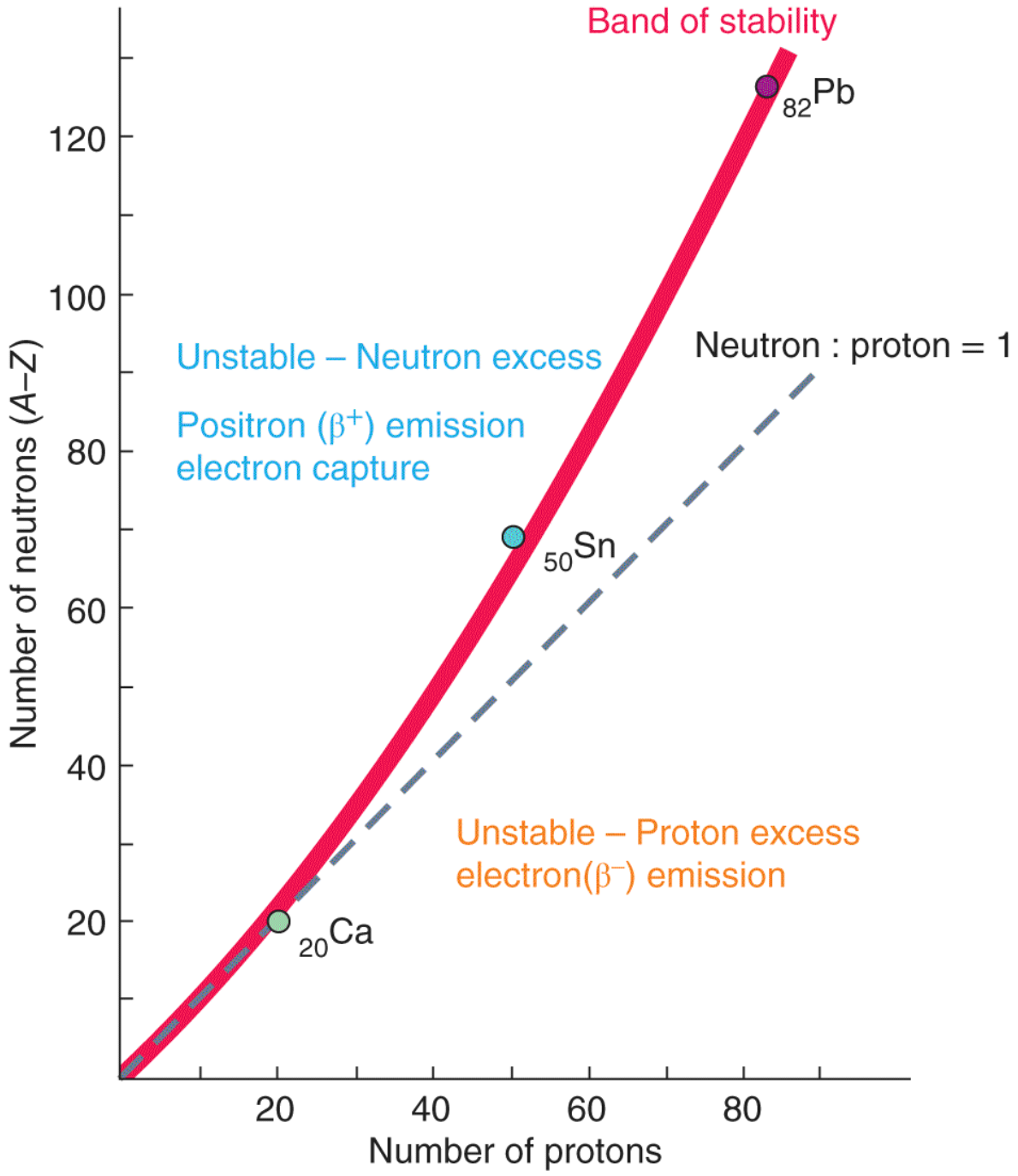


Figure 14.5 The band of stability of the elements.

Theories of nuclear structure predict that certain numbers of protons or neutrons give rise to enhanced stability, while other numbers tend to be associated with reduced stability. Only a small number of stable nuclei have either odd numbers of protons or neutrons. For example, there are only four stable nuclei that contain an odd number of both protons and neutrons, viz. ${}^{2}_{1}\mathrm{H}$, ${}^{6}_{3}\mathrm{Li}$, ${}^{10}_{5}\mathrm{B}$, and ${}^{14}_{7}\mathrm{N}$. Most stable nuclei contain

even numbers of protons or neutrons. Nuclei that contain 2, 8, 20, 28, 50, 82, and 126 protons or neutrons are particularly stable, and these numbers are called *magic numbers*. Nuclei in which both the proton number and neutron number are magic are termed *double magic* and are generally particularly stable. The best known of these are ${}^{4}_{2}\mathrm{He}$, ${}^{16}_{8}\mathrm{O}$, ${}^{40}_{20}\mathrm{Ca}$ and ${}^{208}_{82}\mathrm{Pb}$.

14.2 Artificial Radioactive Atoms

14.2.1 Heavy Elements

Energetic particles, either from naturally radioactive materials, nuclear explosions, or particle accelerators, can be used to bring about atomic transmutations and so form new nuclides. In this way, the number of known elements has been extended above uranium, the heaviest naturally occurring element, with proton number 92. In addition, new elements have been discovered that fill the gaps in the Periodic Table between bismuth and thorium. All these heavy elements are radioactive, and many have very short lifetimes.

Period 6 of the Periodic Table, beginning with the well-known metals caesium, Cs, and barium, Ba, was incomplete near the end of the nineteenth century. This was remedied by the discovery of polonium, Po, in 1898 and radon, Rn, in 1900 although the iodine analogue astatine, At, was not found until 1940. Period 7 of the Periodic Table starts with the alkali metal francium, Fr, discovered in 1939, which is followed by the alkaline earth radium, Ra, discovered in 1898. The *actinoids* are a series of radioactive elements in which the 5f electron orbitals are occupied (Table 14.2) and are analogous to the *lanthanoids*, in which the 4f orbitals are occupied. Neither actinium nor lawrencium behave as genuine actinoids and are sometimes not included in the group but included with the transition metals that follow. The transition metals, in which the 6d orbitals are filled, continue beyond the actinoids, as far as copernicium (Table 14.2). Beyond copernicium the 7p orbitals are filled and all elements up to the noble gas oganesson, with proton number 118 have been reported. Note that in many of these heavier elements, the actual outer electron configurations listed in Table 14.2 are not well defined and uncertainty about this and the corresponding chemical parallels with lighter elements remains.

Table 14.2 Periodic Table group seven elements.

Name	Symbol	Proton number Z	Nucleon number A[a]	Electron configuration[b]	Half-life[c]
s block					
Francium	Fr	87	223	[Rn] $7s^1$	28.1 min
Radium	Ra	88	226	[Rn] $7s^2$	1600 yr
Actinoids					
Actinium	Ac	89	227	[Rn] $6d^1\ 7s^2$	21.8 yr
Thorium	Th	90	232	[Rn] $6d^2\ 7s^2$	1.41×10^{10} yr
Protactinium	Pa	91	231	[Rn] $5f^2\ 6d^1\ 7s^2$	3.25×10^4 yr
Uranium	U	92	238	[Rn] $5f^3\ 6d^1\ 7s^2$	4.47×10^9 yr
Neptunium	Np	93	237	[Rn] $5f^4\ 6d^1\ 7s^2$	2.14×10^6 yr
Plutonium	Pu	94	244	[Rn] $5f^6\ 7s^2$	8.1×10^7 yr
Americium	Am	95	243	[Rn] $5f^7\ 7s^2$	7.38×10^3 yr
Curium	Cm	96	247	[Rn] $5f^7\ 6d^1\ 7s^2$	1.56×10^7 yr
Berkelium	Bk	97	247	[Rn] $5f^9\ 7s^2$	1.38×10^3 yr
Californium	Cf	98	251	[Rn] $5f^{10}\ 7s^2$	898 yr
Einsteinium	Es	99	252	[Rn] $5f^{11}\ 7s^2$	1.29 yr
Fermium	Fm	100	257	[Rn] $5f^{12}\ 7s^2$	100.5 d
Mendelevium	Md	101	258	[Rn] $5f^{13}\ 7s^2$	51.5 d

Name	Symbol	Proton number Z	Nucleon number A[a]	Electron configuration[b]	Half-life[c]
Nobelium	No	102	259	[Rn] $5f^{14}$ $7s^2$	58 min
Lawrencium	Lr	103	262	[Rn] $5f^{14}$ $7s^2$ $7p^1$	3 min
Transition metals					
Rutherfordium	Rf	104	263	[Rn] $5f^{14}$ $6d^2$ $7s^2$	10 min
Dubnium	Db	105	262	[Rn] $5f^{14}$ $6d^3$ $7s^2$	34 s
Seaborgium	Sg	106	266	[Rn] $5f^{14}$ $6d^4$ $7s^2$	21 s
Bohrium	Bh	107	272	[Rn] $5f^{14}$ $6d^5$ $7s^2$	9.8 s
Hassium	Hs	108	277	[Rn] $5f^{14}$ $6d^6$ $7s^2$	11 min
Meitnerium	Mt	109	276	[Rn] $5f^{14}$ $6d^7$ $7s^2$	0.72 s
Darmstadtium	Ds	110	280	[Rn] $5f^{14}$ $6d^9$ $7s^1$	7.6 s
Roentgenium	Rg	111	280	[Rn] $5f^{14}$ $6d^{10}$ $7s^1$	3.6 s
Copernicium	Cn	112	285	[Rn] $5f^{14}$ $6d^{10}$ $7s^2$	11 min
p block					
Nihonium	Nh	113	284	([Rn] $5f^{14}$ $6d^{10}$ $7s^2$ $7p^1$)	~10 s
Flerovium	Fl	114	289	([Rn] $5f^{14}$ $6d^{10}$ $7s^2$ $7p^2$)	~21 s
Moscovium	Mc	115	288	([Rn] $5f^{14}$ $6d^{10}$ $7s^2$ $7p^3$)	~0.28 s
Livermorium	Lv	116	292	([Rn] $5f^{14}$ $6d^{10}$ $7s^2$ $7p^4$)	~0.053 s
Tennessine	Ts	117	294	([Rn] $5f^{14}$ $6d^{10}$ $7s^2$ $7p^5$)	~0.01 s
Oganesson	Og	118	294	([Rn] $5f^{14}$ $6d^{10}$ $7s^2$ $7p^6$)	~0.002 s

[a] Of the most stable isotope.

[b] The configuration is that of the ground state of a gaseous atom. Note that in compounds the electron configuration may be different from that given. Values in parenthesis (xxx) represent configurations that are tentative and based upon the configurations of similar atoms in the same period but lower groups of the Periodic Table. [Rn] represents the configuration of the noble gas radon.

[c] Of most stable isotope.

The search for new heavy elements is an area of continuing nuclear study. The existence of these postulated materials will give important insights into nuclear stability and the chemical underpinning of the Periodic Table. Currently attempts employ a stream of energetic titanium ions fired at a thin foil of californium. The aim is to try to reach the element corresponding to the next magic neutron number of 184 coupled with the proton magic number 114 (but theory is unclear here and 120 or 126 have also been suggested in this context). Only experiment will resolve this uncertainty. The hope is that the nuclides around this magic number might all have enhanced stability, giving rise to the idea of an *island of stability*. If this proves to be the case, nuclides within the island may be stable enough for their physical and chemical properties to be investigated experimentally.

14.2.2 Artificial Radioactivity in Light Elements

When stable nuclei are bombarded by sufficiently energetic radiation, they can transmute into *artificially radioactive* species. Many of the elements with atomic numbers lighter than 82 (bismuth, Bi), which are not normally radioactive, have now been made in radioactive forms.

The first nuclear transmutation described as such was interpreted by Rutherford in 1919. In this reaction energetic α-particles collide with nitrogen atoms to produce an isotope of oxygen and a proton:

$$^{14}_{7}\mathrm{N} + {}^{4}_{2}\mathrm{He} \rightarrow {}^{17}_{8}\mathrm{O} + {}^{1}_{1}\mathrm{H}$$

The reaction that Chadwick used to establish the existence of the neutron, in 1932, was similar, and involved the bombardment of beryllium with energetic α-particles:

$$^{9}_{4}\mathrm{Be} + {}^{4}_{2}\mathrm{He} \rightarrow {}^{12}_{6}\mathrm{C} + {}^{1}_{0}\mathrm{n}$$

The first artificial radioisotope to be produced was made by Irene Curie and Joliot in 1934. The reaction again involved the use of energetic α-particles, colliding this time with boron.

$$^{10}_{5}\mathrm{B} + ^{4}_{2}\mathrm{He} \rightarrow ^{13}_{7}\mathrm{N} + ^{1}_{0}\mathrm{n}$$

The product of the reaction, an isotope of nitrogen, decays by emission of a positron:

$$^{13}_{7}\mathrm{N} \rightarrow ^{13}_{6}\mathrm{C} + ^{0}_{+1}\mathrm{e} + \nu_{\mathrm{e}}$$

Release of a positron is called *β^+-decay*, and to conserve energy and spin, the positron is accompanied by an electron neutrino, ν_e. As with β^--decay, electron neutrinos are not always included in nuclear equations.

Two further examples of this process are:

$$^{38}_{19}\mathrm{K} \rightarrow ^{38}_{18}\mathrm{Ar} + ^{0}_{+1}\mathrm{e}$$

$$^{120}_{51}\mathrm{Sb} \rightarrow ^{120}_{50}\mathrm{Sn} + ^{0}_{+1}\mathrm{e}$$

Positrons, like electrons, do not exist in the nucleus, and are believed to be generated by the transformation of a proton into a neutron; a process requiring energy input:

$$^{1}_{1}\mathrm{H} \rightarrow ^{1}_{0}\mathrm{n} + ^{0}_{+1}\mathrm{e}$$

Positrons have a lifetime of about 10^{-9} seconds, and are annihilated by combination with electrons to produce γ-rays:

$$\mathrm{e}^+ + \mathrm{e}^- \rightarrow \gamma$$

A wide variety of reactions are now utilised for the production of artificial radioisotopes including neuron and proton bombardment. Some examples are:

$$^{6}_{3}\mathrm{Li} + ^{1}_{0}\mathrm{n} \rightarrow ^{4}_{2}\mathrm{He} + ^{3}_{1}\mathrm{H}$$

$$^{14}_{7}\mathrm{N} + {}^{1}_{1}\mathrm{H} \rightarrow {}^{11}_{6}\mathrm{C} + {}^{4}_{2}\mathrm{He}$$

$$^{58}_{26}\mathrm{Fe} + 2{}^{1}_{0}\mathrm{n} \rightarrow {}^{60}_{27}\mathrm{Co} + {}^{0}_{-1}\mathrm{e}$$

Cobalt-60 is used in cancer therapy. It breaks down according to the reaction:

$$^{60}_{27}\mathrm{Co} \rightarrow {}^{60}_{28}\mathrm{Ni} + {}^{0}_{-1}\mathrm{e}$$

The emission of γ-rays during a nuclear reaction does not change either mass number or atomic number, because they are high-energy photons. This often takes place after an α- or β-emission because these latter processes frequently leave the daughter nucleus in an excited state. An example is:

$$^{16}_{7}\mathrm{N} \rightarrow {}^{16}_{8}\mathrm{O} + {}^{0}_{-1}\mathrm{e} + \gamma$$

The daughter nucleus, $^{16}\mathrm{O}$ in this case, is formed in an excited state. Stability is gained when the nucleus subsequently loses energy by way of γ-emission.

14.3 Nuclear Decay

14.3.1 The Rate of Nuclear Decay

One of the most important properties of a radioactive nuclide is its lifetime. At present it is not possible to predict theoretically when any particular nucleus in a sample will decay. However, the number of nuclides in a sizable sample that will decompose in a given time can be measured, and it is found that this rate of decay is a characteristic of a given isotope. The rate of decay of an isotope is constant and unvarying. That is, if a fraction of a radioactive nuclide decays in a certain time interval *t*, then the same fraction of the remainder will decay in another increment of time *t*. Nuclear reactions are not affected by outside influences such as temperature and pressure, and it is not possible to significantly alter this constant rate of radioactive decay. For example, radioactive strontium-90, an important product of nuclear fission, undergoes β-decay to yield the daughter atom yttrium:

$$^{90}_{38}\mathrm{Sr} \rightarrow {}^{90}_{39}\mathrm{Y} + {}^{0}_{-1}\mathrm{e}$$

It is found that it will take 29 years for ½ of the sample to decay and another 29 years for half of the remaining strontium-90 to decay and so on.

Time	0	29	58	87	Years
Amount of Sr-90	1	1/2	1/4	1/8	Grams

The lifetime of a radioactive substance is usually quoted in terms of the time in which half of the sample decays, called its *half-life*, $t_{1/2}$ (Table 14.3).

Table 14.3 Half-lives of some radioisotopes.

Nuclide	Half-life	Decay mode
$^{3}_{1}H$	12.33 yr	β^-
$^{14}_{6}C$	5730 yr	β^-
$^{22}_{11}Na$	2.602 yr	β^+, γ
$^{47}_{20}Ca$	4.536 d	β^-, γ
$^{59}_{26}Fe$	44.496 d	β^-, γ
$^{60}_{27}Co$	5.271 yr	β^-, γ
$^{90}_{38}Sr$	28.5 yr	β^-
$^{131}_{53}I$	8.040 d	β^-, γ
$^{133}_{54}Xe$	5.245 d	β^-, γ
$^{137}_{55}Cs$	30.1 yr	β^-, γ
$^{198}_{79}Au$	2.639 5 d	β^-, γ
$^{222}_{86}Rn$	2.825 d	α, γ
$^{226}_{88}Ra$	1600 yr	α, γ, X[a]
$^{228}_{88}Ra$	5.75 yr	β^-, γ
$^{235}_{92}U$	7.037×10^8 yr	α, γ, X[a]
$^{238}_{92}U$	4.468×10^9 yr	α, γ

[a] X-ray emission.

The number of atomic disintegrations that occur in a radioactive material per second, the *rate of decay*, is called its *activity*. A rate of decay of 1 disintegration per second is the SI unit of activity, 1 *becquerel* (Bq). The *specific activity* of a material is the activity of 1 g of that material. Radium has a specific activity 3.7×10^{10} Bq g^{-1}.

The disintegration of atomic nuclei can be expressed by the *first-order* rate equation:

$$-\mathrm{d}N/\mathrm{d}t = \lambda N \tag{14.1}$$

where $\mathrm{d}N/\mathrm{d}t$ is the rate of disintegration and N is the number of atoms present at any given instant. The proportionality constant, λ, usually called the *decay constant*, is different for each radioactive isotope. To relate the number of atoms N_0 that exits at time $t = 0$ to the amount present at any later time, N, Eq. (14.1) Figure 14.6, must be integrated, to give:

$$N = N_0\, \mathrm{e}^{-\lambda t} \quad \text{or} \quad N_0 = N\, \mathrm{e}^{\lambda t} \tag{14.2}$$

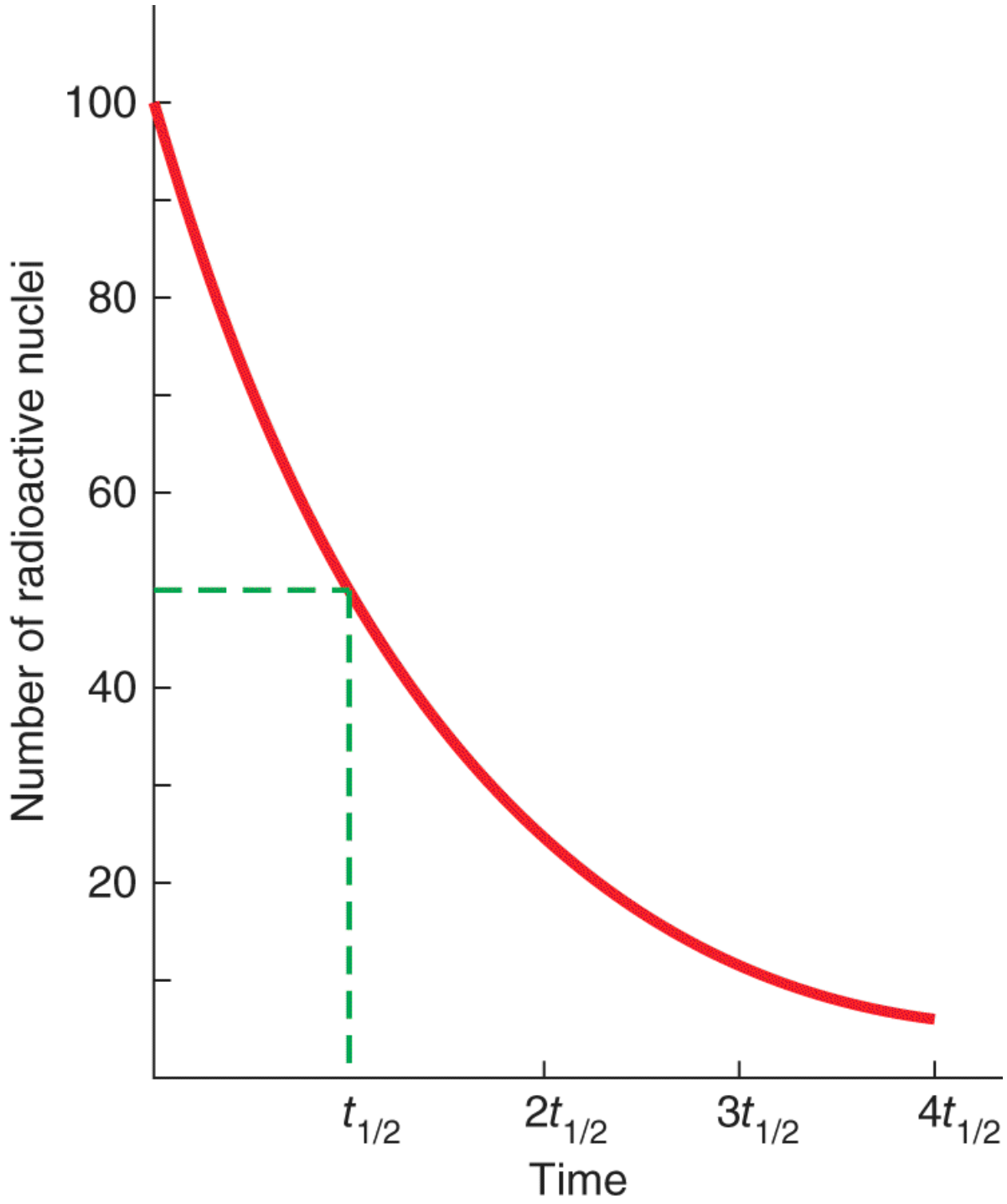


Figure 14.6 The exponential rate of radioactive decay.

Equation (14.2) is often written in the logarithmic form:

$$\ln N - \ln N_0 = -\lambda t$$

or

$$\ln[N_0/N] = \lambda t \tag{14.3}$$

The value of the decay constant can be determined from a plot of ln N vs. t, the slope of the straight-line graph being equal to $-\lambda$. Substitution of $N = ½\ N_0$ into Eq. (14.3) shows that the half-life is given by:

$$t_{½} = 0.693/\lambda \tag{14.4}$$

The fraction of the radio-isotope remaining after n half-lives have elapsed is:

$$N/N_0 = [½]^n$$

The specific activity of a material can be related to its half-life in the following way.

$$\text{Activity} = -\mathrm{d}N/\mathrm{d}t = \lambda N$$

The specific activity is the activity divided by the mass in grams, m:

$$\text{Specific activity} = \text{activity}/m = \lambda N/m$$

Substituting for λ from Eq. (14.4) gives:

$$\text{Specific activity} = 0.693\ N/t_{½}\ m$$

The number of atoms, N, remaining at any time, and their mass, m, is related in the following way. The number of moles of material present is given by:

$$\text{Moles of material} = N/N_A$$

where N_A is the Avogadro constant. The mass of this quantity is:

$$m = M\ [N/N_A]$$

where M is the isotopic mass of the isotope. Hence:

$$N/m = N_A/M$$

and

$$\text{Specific activity} = 0.693N_A/Mt_{1/2}$$
$$= 4.2 \times 10^{23}/Mt_{1/2}$$

14.3.2 Radioactive Dating

One of the principal problems confronting geologists and archaeologists is the accurate dating of the materials under examination. Because of the fixed rate of decay of radioactive isotopes, naturally occurring radioactive minerals can be used to for this purpose. In fact, it was Rutherford who first suggested that the then newly discovered radioactive atoms could find a use in the absolute determination of the age of rocks and minerals. A decay series widely used for dating rocks involves ^{87}Rb, which transforms to stable (non-radioactive) ^{87}Sr by β^--emission:

$$^{87}_{37}\text{Rb} \rightarrow {}^{87}_{38}\text{Sr} + {}^{0}_{-1}\text{e}; \quad \lambda = 1.46 \times 10^{-11}\ \text{yr}^{-1}$$

This example can be used to illustrate the method of radioactive dating. In principle the amount of ^{87}Sr present in a specimen is measured and the age of the sample is then obtained from Eqs. (14.2) and (14.3) knowing the value of the decay constant. However, in practical terms it is easier to measure the amount of the radioactive nuclide remaining, N_r rather than the initial concentration N_0. Knowing that one disintegration removes one radioactive nuclide and leaves one product nuclide:

$$N_r = N_0 - N$$

so that Eqs. (14.2) and (14.3) become:

$$N_r = N(e^{\lambda t} - 1)$$

$$\ln[(N + N_r)/N] = \lambda t$$

that is:

$$^{87}\text{Sr} = {}^{87}\text{Rb}\,(e^{\lambda t} - 1)$$

where the amounts of ^{87}Sr and ^{87}Rb are those present in the sample. These quantities can be obtained using a mass spectrometer. However, a practical difficulty arises because mass spectrometry gives a more accurate measurement of the ratios of nuclides rather than absolute values and for this purpose it is necessary to choose another nuclide to act as a reference. For Rb/Sr decay the isotope ^{86}Sr is generally used, because it is not produced from a radioactive event and it is not radioactive, allowing the overall abundance to be taken as constant. The mass spectrometric results then provide the $^{87}Sr/^{86}Sr$ and $^{87}Rb/^{86}Sr$ ratios, to give:

$$^{87}Sr/^{86}Sr = (^{87}Rb/^{86}Sr)\,(e^{\lambda t} - 1)$$

There is one more correction to make. There may be some ^{87}Sr present when the material was first formed. This must clearly be subtracted from the amount measured if a true time is to be obtained, that is:

$$(^{87}Sr/^{86}Sr) - (^{87}Sr/^{86}Sr)_0 = (^{87}Rb/^{86}Sr)\,(e^{\lambda t} - 1)$$

or

$$(^{87}Sr/^{86}Sr) = (^{87}Rb/^{86}Sr)\,(e^{\lambda t} - 1) + (^{87}Sr/^{86}Sr)_0$$

where $(^{87}Sr/^{86}Sr)_0$ represents the initial ratio when the sample was formed. This is the equation of a straight line and a plot of $(^{87}Sr/^{86}Sr)$ vs. $(^{87}Rb/^{86}Sr)$ for a number of samples will have a slope of $(e^{\lambda t} - 1)$ and an intercept of $(^{87}Sr/^{86}Sr)_0$ (Figure 14.7). The age of the material can then be determined from the slope of the graph.

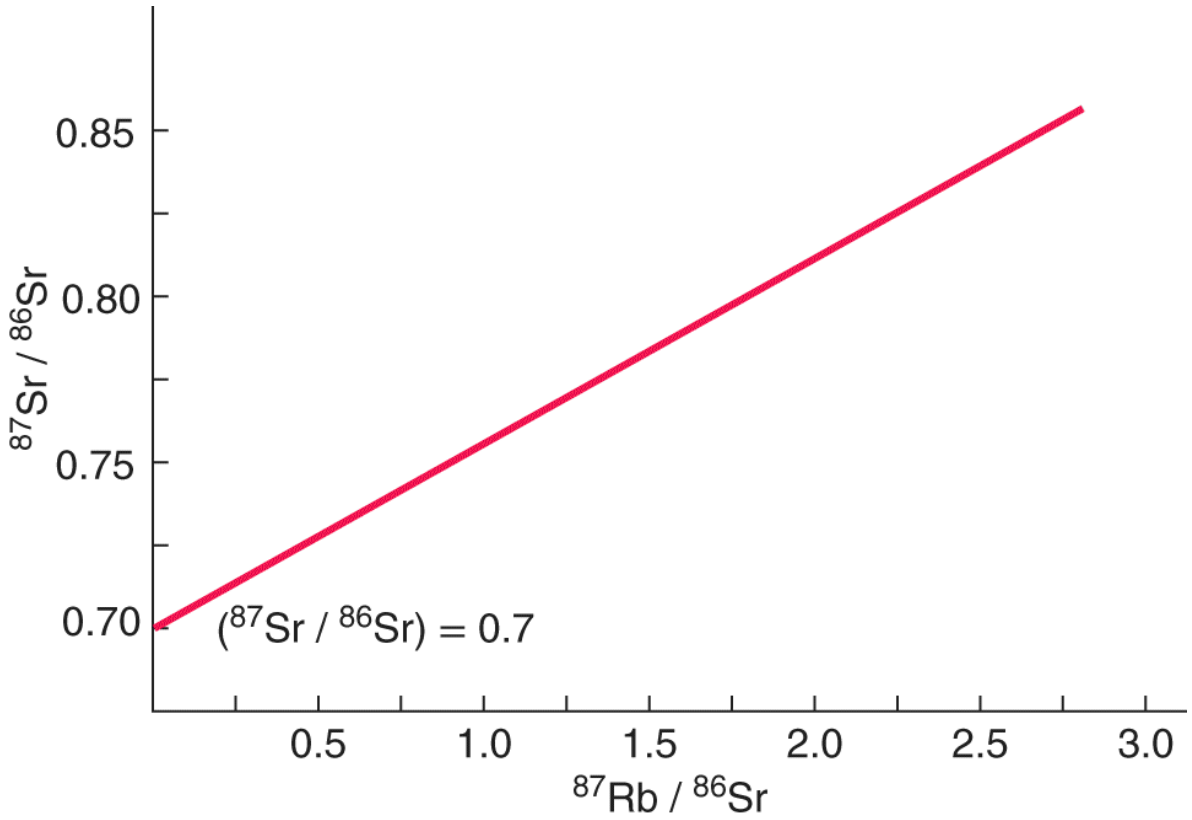


Figure 14.7 Plot of isotope ratios $^{87}Sr/^{86}Sr$ vs. $^{87}Rb/^{86}Sr$, schematic.

The Rb/Sr pair cannot be employed in all situations and a number of other decay processes are widely used. One of the most important involves the chain reaction of the uranium isotopes that end in lead (Figures 14.1 and 14.3).

$^{238}U \rightarrow {}^{206}Pb$, measuring $^{206}Pb/^{204}Pb$

$^{235}U \rightarrow {}^{207}Pb$, measuring $^{207}Pb/^{204}Pb$

The age of the earth, approximately 4550 million years, is largely reached via uranium decay dating. The uranium method frequently relies upon small crystals of zircon, $ZrSiO_4$, that are present at the site to be dated because zircon often contains small amounts of uranium substituted for zirconium. The uranium decay chains then form the basis of the analytical method similar to that set out earlier.

For $^{238}U \rightarrow {}^{206}Pb$:

$$(^{206}Pb/^{204}Pb) - (^{206}Pb/^{204}Pb)_0 = (^{238}U/^{204}Pb)\,(e^{\lambda 1 t} - 1)$$

For $^{235}U \rightarrow {}^{207}Pb$:

$$(^{207}Pb/^{204}Pb) - (^{207}Pb/^{204}Pb)_0 = (^{235}U/^{204}Pb)\,(e^{\lambda 2t} - 1)$$

where λ_1 and λ_2 are the relevant decay constants for the two decomposition chains. Combining these two equations:

$$\begin{aligned}&[(^{207}Pb/^{204}Pb) - (^{207}Pb/^{204}Pb)_0]/\\&[((^{206}Pb/^{204}Pb) - (^{206}Pb/^{204}Pb)_0)]\\&= [(^{235}U)(e^{\lambda 2t} - 1)]/[(^{235}U)(e^{\lambda 1t} - 1)]\end{aligned}$$

A plot of ($^{207}Pb/^{204}Pb$) against ($^{206}Pb/^{204}Pb$) allows the value of *t*, the age of the sample, to be determined from the slope of the graph, provided that the ratio of the isotopes $^{238}U/^{235}U$ is known. For many years this was considered to be invariant in all rocks throughout the solar system (including the moon and meteorites) at 137.88 but recently the figure has been revised to 137.818, meaning that a slight recalibration is needed in earlier studies.

Another widely decay used reaction is:

$$^{40}_{19}K \rightarrow {}^{40}_{18}Ar + {}^{0}_{-1}e; \quad t_{1/2} = 1.3 \times 10^9 \text{ years}$$

This reaction gives dates over the same time range as the U-238–Pb-208 analysis, but can use micas, such as biotite, $K(MgFe)_3(AlSi_3O_{10})(OH)_2$, that naturally contain potassium. As micas have a different distribution to uranium-containing minerals, these allow dating to be applied to rocks and sediments lacking in uranium containing compounds.

To date organic materials less than about 50 000 years old, the radioactive decay of carbon-14 is preferred. Carbon-14 is produced at a steady rate in the Earth's upper atmosphere, due to the interaction of cosmic ray neutrons with nitrogen:

$$^{14}_{7}N + {}^{1}_{0}n \rightarrow + {}^{14}_{6}C + {}^{1}_{1}H$$

Carbon-14 subsequently degrades by β-decay:

$$^{14}_{6}\mathrm{C} \rightarrow {}^{14}_{7}\mathrm{N} + {}^{0}_{-1}\mathrm{e}; \quad t_{1/2} = 5730 \text{ years}$$

Carbon-14 is distributed throughout the atmosphere in the form of carbon dioxide, CO_2, molecules, and due to atmospheric diffusion a fairly constant proportion of all CO_2 is radioactive due to the presence of carbon-14. Living plants and animals absorb CO_2 and so incorporate carbon-14 into their tissues. The relative quantity of carbon-14 in an organism remains constant until it dies. At this point the carbon-14 begins to decay at its normal rate. A measurement of the radioactivity of the once-living samples, such as bones, burnt charcoal or plant remains, can then be used to determine the age of the sample itself using Eq. (14.2). The raw data obtained in this way has an inherent error of 10–20%, because of naturally occurring fluctuations in the amount of carbon-14 present in the atmosphere due to irregular variations in solar activity. In order to make the correspondence between the carbon-14 date and the calendar date precise, the results must be compared with independent age measurements such as the annual growth rings on trees. Such calibration curves are being continually refined, allowing for more precise dating of these relatively recent past events.

14.4 Nuclear Energy

14.4.1 The Binding Energy of Nuclides

When protons and neutrons are (conceptually) brought together to form an atomic nucleus, its mass is less than the combined masses of the protons and neutrons. This difference, when expressed as energy, is called the *binding energy* of the nucleus. Writing the formation of a nucleus ${}^{A}_{Z}\mathrm{X}$ as a pseudo-chemical reaction:

$$Z \text{ protons} + (A - Z) \text{ neutrons} \rightarrow {}^{A}_{Z}\mathrm{X}$$

the difference between the mass of the products and the mass of the reactants Δm is given by:

$$\Delta m = \text{mass}[{}^{A}_{Z}\mathrm{X}] - \text{mass}[Z \text{ protons} + (A - Z) \text{ neutrons}]$$

This difference mass is converted into energy using the Einstein equation:

$$E = \Delta m c^2$$

where c is the speed of light.

In making these calculations it is simplest to use the masses of the particles in atomic mass units, u. The mass of an atom in atomic mass units is numerically equal to its isotopic mass, based on 1/12 of the mass of the isotope carbon-12. As the mass of one atom of carbon-12 is $1.992\,6 \times 10^{-26}$ kg, $1u = 1.660\,5 \times 10^{-27}$ kg. Moreover, it is usual to use the mass of the hydrogen isotope ${}^{1}_{1}\mathrm{H}$ rather than the mass of the proton, in making the calculations. This is because the mass of the isotopes is also obtained from the relative molar mass of atoms, and so the mass of the electrons present on the ${}^{1}_{1}\mathrm{H}$ isotopes cancels neatly with the mass of the electrons on the ${}^{A}_{Z}\mathrm{X}$ isotope.

The isotopic binding energy is often quoted per nucleon (Figure 14.8). The curve is smooth over much of the range. However, for the lightest elements, a series of peaks occur at ^{4}He, ^{12}C, and ^{16}O. The curve rises to

a maximum at ^{56}Fe, the most stable isotope, and decreases slightly thereafter. Isotopes of lower nucleon number than ^{56}Fe will release energy in fusion reactions (Section 14.4.6). Isotopes of higher nucleon number than ^{56}Fe will release energy on fission (Section 14.4.2).

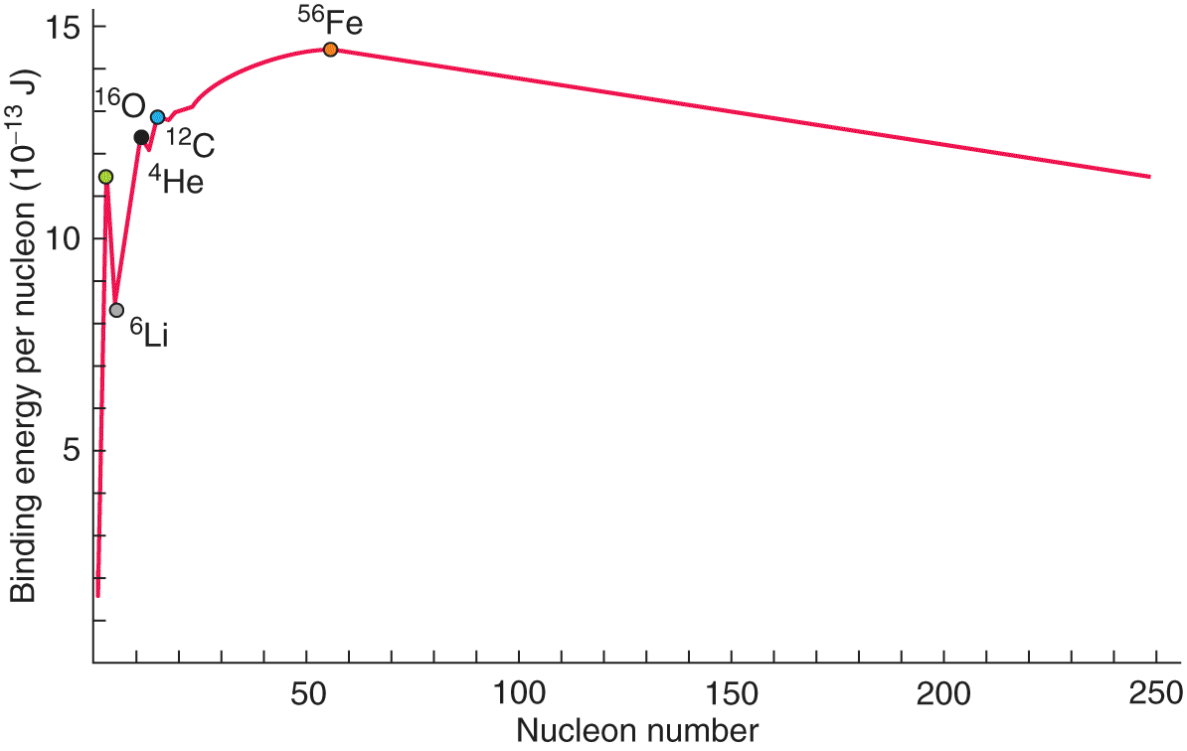


Figure 14.8 The isotopic binding energy per nucleon.

14.4.2 Nuclear Fission

Nuclear fission is the breaking apart of atomic nuclei into two or more pieces. This can take place *spontaneously* in the case of the heaviest atoms. Neutron bombardment of atoms can also cause the nuclei to break apart. This process is called *induced nuclear fission* and susceptible atoms are said to be *fissionable*. Some nuclides can undergo fission with slow (not very energetic or *thermal*) neutrons. These atoms are called *fissile*.

The neutron, being uncharged, is not repelled by the positive charge on the nucleus, and makes an ideal nuclear probe. Soon after the discovery of the neutron, many experiments were carried out to try to make new elements that were more massive than uranium by bombarding heavy atoms, notably U itself, with neutrons. Two such elements that can be made this way are neptunium, Np, and plutonium, Pu.

$$^{238}_{92}\mathrm{U} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{239}_{93}\mathrm{Np} + {}^{0}_{-1}\mathrm{e}$$

Neptunium has a half-life of about two days and decays to plutonium-239 by way of β-decay:

$$^{239}_{93}\mathrm{Np} \rightarrow {}^{239}_{94}\mathrm{Pu} + {}^{0}_{-1}\mathrm{e}$$

Plutonium has a half-life of about 24 000 years.

However, the experiments most often resulted in fission of the heavy nuclei into two more or less equal parts during the bombardment. For example, fission of uranium-235 can produce krypton and barium:

$$^{235}_{92}\mathrm{U} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{92}_{36}\mathrm{Kr} + {}^{141}_{56}\mathrm{Ba} + 3{}^{1}_{0}\mathrm{n} \qquad (14.5)$$

In practice a range of fission products with masses similar to krypton and barium are formed when uranium is irradiated with neutrons. This reaction is important, as it is used to produce nuclear power. There are two vital features that make this application possible; firstly, the amount of energy liberated, and secondly, the number of neutrons produced.

The energy produced in all fission reactions is derived from the difference in masses of the reactants and the products. This mass difference is liberated as heat. The mass difference for Eq. (14.5) is -3.198×10^{-28} kg. The negative value arises because the reactants are heavier than the products, and this is the amount of mass that is converted into energy. It amounts to 2.878×10^{-11} J. The amount of energy per gram of U-235 is then calculated to be 7.5×10^{10} J. In contrast to this, 1 g of coal burnt in air produces about 3×10^{4} J. That is, uranium-235 fission produces about one million times more energy than burning fossil fuels.

The second important feature of Eq. (14.5) is the number of neutrons emitted. When more neutrons are emitted than are produced, an ever-accelerating reaction, called a *chain reaction*, can result (Figure 14.9). Suppose that the first disintegrating nucleus is surrounded by sufficient U-235 so that all of the neutrons are absorbed and none are lost through the surface of the material. These will then undergo fission to produce more neutrons, and so on, in a rapidly escalating fashion. Unless controlled, in a fraction of a second all of the U-235 will have transformed into fission products, with the liberation of huge amounts

of energy. This is the principle upon which fission atomic weapons operate. The smallest amount of material for which more neutrons are produced than are lost through the surface is called the *critical mass*. A quantity of U-235 smaller than the critical mass will not support a chain reaction because too many neutrons escape without hitting another U-235 nucleus.

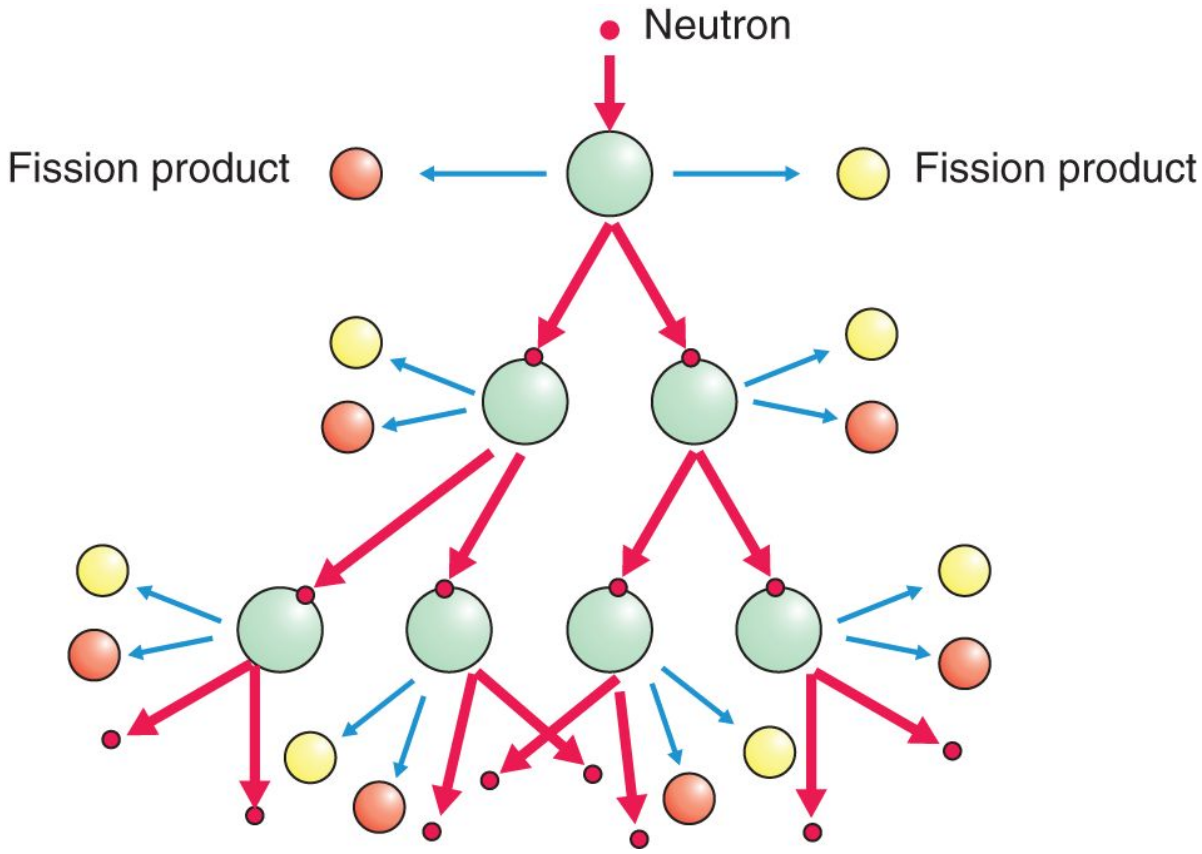


Figure 14.9 A fission chain reaction, schematic. In the chain illustrated, each nucleus emits two neutrons that both cause further fission.

For a chain reaction to occur it is not enough simply to have a quantity of material greater than the critical mass. The key to the continuation of the reaction is that each dividing nucleus must emit more than one neutron *that reacts with another nucleus*. To ensure that this occurs, it is necessary for the uranium to be pure, or at least not to contain appreciable quantities of substances that absorb neutrons. Moreover, the isotope U-235 is fissile and reacts best with thermal neutrons and the arrangement of the uranium-235 must incorporate a *moderator*, which is a material for slowing down the energetic neutrons released by the fission.

14.4.3 Thermal Reactors for Power Generation

The fission of U-235 is used in a nuclear reactor to produce nuclear power. Because U-235 interacts with low energy thermal neutrons, the reactors are usually called *thermal reactors* (Figure 14.10). The fuel used in almost all nuclear power reactors is uranium dioxide. Natural uranium ores consist of approximately 99% of the U-238 isotope, about 0.71% U-235 and small amounts of U-234. For power production, the amount of U-235 in the UO_2 is increased to 2–4%, the resulting material being *enriched* uranium dioxide. This is achieved by converting the uranium present in the ores to a gas, uranium hexafluoride. The hexafluoride is then spun at high speeds in a gas centrifuge. This has the effect of separating the lighter, $^{235}UF_6$, and heavier, $^{238}UF_6$, molecules, on the basis of the slight difference in their masses. The hexafluoride mixture that is enriched in the U-235 isotope is then converted back into UO_2 and this is made into fuel pellets. The moderator employed is pressurised water, which also acts as a coolant, transferring the heat produced to steam generators.

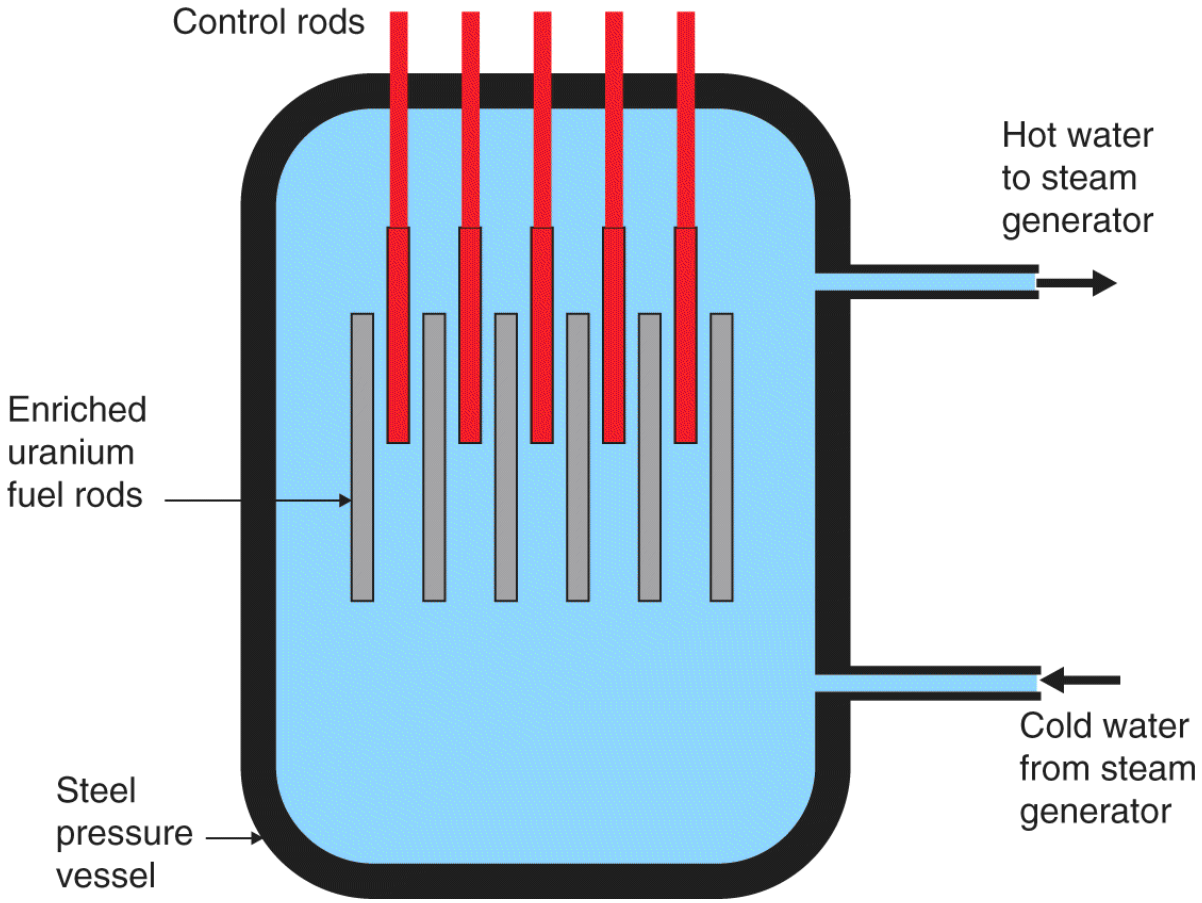


Figure 14.10 A pressurised water nuclear reactor for power generation, schematic.

Uranium dioxide has a number of properties that make it suitable for a fuel. The crystal structure, the fluorite (CaF_2) type (Figure 5.24), is stable to temperatures in excess of 2800 °C. Moreover, because it is a ceramic oxide, the material is refractory, chemically inert, and resistant to corrosion Enrichment does not change these features. The oxide powder is pressed into pellets and sintered to a density of about 95% maximum by traditional ceramic processing technology, but carried out in conditions that minimise risks from radiation effects. The pellets are contained in zirconium alloy (Zircaloy) containers, which are then introduced into the reactor. During operation the outer surface of the Zircaloy is at a temperature of approximately 400 °C, while the centre of the fuel rod can be at 1700 °C. This can lead to cracks and grain boundary growth as well as causing differential thermal expansion problems. The moderator, which surrounds the fuel rods, can be graphite, ordinary water, or heavy water (This latter substance is water in which the ordinary hydrogen isotope, ^1H, is replaced by the heavier

deuterium isotope, ^2H, given the symbol D, which has a nucleus containing one proton and one neutron. Heavy water is often written as D_2O.). Reactors also contain control rods of cadmium and boron that absorb neutrons. In the event of an accident, the control rods can be inserted into the reactor, which has the effect of stopping the chain reaction and closing the reactor down.

Some reactors, called *fast reactors*, do not use thermal neutrons and have no need of extensive systems of moderators in place. The drawback of these is that they need fuel that is more highly enriched in fissile material because the neutron absorption process has a lower efficiency (Also see Section 14.4.5.).

The majority of current nuclear power plants use variations of the scheme outlined, in what are termed Generation II reactors. Generation III reactors offer incremental design improvements on Generation II technology. Generation IV reactors are the subject of much continuing study. These are being designed to operate with greater efficiency and safety and are intended to operate at temperatures in excess of 1000 °C. Six Generation IV technologies have been proposed; gas-cooled fast reactors, lead-cooled gas reactors, molten salt reactors, sodium-cooled fast reactors, supercritical water-cooled reactors, and very high temperature gas reactors. All of these designs make substantial demands upon the construction materials. To give just one example, the higher operating temperatures of the reactors will necessitate new thermocouple materials that will allow accurate temperature measurement while being able to withstand intense neutron bombardment.

14.4.4 Fuel for Space Exploration

Space exploration relies heavily on solar energy when the spacecraft is in the inner solar system. However, solar power is insufficient for spacecraft that have to journey to the outer planets. Chemical energy sources, typically batteries, tend to be heavy and have rather short lifetimes for missions that are to last many years. The solution adopted is to combine a radioactive solid with a thermoelectric generator (Section 13.4.4). The advantages of this solution are that there are no liquids to spill, no moving parts to wear, and a nuclear isotope with a long half-life will continue to provide power over the lifetime of the craft.

The chosen fuel, used in many space missions, including the Mars Curiosity rover that landed 6 August 2012, is the isotope ^{238}Pu. This is an α-emitter with a half-life of 87.4 years, which provides a power of about 0.5 W g^{-1}. The fuel is the solid oxide $^{238}PuO_2$. Chemically it is similar to the uranium dioxide used in thermal reactors, and adopts the same fluorite crystal structure. It is inert chemically and stable up to the melting point, approximately 2500 °C. The oxide is pressed and sintered into pellets under conditions that lead to high density and low, but not zero, porosity. This is to ensure dimensional stability of the pellets over the lifetime of the spacecraft, because as ^{238}Pu is an α-emitter, the resulting helium gas must be allowed to escape.

14.4.5 Fast Breeder Reactors

The amount of U-235 present in a nuclear fuel rod is gradually depleted, and ultimately there is insufficient present for the economic generation of power. A fast breeder reactor uses the interaction of U-238 with energetic (fast) neutrons, to generate the plutonium isotope Pu-239. As Pu-239 can be used as a nuclear fuel, a breeder reactor produces more fuel than it consumes. The sequence of steps is:

$$^{238}_{92}U + {}^{1}_{0}n \rightarrow {}^{238}_{92}U \quad (t_{1/2} = 24\ \text{minutes})$$

$$^{239}_{92}U \rightarrow {}^{239}_{93}Np + {}^{0}_{-1}e \quad (t_{1/2} = 2.35\ \text{days})$$

$$^{239}_{93}Np \rightarrow {}^{239}_{94}Pu + {}^{0}_{-1}e \quad (t_{1/2} = 24\,000\ \text{years})$$

Pu-239 has a half-life of 24 000 years and can be collected for use in fission reactions. The high neutron flux needed is obtained from a reactor using uranium-235 and no moderator. As each decay of a uranium nucleus produces more than two neutrons, it is possible for the reactor to produce more plutonium-239 than it consumes uranium-238. However, the returns are not great, and it would take about 20 breeder reactors running for one year to produce enough plutonium to run a further reactor for one year.

The fuel in fast breeder reactors is an oxide, as with a thermal reactor. The material chosen is a solid solution of uranium and plutonium dioxides, $U_xPu_{1-x}O_2$. This material shares the same fluorite crystal

structure as uranium dioxide and plutonium dioxide and the solid solution is readily prepared and stable.

14.4.6 Fusion and Solar Cycles

The binding energy curve (Figure 14.8) shows that far more energy should be released when the lightest atoms are built up into heavier ones than when heavy atoms are broken down into lighter ones. These building-up reactions are called *fusion reactions*. Of those available, the production of helium from hydrogen would appear to be the reaction that would yield the most energy per atom.

As hydrogen is commonly available, and as fusion reactions do not produce radioactive waste problems, there has been considerable research into the construction of controlled fusion reactors. In order for fusion reactions to occur, extremes of temperature must be achieved, some millions of degrees, while keeping the reactants confined and under control. Intensive study of this problem has not yet produced a working reactor, but the possibility of producing essentially limitless clean power will continue to be a goal for research.

Uncontrolled fusion reactions form the basis of the 'hydrogen bomb'. The temperature needed to initiate fusion is brought about by a fission bomb. A number of competing and cooperative reactions take place during fusion. Typical reaction schemes for fusion reactions are:

$$^{2}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} \rightarrow {}^{3}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}$$

$$^{2}_{1}\mathrm{H} + {}^{3}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}$$

These involve the hydrogen isotopes deuterium, $^{2}\mathrm{H}$, and tritium, $^{3}\mathrm{H}$, as reactants. Tritium has a low natural abundance, and is generated in the reaction by use of the isotope lithium-6.

$$^{6}_{3}\mathrm{Li} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{3}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He}$$

The reacting material is lithium-6 deuteride, $^{6}\mathrm{Li}^{2}\mathrm{H}$, which forms the core inside a fission bomb. The energy released by such reactions can be calculated by the methods described above.

The process that powers the Sun and other stars is fusion. It seems that the early universe contained clouds of hydrogen atoms dispersed throughout space. Gravitational forces gradually caused these to collapse into dense regions. Ultimately it is supposed that when the collapse produced a sufficiently high-pressure core, the temperature reached the order of 10^7 °C and hydrogen fusion started. This process is taking place in the Sun today. Under the intense conditions within the Sun's core, the fusion of protons into helium nuclei takes place following the reaction scheme:

$$4\,^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He} + 2\,{}^{0}_{+1}\mathrm{e}$$

The energy production is enormous, about 10^{26} J s^{-1}.

In older stars the hydrogen is eventually consumed and the rate of fusion slows. At this stage gravitational collapse again occurs, resulting in an increase in pressure and temperature until helium fusion starts. The products of this reaction are mostly C and O. When the He supply becomes depleted, smaller stars may explode under gravitational collapse. Stars that are larger than about 20 solar masses, that is, about 20 times the mass of the Sun (one solar mass is about 2×10^{30} kg), can collapse in a relatively controlled fashion until core temperatures reach the order of 10^9 °C. At this point carbon and oxygen fusion begins. The products are now the elements close to Si in the Periodic Table, known as the *silicon peak elements*. Further exhaustion and collapse increase the temperature to the order of 3×10^9 °C, at which point the Si elements fuse to give the *iron peak elements*, Cr, Mn, Fe, Co, and Ni. The most abundant elements in the universe, which are, in order of abundance, H, He, O, Ne, N, C, Mg, Si, Fe, and S, mirror these processes.

The binding energy curve (Figure 14.8) indicates that at the iron peak fusion no longer produces energy as ^{56}Fe is the element with the greatest binding energy and so further fusion will consume rather than realise energy. At this point in the life of a star, the nuclear reactions diminish, gravitational collapse raises the internal pressure, and ultimately the star explodes as a supernova. At this point vast numbers of nuclear reactions take place and enormous numbers of other atoms are synthesised.

14.5 Nuclear Waste

The problems with unwanted nuclear material, irrespective of its origin, are the same. Although some nuclear materials are poisonous, at the heart of the matter are the facts that radioactive isotopes behave chemically in an identical way to non-radioactive counterparts, and they emit damaging radiation that can disrupt cells, leading to severe health problems for living organisms. Even these features would not pose grave problems, though, if radioisotopes were short lived, or the decay process could be accelerated. Unfortunately, some of the more important by products of fission are long lived, and nuclear decay rates cannot be altered by any physical or chemical means at our disposal.

In a nuclear accident, these problems have to be dealt with as an emergency, while in the case of decommissioning of nuclear plant, or disposal of laboratory chemicals, the problems can be approached within a longer timescale.

14.5.1 Nuclear Accidents

There have been few serious nuclear accidents, the most recent being at Chernobyl, in the (then) USSR, when a nuclear reactor caught fire on 26 April 1986 and at Fukushima Daiichi, Japan, on 11 March 2011 following an earthquake and tsunami. In the Chernobyl accident the lid of the reactor was blown off and the explosion and fire sent radioactive material high into the atmosphere. The reactor cooling system was closed down and the chain reaction came close to an uncontrolled nuclear explosion. In the event, steam generated in the accident blew open the reactor. Subsequent reactions between steam and graphite and steam and the Zircaloy fuel cladding produced hydrogen, which then caught fire. The reactor core and the graphite burnt with a temperature of about 1600 °C. As the fuel was uranium dioxide, UO_2, with a melting point of 2856 °C, no melting of fuel seems to have occurred, although reaction with other materials including reaction products can lower melting points considerably (Sections 1.2.1 and 1.4.1).

At Fukushima Daiichi the reactors shut down automatically following an earthquake, and emergency power supplies took over the control electronics and cooling water pumping system. Unfortunately, the resulting tsunami flooded the emergency power generator location and all cooling was lost. The reactors started to heat up due to the normal decay of materials in the core and hydrogen produced by reaction of

Zircaloy and water subsequently exploded and damaged buildings with the result that multiple fires broke out. Full melt-down of three reactors occurred resulting in a 'larva flow' of molten fuel assemblies, combined with concrete, steel, sand, and so on. One product of this reaction mixture was zircon, $U_xZr_{1-x}SiO_4$, the same crystals that are used in uranium dating (Section 14.3.2).

Further problems can arise at the site of an accident during the clean-up. Uranium dioxide is easily transformed into the water-soluble uranyl cation $(UO_2)^{2+}$. The use of water, including seawater, to control fires can then lead to contamination of local water supplies. However, the major problem in accidents, apart from local radioactivity, is the formation of volatile fission products. These may be widely distributed by winds leading to pollution of large areas. The main volatile elements released at Chernobyl were the noble gases (especially krypton and xenon), ^{131}I, and ^{137}Cs. At Fukushima, radioactive calcium was important locally, but iodine-131 and caesium-137 were distributed globally. The caesium isotopes cesium-134 and cesium-137 pose a particular problem. The half-life of caesium-134 is 2.06 years, decaying via β-emission. The half-life of caesium-137 is 30.17 years, which also decays via β-emission. Plants readily take up caesium, which is chemically similar to its neighbouring alkali metal potassium, and which is a trace element vital to plant growth. The caesium can then transfer to meat, milk, eggs, etc. by way of grazing animals, and so enter the human food chain.

14.5.2 The Storage of Nuclear Waste

The storage of nuclear waste is one of the more significant challenges facing materials engineers. A storage facility must last for thousands of years, and even small amounts of leakage are unacceptable. Moreover, radioactive waste is not a passive material. The heat generated by the radioisotopes is significant, damage from radiation greatly increases the rates of chemical changes, and the products of nuclear decay result in significant pressure and volume changes that can lead to container damage.

Nuclear waste is divided into three categories. *High-level waste*, which is the most radioactive component, forms about 0.2% of the whole. It is mainly derived from weapons applications and spent nuclear fuel rods. In addition, there is about 20% *intermediate level waste*, which arises from similar sources and is increased by materials used in reprocessing.

This component is not very radioactive, and does not liberate large amounts of heat. The remainder, described as *low-level waste*, is a material that is slightly radioactive. Apart from military and nuclear energy sources, this material comes from hospitals, research laboratories, and industry, and includes contaminated paper towels, gloves, and laboratory equipment.

Spent fuel rods from nuclear power stations are a major source of nuclear waste. Nuclear fuel is composed of uranium dioxide. After some years of use, when 1–4% of the uranium has undergone fission, the performance of the fuel rods falls and they are replaced. The spent fuel rods consist of uranium dioxide together with fission products, typically the gases krypton and xenon, volatile elements such as caesium and iodine, and metals such as barium, technetium, molybdenum, and ruthenium and the lanthanoids. In addition, the elements plutonium and americium, generated by neutron capture, are present in significant quantities. These products are very diverse in nature and are deposited in various forms. The metals Mo, Tc, Ru, Rh, and Pd are distributed as particles throughout the fuel. Oxides of Rb, Cs, and Ba form precipitates, while those of Sr, Zr, Nb, the lanthanoids, and transuranium elements form solid solutions with the UO_2 fuel. Gas bubbles are also present due to the production of elements such as Kr and Xe.

The spent fuel rods are far more radioactive than the unused rods. On removal from the reactor, these *hot* fuel rods are placed into ponds of water for 10 years or so, to cool down. During this period, many of the radioactive elements decay, as most have short half-lives, i.e.

$$^{142}_{56}\mathrm{Ba} \rightarrow {}^{142}_{57}\mathrm{La} + \beta^- \quad (t_{1/2} = 11 \text{ minutes})$$

After 10 years, the major radioactive materials present are the long-lived isotopes ^{90}Sr ($t_{1/2}$, 28.5 years), ^{137}Cs ($t_{1/2}$, 30.1 years), and 239 Pu ($t_{1/2}$, 24 000 years). At this stage the fuel can be reprocessed to regain uranium and plutonium and to reduce the amount of material that has to be safely stored to manageable amounts. The result is a relatively small amount of high-level waste. In addition, there is a considerable amount of intermediate level waste, which arises from the Zircaloy cases, graphite, stainless steel containers, and components and materials used in the reprocessing.

The major materials problems in radioactive waste disposal are associated with the storage of the high-level waste. Generally waste disposal is broken down into three stages. The initial stage, *immobilisation* of the radioactive isotopes, involves trapping the radioactive isotopes in a stable solid, such as glass, cement, or ceramic. The second stage is to seal the solid into a metal container. Finally, the container must be buried in a geologically stable area, and isolated with an impermeable barrier.

There are materials problems associated with all of these steps. The most widely explored solid for waste immobilisation is borosilicate glass. Unfortunately, glass is damaged by radiation effects, which accelerate devitrification and cause volume changes leading to cracking or erosion. Radiation combined with the heat produced by radioactive decay can enhance chemical reactivity, even in such non-reactive materials such as cement or ceramic oxides, again leading to physical disintegration.

The durability of the canister material also poses a problem. At present the favoured material is stainless steel. The conditions that the canister must tolerate include temperatures of up to 200 °C, water vapour, liquid water, and corrosion products from the immobilising solid. The time scale of thousands of years increases the durability problem enormously. Additionally, the presence of other metallic elements, both inside and outside of the container, can lead to electrochemical corrosion.

The ideal backfill material to isolate sealed canisters is a clay-like substance. This is because clays are able to absorb cations, which are then incorporated in the crystal structure. Thus, the clay acts as a further barrier to dispersal in the event that a storage canister is breached and a radioactive solution is formed. Another advantage is that both cation and water absorption cause clay particles to swell, thus increasing the pressure of the backfilling material and further impeding the movement of solutions containing radioactive ions.

The many problems associated with the safe storage of high-level radioactive waste are an ongoing area of intensive materials research.

Further Reading

A unique insight into the history of the birth and development of atomic physics and radiochemistry can be gained by reading the lectures given by the key Nobel Prize winners, in book form and also at www.nobel.se. Also, of general interest are

Chown, M. (1999). *The Magic Furnace*. London: Jonathan Cape.

Corfield, R. (2001). *Architects of Eternity*. London: Hodder Headline.

Lewis, C. (2000). *The Dating Game*. Cambridge: Cambridge University Press.

The Search for New Heavy Elements:

Clery, D. (2011). Which way to the island? *Science* 333: 1377–1379.

Radioactive Dating:

Hiess, J., Condon, D.J., McLean, N. et al. (2012). $^{238}U/^{235}U$ systematics in terrestrial uranium-bearing minerals. *Science* 335: 1610–1614. (Uranium dating).

Pickering, R., Dirks, P.H.G.M., Jinnah, Z. et al. (2011). *Australopithecus sediba* at 1.977 Ma and implications for the origins of the genus *Homo*. *Science* 333: 1421. and especially the Supplementary information. (Uranium dating).

Pike, A.W.G., Hoffmann, D.L., García-Diez, M. et al. (2012). U-series dating of Paleolithic art in 11 caves in Spain. *Science* 336: 1409–1413.

Stirling, C.H. (2012). Keeping time with Earth's heaviest element. *Science* 335: 1585–1586. (Uranium dating).

Wu, X., Zhang, C., Goldberg, P. et al. (2012). Early pottery at 20,000 years ago in Xianredong Cave, China. *Science* 336: 1696–1700. (Carbon dating).

For an accurate table of particle rest mass see: http://physics.nist.gov/PhysRefData/Compositions/index.html.

Nuclear Reactors:

Lewis, E.E. (2008). *Fundamentals of Nuclear Reactor Physics*. Amsterdam: Academic Press.

Full details of the use of radioisotope power generation in space exploration will be found at www.nasa.gov.

Nuclear Waste:

Burns, P.C., Ewing, R.C., and Navrotsky, A. (2012). Nuclear fuel in a reactor accident. *Science* 335: 1184–1188.

Yoshida, N. and Kandra, J. (2012). Tracking the Fukushima radionuclides. *Science* 336: 1115–1116.

Problems and Exercises

Quick Quiz

14.1 The atomic number of an element defines:

a. The number of protons in a nucleus.

b. The number of neutrons in a nucleus.

c. The number of protons plus neutrons in the nucleus.

14.2 The number of electrons surrounding a neutral atom is the same as the:

a. Mass number.

b. Proton number.

c. Nucleon number.

14.3 Isotopes of an element have the same numbers of:

a. Neutrons.

b. Nucleons.

c. Protons.

14.4 An atomic species that has specified nucleon number and proton number is named:

a. An isotope.

b. A nuclide.

c. A radioisotope.

14.5 When a radioactive element emits an α-particle the daughter element has a nucleon number:

a. 4 less that the parent atom.

b. 2 less than the parent atom.

c. The same as the parent atom.

14.6 When a radioactive element emits a β-particle the daughter element has a nucleon number:

a. 4 less that the parent atom.

b. 2 less than the parent atom.

c. The same as the parent atom.

14.7 The total number of different gaseous elements produced in all the different radioactive series is:

a. One.

b. Two.

c. Four.

14.8 The only non-naturally occurring radioactive decay series, the neptunium series, has a formula:

a. $4x + 1$.

b. $4x + 2$.

c. $4x + 3$.

14.9 The elements with a partly filled 5f electron shell are named the:

a. Lanthanoids.

b. Acinoids.

c. Uranoids.

14.10 During β^+-decay, the particle ejected from the nucleus is:

a. A neutron.

b. An electron.

c. A positron.

14.11 Radioactive nuclei below the band of stability tend to emit:

a. Positrons.

b. Electrons.

c. Neutrons.

14.12 The half-life of strontium-90 is 28.1 years. A person ingests 0.01 g of this isotope which is incorporated into the bones. How long will it take before the level reaches 0.001 25 g?

a. 28.1 years.

b. 56.2 years.

c. 84.3 years.

14.13 The binding energy of a nucleus is:

a. The chemical bonding energy between the protons and neutrons.

b. The energy of interaction between the protons in the nucleus.

c. This mass of the nucleus minus the total masses of the neutrons and protons.

14.14 Fissile materials undergo nuclear fission:

a. Spontaneously.

b. Under bombardment by not very energetic neutrons.

c. Under bombardment by very energetic neutrons.

14.15 The purpose of a moderator in a nuclear reactor is to:

a. Slow the neutrons down.

b. Initiate the chain reaction.

c. Control the energy produced.

14.16 Thermal reactors for the production of nuclear power use a fuel of:

a. Uranium.

b. Uranium dioxide.

c. Uranium hexafluoride.

14.17 The purpose of the control rods in a nuclear reactor is to:

a. Slow down neutrons.

b. Generate neutrons.

c. Absorb neutrons.

14.18 A fast breeder reactor produces:

a. Uranium.

b. Plutonium.

c. Neptunium.

14.19 Fusion reactions for the production of power envisage utilisation of the reaction of:

a. Helium to form hydrogen.

b. Hydrogen to form helium.

c. Lithium-6 to form hydrogen.

14.20 The process that powers the Sun and stars is:

a. Fusion.

b. Fission.

c. A mixture of fusion and fission.

14.21 Intermediate level nuclear waste constitutes how much of the total:

a. Less than 1%.

b. Approximately 20%.

c. Approximately 80%.

Calculations and Questions

14.1 Write the nuclear symbol, and the number of protons, neutrons and nucleons in:

a. rhenium-166;

b. barium-140;

c. oxygen-18;

d. boron-14.

14.2 Write the nuclear symbol, and the number of protons, neutrons, and nucleons in:

a. mercury-181;

b. iridium-169;

c. iodine-117;

d. zirconium-98.

14.3 Write nuclear equations for the decay of the following radioactive isotopes; the particles emitted in the decay are given in

brackets:

a. $^{27}_{14}\mathrm{Si}$ (positron);

b. $^{28}_{13}\mathrm{Al}$ (electron);

c. $^{24}_{11}\mathrm{Na}$ (electron);

d. $^{17}_{9}\mathrm{F}$ (positron).

14.4 Write nuclear equations for the decay of the following radioactive isotopes; the particles emitted in the decay are given in brackets:

a. $^{24}_{8}\mathrm{O}$ (electron);

b. $^{47}_{23}\mathrm{V}$ (positron);

c. $^{32}_{15}\mathrm{P}$ (electron);

d. $^{39}_{20}\mathrm{Ca}$ (positron).

14.5 Write nuclear equations for the decay of the following radioactive isotopes; the particles emitted in the decay are given in brackets:

a. $^{243}_{100}\mathrm{Fm}$ (alpha);

b. $^{241}_{95}\mathrm{Am}$ (alpha);

c. $^{241}_{94}\mathrm{Pu}$ (electron);

d. $^{237}_{92}\mathrm{U}$ (electron).

14.6 Determine the half-life of a radioactive isotope from the variation of the number of radioactive disintegrations observed, in counts per minute, over a period of time, given in the following table.

Time (min)	**0**	**2**	**4**	**6**	**8**	**10**	**12**	**14**
Activity (counts per min)	3160	2512	1778	1512	1147	834	603	579

14.7 A sample of radioactive sodium-24 with a half-life of 15.0 hours is used to measure the diffusion coefficient of Na in NaCl.

How long will it take for the activity to drop to 0.1 of its original activity?

14.8 What is the specific activity of radium-226, which has a half-life of 1600 years?

14.9 What is the specific activity of plutonium-241, which has a half-life of 14.35 years?

14.10 What is the specific activity of neptunium-233, which has a half-life of 36.2 minutes?

14.11 A purified sample of a radioactive compound is found to have an activity of 1365 counts per minute at 10 a.m. but only 832 counts per minute at 1 p.m. What is the half-life of the sample?

14.12 A 250 mg piece of carbon from an ancient hearth showed 1530 carbon-14 disintegrations in 36 hours. 250 mg of fresh carbon from charcoal gave 8280 disintegrations in the same time. What is the date of the carbon sample? The half-life of carbon-14 is 5.73×10^3 years.

14.13 A cellar of dimensions 2 m × 3 m × 2.5 m is found to show an activity of 0.37 Bq dm^{-3}, due to the presence of radon.

a. How many nuclei decay per minute in the cellar?

b. How long will it take for the activity to fall to 0.015 Bq, if no more radon leaks into it? The half-life of radon is 3.8 days.

14.14 The suggested upper limit for radon concentration in a building is equivalent to an activity of 200 Bq m^{-3}.

a. How many radon atoms are needed per dm^3 of air to give this figure?

b. How long will it take for the activity to fall to one tenth of this value in a sealed room? The half-life of radon is 3.8 days.

14.15 In 1986 a nuclear reactor at Chernobyl exploded, depositing caesium-137 over large areas of northern Europe. An initial activity of 1000 Bq m^{-2} on vegetation was found over parts of Scotland.

a. How many grams of caesium-137 were deposited per square metre of vegetation?

b. What was the activity after 10 years? The half-life of caesium-137 is 30.1 years.

14.16 Calculate the binding energy per nucleon for $^{4}_{2}\mathrm{He}$. The masses are $^{1}_{1}\mathrm{H}$, 1.0078 u; $^{1}_{0}\mathrm{n}$, 1.008 7 u; $^{4}_{2}\mathrm{He}$, 4.0026 u.

14.17 Calculate the binding energy per nucleon for $^{23}_{12}\mathrm{Mg}$. The masses are $^{23}_{12}\mathrm{Mg}$, 22.9941 u; $^{1}_{1}\mathrm{H}$, 1.0078 u; $^{1}_{0}\mathrm{n}$, 1.0087 u.

14.18 Calculate the binding energy per nucleon for $^{34}_{16}\mathrm{S}$. The masses are $^{34}_{16}\mathrm{S}$, 33.967 865 u, $^{1}{}_{1}\mathrm{H}$, 1.007 8 u; $^{1}_{0}\mathrm{n}$, 1.008 7 u.

14.19 On decay, an atom of radium-226 emits one α-particle and is converted into an atom of radon-222. A quantity of radium-226 produced 4.48×10^{-6} dm^3 of helium at 273 K and 1 atm pressure. Determine:

a. the mass of radium-226 that decayed,

b. the mass of radon-222 produced, if no radon decays.

c. One mole of helium gas occupies 22.4 dm^3 at 273 K and 1 atm pressure.

14.20 Calculate the energy released per mole in the fission reaction:

$$^{235}_{92}\mathrm{U} + ^{1}_{0}\mathrm{n} \rightarrow ^{102}_{42}\mathrm{Mo} + ^{128}_{50}\mathrm{Sn} + 6\,^{1}_{0}\mathrm{n}$$

The masses are $^{235}_{92}\mathrm{U}$, 235.0439 u; $^{102}_{42}\mathrm{Mo}$, 101.910 25 u; $^{128}_{50}\mathrm{Sn}$, 127.910 47 u; $^{1}_{0}\mathrm{n}$, 1.0087 u.

14.21 Energy generation in the Sun is by the reaction of hydrogen to form helium. One reaction is:

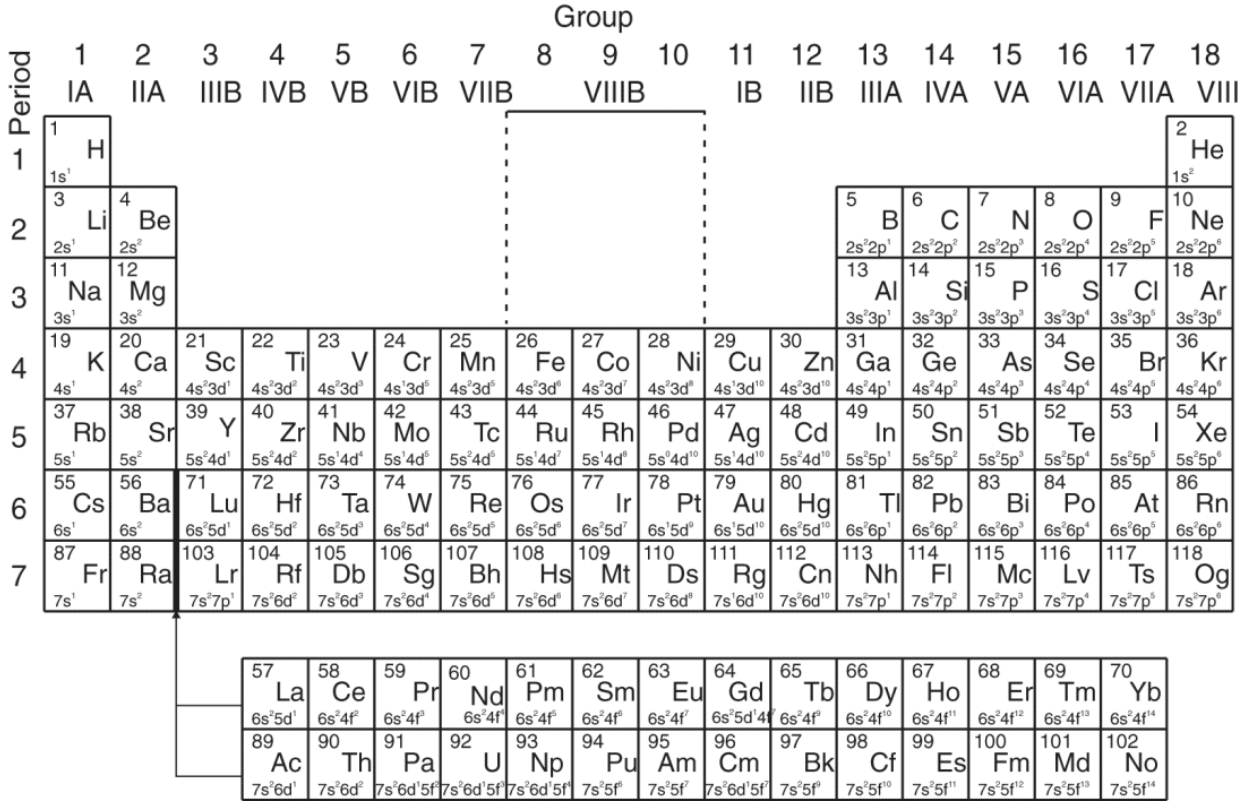
$$4\,^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He} + 2\,^{0}_{+1}\mathrm{e}$$

a. Calculate the energy liberated per fusion reaction.

b. It is estimated that the sun produces approximately 10^{26} J s^{-1}. How many fusion reactions per second are required for this?

The masses are ${}_{1}^{1}\mathrm{H}$, 1.0078 u; ${}_{2}^{4}\mathrm{He}$, 4.0026 u; ${}_{+1}^{0}\mathrm{e}$, 5.486×10^{-4} u.

Appendix A
Periodic Table



Appendix B
Energy Levels and Terms of Many-Electron Atoms

B.1 Derivation of Atomic Terms

The Russell–Saunders terms of a many-electron atom are derived by adding the individual spin quantum numbers of the electrons to yield a total spin quantum number S, and adding the individual orbital angular momentum quantum numbers of the electrons to give a total orbital angular momentum quantum number, L. For example, the total spin angular momentum quantum number, $S(2)$, for two electrons is given by adding the individual quantum numbers, thus:

$$S(2) = (s_1 + s_2), (s_1 + s_2 - 1), \ldots |s_1 - s_2|.$$

As s_1 and s_2 are both equal to ½,

$$S(2) = 1 \text{ or } 0.$$

Note that this is a maximum number of values. If two electrons are spin paired, then only the value 0 applies. If the electrons are in different orbitals, say, an s and a p orbital, they can have parallel or antiparallel spins, making both 1 and 0 possible.

In order to obtain the value of S for three electrons, $S(3)$, the value for two electrons, $S(2)$, is added to the spin quantum number of the third electron, s_3, thus:

$$S(3) = (S(2) + s_3), (S(2) + s_3 - 1), \ldots |S(2) - s_3|$$

Both of the values for $S(2)$ are permitted, so we obtain:

$$\text{(a)}, \quad S(2) = 1, \quad S(3) = 1 + {}^1\!/_2,$$

$$1 + {}^1\!/_2 - 1 = {}^3\!/_2, {}^1\!/_2$$

$$\text{(b)}, \quad S(2) = 0, \quad S(3) = 0 + {}^1\!/_2$$

$$S(3) = {}^3\!/_2, {}^1\!/_2$$

This procedure is called the *Clebsch–Gordan rule*. It is used to obtain the S values for increasing numbers of spins. It will be found that for an even number of electrons, S values are integers, and for an odd number of electrons, S values are half-integers.

The total angular momentum quantum number, L, is obtained in a similar fashion. For two electrons with individual angular momentum quantum numbers l_1 and l_2, the total angular momentum quantum number, $L(2)$, is:

$$L(2) = (l_1 + l_2), (l_1 + l_2 - 1), \ldots |l_1 - l_2|$$

In the case of three electrons, the Clebsch–Gordan rule is applied thus:

$$L(3) = (L(2) + l_3), (L(2) + l_3 - 1), \ldots |L(2) - l_3|$$

using every value of $L(2)$ obtained previously. Values of the quantum number L are given letter symbols as described.

For example, the terms arising from the two p electrons on carbon, C, with $l_1 = l_2 = 1$, are obtained in the following way.

$$S = (1/2 + 1/2), \ (1/2 - 1/2) = 1, \ 0$$

$$2S + 1 = 3 \text{ or } 1$$

$$L = (1 + 1), \ (1 + 1 - 1), \ (1 - 1)$$

$$= 2, \ 1, \ 0 \ (\text{D}, \ \text{P}, \ \text{S})$$

The total number of possible terms for the two p electrons is given by combining these values. The possible terms for two p electrons are therefore:

$$^3\text{D}, \ ^3\text{P}, \ ^3\text{S}, \ ^1\text{D}, \ ^1\text{P}, \ ^1\text{S}$$

Not all of these possibilities are allowed for any particular configuration, because the Pauli exclusion principle limits the number of electrons in each orbital to two with opposed spins. When this is taken into account, the allowed terms are:

$$^3\text{P}, \ ^1\text{D}, \ ^1\text{S}$$

Similarly, an atom with two d electrons, with a configuration of $3d^2$ will again have:

$$S = (1/2 + 1/2),\ (1/2 - 1/2) = 1,\ 0$$

The possible values of L are obtained by using the values $l_1 = l_2 = 2$, to give:

$$L = (2 + 2),\ (2 + 2 - 1),\ (2 + 2 - 2),$$
$$(2 + 2 - 3),\ (2 - 2) = 4,\ 3,\ 2,\ 1,\ 0$$
$$(\mathrm{G},\ \mathrm{F},\ \mathrm{D},\ \mathrm{P},\ \mathrm{S})$$

Combining all of these gives ^3G, ^3F, ^3D, ^3P, ^3S, ^1G, ^1F, ^1D, ^1P, ^1S

Taking into account forbidden configurations gives the allowed terms as:

$$^3F,\ ^3P,\ ^1G,\ ^1D,\ ^1S$$

The actual energies of the terms are difficult to obtain simply, and they must be calculated using quantum mechanical procedures. However, the lowest energy term, the ground state term, is easily found, as follows.

B.2 The Ground State Term of an Atom

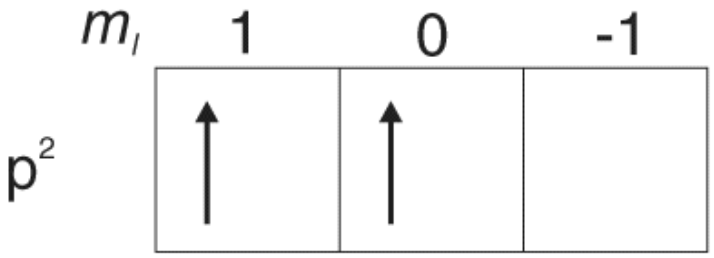
The lowest energy term, the ground state term, can be found using Hund's first and second rules.

1. The term with the lowest energy has the highest multiplicity, equivalent to the highest total spin quantum number, S.
2. For terms with the same value of multiplicity, the term with the highest value of L is lowest in energy.

There is a simple method of determining the ground state of any atom or ion. The procedure is:

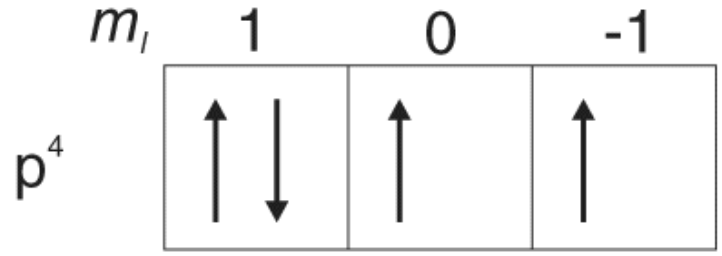
1. Draw a set of boxes corresponding to the number of orbitals available. For a p electron, this is three.
2. Label each box with the value of m_l, highest on the left and lowest on the right.
3. Fill the boxes with unpaired electrons, from left to right. When each box contains one electron, start again at the left.
4. Sum the m_s values of each electron, +1/2 or −1/2. This is equal to the maximum value of S.
5. Sum the m_l values of each electron to give a maximum value of L.
6. Write the ground term ${}^{2S+1}L$.

Using this technique, the ground term of $2p^2$ and $2p^4$ configurations is ${}^3P$, as follows:



$S = 1/2 + 1/2 = 1; \quad 2S + 1 = 3$
$L = 1 + 0 = 1$

Term ^3P



$S = 1/2 + 1/2 \ + 1/2 - 1/2 = 1; \quad 2S \ + 1 = 3$
$L = 1 + 0 + -1 + 1 = 1$

Term ^3P

B.3 Energy Levels of Many Electron Atoms

The term symbol does not account for the true complexity found in most atoms. This arises from the interaction between the spin and the orbital momentum (spin-orbit coupling), that is ignored in Russell–Saunders coupling. For this the quantum number, J, described above, is needed. It is given by:

$$J = (L + S), (L + S - 1), \ldots |L - S|$$

where $|L - S|$ is the modulus (absolute value, regardless of whether + or −). Thus, the term $^3\mathrm{P}$ has J values given by:

$$J = (1 + 1),\ (1 + 1 - 1), \ldots (1 - 1) = 2,\ 1,\ 0$$

The new quantum number is incorporated, as a subscript to the term, now written $^{2S+1}L_J$ and this is no longer called a term symbol, but a *level*. Each value of J represents a different energy level. It is found that a singlet term always gives one energy level, a doublet, two, a triplet three, and so on. Thus, the ground state term $^3\mathrm{P}$ is composed of three levels $^3\mathrm{P}_0$, $^3\mathrm{P}_1$, and $^3\mathrm{P}_2$. The separation of these energy levels is controlled by the magnitude of the interaction between L and S. Hund's third rule allows the values of J to be sorted in order of energy. The level with the lowest energy is that with lowest J value if the valence shell is *up to half full*, and that with the highest J value if the valence shell is *more than half full*.

Index

Note: Page numbers in *italic* refer to figures, page numbers in **bold** refer to tables.

3D printing, sintering 196, 199, 227

a

accelerating oxidation, metal oxide film microstructure 223
addition polymers 134–135
additive manufacturing, sintering 199
additive primary colours 459
adiabatic cooling, magnetocaloric effect 533–534
advection diffusion 175
aerogels 40
aggregate 153
air electrode 251
aliovalent substitution 94
allotropes 53–54
alloying 301
alloys
  binary alloys 214
  high temperature oxidation 225
  internal oxidation 225
  Ni–Ti alloys 211–212
  shape-memory alloys 212–214
  substitutional alloy formation 112–114
alloy solid solutions, metals 112–115
alpha (α) particles 545
aluminosilicates 120
ambipolar diffusion, high temperature oxidation of metals 224
amorphous solids 40
amphiboles 120, *121*

anion packing, crystal structures 84–86
annealing 44
anodes, galvanic cells 239–241
antiferromagnetic compounds 370–371
antiferromagnetic materials, superexchange 386–387
antiphase boundaries (APB's), crystal structures 98, 99
Arrhenius equation 171, *172*
  impurities effect 179
  ionic conductivity 183
  point defects 178–180
atactic polymers 141
atomic energy levels 3–6, 8–11
atomic magnetism 368–369
atomic structure *see* electron structure of atoms
atomistic simulation 393
  ionic bonding 14–15
attenuation 485–486
  Beer–Lambert–Bouguer law 485
  Beer–Lambert Law 486
  Beer's law 485
  Lambert's law 485
  linear absorption coefficient 485
  linear attenuation coefficient (extinction coefficient) 485
  molar (decadic) attenuation coefficient 486
  (Napierian) linear absorption coefficient 485
  (Napierian) linear attenuation coefficient 485
  (Napierian) linear scattering coefficient 485

attenuation in glass fibres 493–494
austenite finish temperature 209
austenite–martensite transition 207–210
austenite start temperature 209
auxetic materials 310–312, *313*

b

Bain model 207–208
band gap 406
  band gap measurement 412–413
bandpass filters, reflection 483
bands in ionic and covalent solids, metallic bonding 27–28
Bardeen–Cooper–Schrieffer (BCS) theory, superconductivity 431–432, 434–435

batteries 247–255
see also galvanic cells
fuel cells 247, 250–252
alkaline fuel cells (AFC) 250, *251*
direct methanol conversion fuel cells (DMCFC) 251
molten carbonate fuel cells (MCFC) 250, *251*
phosphoric acid fuel cells (PAFC) 250, *251*
proton exchange membrane fuel cells (PEMFC) 250, *251*
primary batteries 248–250
carbon-zinc batteries 248, *249*
'dry' and alkaline primary batteries 248, *249*
lithium–air batteries 249–250
lithium-ion primary batteries 249
primary cells 247
secondary batteries 252–255
dual-ion batteries 254–255
lead-acid batteries 252–253
lithium-ion batteries 253–254
secondary cells 247
Beer–Lambert–Bouguer law 485
Beer–Lambert Law 486
Beer's law 485
Berkovich indenter 312–314
beta (**β**)-alumina oxides, superionic conductors 185–188
beta (β) particles 545
Bethe–Slater diagram 381
bi-layer graphene, superconductivity 444–445

binary alloys, equilibrium solidification of simple binary alloys 214
binary phase diagrams *see* two-component (binary) systems
biomolecular materials, piezoelectrics 342
Biot–Fourier equation 518
black bodies 460–461
black body emission spectrum 460–461
Bloch functions 24–25
Bloch wall 383, *384*
body-centred cubic (bcc, A2) structure 70
Bohr magneton 368–369
Boltzmann statistics 515–516
bonding *see* covalent bonding; ionic bonding; metallic bonding; weak chemical bonds
Bose–Einstein statistics 516
bosons 516
bound magnetic polaron 395–396
Bragg's Law 488
Bragg stacks 490
Bravais lattices, crystal systems 66–68
breakaway oxidation, metal oxide film microstructure 223
Brillouin function 377
Brillouin zones 26, 27
Brinell hardness number (BHN) 283
brittle materials
  deformation and fracture 291–294, 297–299
  ductile to brittle transitions 297
Buckingham potential 15

buckminsterfullerene 41, *42*
buckyballs 41, *42*
bulk modulus, elastic moduli 289–290
burnt lime 153

c

calcia-stabilised zirconia 246–247
capacitors, dielectrics 327–329
carbon dating 546, 555–557
cathodes, galvanic cells 239–241
cell potential, galvanic cells 241, 243
cement 153–156
cemented carbide materials 152–153
cement paste 153
ceramic magnets 389–390
ceramic–matrix composites 153

ceramics 118–126
diversity 118–119
nomenclature 118–119
non-silicate ceramics 122–125
preparation and processing 125–126
principal properties 126
silicate structures 119–122, **124**
aluminosilicates 120
amphiboles 120, *121*
chrysotile asbestos 120, *123*
feldspars 120
kaolinite 120, *123*
pyroxenes 120
talc 120, *123*
ultramarines 121
zeolites 121
chain growth, polymers 144–147
chemical adsorption (chemisorption), metal oxide film microstructure 222
Chernobyl nuclear accident 563–564
chrysotile asbestos 120, *123*
circularly polarised light 456–457
Clausius–Mossotti equation 331, 333–334
CMR (colossal magnetoresistance), magnetic defects 398–399
coactivators 463
cobaltites, magnetic defects 396–399
Coble creep 308–309

COFs (covalent organic frameworks) 147–149, *150*, 151
cold working 302
colossal magnetoresistance (CMR), magnetic defects 398–399
colour and appearance, light 459–460
compensated semiconductors 407
compliance, solids 286
composite materials 152–156, 315–317
  cement 153–156
  concrete 153–156
  elastic modulus of fibre reinforced composites 315–316
  fibre-reinforced materials 152–153
  Young's modulus 315–317
compressional waves 290–291
compressive strength, solids 277, *278*, 279
concentration dependence 243
concentration profiles 168, 170
conchoidal fracture 291
concrete 153–156
condensation polymers 135
conducting polymers 427–431
conduction band (CB) 406
conductivity, electronic *see* electronic conductivity
congruent melting 52
contact potential 419–421
continuous fibre composite 315
continuous transitions 201

contraction
thermal contraction 524–527
zero thermal contraction materials 526–527
Coolidge process 297
Cooper pairs 434–435
‘copper-bottomed’ 258
coring 215
correlation factors
ionic conductivity 184
self-diffusion 180–181, 184
corrosion 255–260
‘copper-bottomed’ 258
crevice corrosion 259, *260*
direct corrosion, high temperature oxidation 220–222
dissimilar metal corrosion 257–259
electrochemical reaction 255–256
galvanizing 258
Pourbaix diagrams 264–268
reaction of metals with water and aqueous acids 256–257
single metal electrochemical corrosion 259–260
corrosion fatigue 305
Coulomb interaction 13
Coulomb repulsion 380
covalent bonding 15–21, 28, 29, 31
bond energies 18–21
bond geometry 15–18
covalent organic frameworks (COFs) 147–149, *150*, 151

creep 305–309
Coble creep 308–309
creep rupture 309
diffusion creep 308
dislocation creep 307–308
elongation 305–309
Nabarro–Herring creep 309
power law creep 307–308
primary creep 307
secondary creep 307
steady-state creep 307
tertiary creep 309
cross-linking 44–45
polymers 137–138, 144
crystal density 80–82
density estimation 80–81
density of crystals with a variable composition 81–82
density of NaCl 81
crystal geometry 65–68
crystal lattices 66–68
crystal systems 65–66
symmetry and crystal classes 68
crystal glasses 68
crystal growth 44
crystal lattices 66–68
crystal-like structures, diffraction 488–490

crystalline silicates 118–126
see also ceramics
crystalline solids 37, *38*
crystallographic shear (CS) plane, crystal structures *98*, 99
crystal planes and directions 74–79, **80**
directions 78–79
hexagonal crystals 76–77, *78*
interplanar spacings 79, *80*
Miller–Bravais Indices 76–77, *78*
Miller Indices 74–77, *78*
crystals, relative permittivity 332–334

crystal structures 69–74
antiferroelectric crystals 343–345
atomic coordinates 69
body-centred cubic (bcc, A2) structure 70
diamond structure 71
dislocations 95–96
edge dislocations 95–96
plastic deformation 95
screw dislocations 95–96, 97
face-centred cubic (fcc, A1) structure 70
ferroelectric crystals 343–345
graphite structure 71
halite structure 71–72
hexagonal close-packed (hcp, A3) structure 70
lattice parameters 74
linear defects 95–96
metals 109–110
perovskite structure 72–73
planar defects 96–99
antiphase boundaries (APB's) 98, 99
crystallographic shear (CS) plane 98, 99
grain boundaries 97–99
twin boundaries 98, 99

point defects 87–95
crystals of the elements 88–89
Frenkel defects 90–91
non-stoichiometric compounds 91–93
point defect notation 93–95
Schottky defects 90–91
solid solutions 89–90
precipitates 99
spinel structure 73–74
structural relationships 82–87
anion packing 84–86
cubic close-packing (ccp) 84–86
hexagonal close-packing (hcp) 82–84
ionic structures in terms of anion packing 84–86
polyhedral representations 86–87, *88*
sphere packing 82–84
unit cells 65–68, 69–71
Vegard's law 74
volume defects: precipitates 99
crystal symmetry, piezoelectric effect and pyroelectric effect 335–337
crystal systems 65–66
Bravais lattices 66–68
cubic close-packing (ccp), crystal structures 84–86
cubic spinel ferrites, ferrimagnetic materials 387–388
cuprate high temperature superconductors, superconductivity 437–444

Curie point 347–349, *349*, 370
Curie temperature 347–349, *349*, 370
Curie–Weiss behaviour 347, *348*, *349*
Curie–Weiss constant 370–371

d

Daniell cells 239, *240*, 242, 243
data recording, nanostructures 391
dating, radioactive 546, 555–557
Debye temperature 516–517
deformation and fracture 291–304
  brittle fracture 291–294
  brittle materials 297–299
  computation of deformation and fracture 303–304
  ductile materials 297–299
  plastic deformation of metals 294–297
  plastic deformation of polymers 299–301
  strengthening 301–303
dehydration 45
dendritic growth 44
density functional theory 31
density of states function 24
depletion region 421
devitrification 45
diamond, hardness 284
diamond structure 71
dielectric mirrors 483

dielectrics 327–334
capacitors 327–329
polarisability 330–332
polarisation 329–330
relative permittivity 327–330
relative permittivity of crystals 332–334
diffraction 486–493
crystal-like structures 488–490
diffraction by an aperture 486–487
diffraction gratings 487–488
holograms 490–493
formation 490–491
hologram recording media 492
diffuse reflection 459

diffusion 167–184
advection diffusion 175
ambipolar diffusion, high temperature oxidation of metals 224
diffusion coefficients 167–169, 171–175
impurities effect 171
ionic conductivity 183–184
temperature variation 171
diffusion equations 168–169, **170**
diffusion in solids 175–176
diffusion profiles 168, 170
drift diffusion 175
Einstein (or Einstein–Smoluchowski) diffusion equation 174
exchange diffusion 175–176
film growth via diffusion, high temperature oxidation of metals 223–225
Frenkel defects
diffusion in crystals 179, **180**
diffusion in solids 176
interstitialcy diffusion 175, 176
interstitial diffusion 175, 176, 181
non-steady-state diffusion 169–171
random walk diffusion 171–175
ring diffusion 175–176
Schottky defects
diffusion in crystals 179, **180**
diffusion in solids 176

self-diffusion 167–169, **172**, 174–178
correlation factors 180–181
in crystals 178
interstitial diffusion 181
in one dimension 176–178
in solids 175–176
steady-state diffusion 169–171
tracer diffusion 167–169
vacancy diffusion 175, 176
diffusion creep 308
diffusons 522
diluted magnetic semiconductors 395
direct lattices, crystal systems 66–68
discontinuous fibre composite 315
dislocation creep 307–308
dislocations, crystal structures 95–96
dispersion and optical fibre design 494–496
displacive transitions 201–203
Jahn–Teller effect 202
doping 413, 425–426, 429
drift diffusion 175
drift mobility 410
ductile materials
deformation and fracture 297–299
ductile to brittle transitions 297
ductile solids 280
ductility, metals 301

dye-sensitised solar cells 497–499
dynamic equilibrium, binary alloys 214

e

Einstein (or Einstein–Smoluchowski) diffusion equation 174
Einstein coefficient for absorption of radiation 470
Einstein coefficient for spontaneous emission 469–470
Einstein coefficient for stimulated emission of radiation 470
elastic moduli 277, 280, 282, 285–291
  bulk modulus 289–290
  Lamé modulus 289–290
  longitudinal or axial modulus 289
  Poisson's ratio **287**, 288–289
  relationships between 290
  shear modulus 289–290
  ultrasonic waves in elastic solids 290–291
  Young's modulus 286–288
elastic modulus of a two-phase system 316–317
elastic modulus of fibre reinforced composites 315–316
elastic region 280
elastomers 134, 142–144
  strength 280, 281
electrolysis 260–264
  amount of product produced 262–263
  electrolytic cells 260–261
  electrolytic preparation of titanium 263–264
  electroplating 261–262
  Faradays laws of electrolysis 262–263
  FFC Cambridge Process 263–264

electromagnetic spectrum 455, *456*
electronegativity 20, 29, 112
electron–electron repulsion 423–425
electron flow, galvanic cells 239–241

electronic conductivity 405–447
conducting polymers 427–431
insulators 405–407
metal–insulator transitions 422–427
electron–electron repulsion 423–425
metals and insulators 422–423
modification of insulators 425–426
transparent conducting oxides (TCOs) 426–427
metals 405–411
resistivity 410–411
semiconductors 405–407, 411–422
band gap 406
band gap measurement 412–413
characterisation 416–419
extrinsic semiconductors 407, 413–416
intrinsic semiconductors 405–406, 411–412
p–n junction diode 419–422

superconductivity 431–445
Bardeen–Cooper–Schrieffer (BCS) theory 431–432, 434–435
bi-layer graphene 444–445
cuprate high temperature superconductors 437–444
effect of magnetic fields and current 432–434
Josephson junctions 435–437
lanthanum cuprate, La_2CuO_4 437–438
neodymium cuprate, Nd_2CuO_4 438–439
perovskite-related structures and series 440–444
SQUID (superconducting quantum interference device) 436–437
superconductors 431–432
transition temperature 431
type II superconductors 433–434
type I superconductors 433
unconventional superconductors 431–432
yttrium barium copper oxide, $YBa_2Cu_3O_7$ 439–440
electronic polarisability 330
electron structure of atoms 3–11
atomic energy levels 3–6, 8–11
electron configuration 8–9
electron spin 8–9
hydrogen 3–4
many electron atoms 4–6
energy levels and terms 573–575
orbital shapes 6–8

elongation
  creep 305–309
  gradual elongation 305–309
  polymers 299–300, *301*
energy bands
  metallic bonding 21–22, 24–28
  molecular orbitals 21–22
energy conversion 496–501
  dye-sensitised solar cells (DSSCs) 497–499
  Grätzel cells 497–499
  perovskite solar cells 499–501
  photoconductivity 496–497
  photovoltaic solar cells 496–497
engineering (or nominal) strain 279
epoxy resins 137–138
equilibrium heat treatment of steel in the Fe–C phase diagram 218–219, *220*
equilibrium solidification of simple binary alloys 214
eutectic point 49–51, 54
  solidification in systems containing a eutectic point 216–218
eutectoid point 54
exchange–correlation energy 31
exchange diffusion 175–176
exchange of carriers 421
exhaustion range 407

expansion 522–524
thermal expansion 522–524, 525
interatomic potentials 523–524
extrapolated Curie temperature 347, *349*
extrinsic ferroelectricity 356
extrinsic semiconductors 407, 413–416

f

face-centred cubic (fcc, A1) structure 70
Faraday constant 263
Faradays laws of electrolysis 262–263
fast breeder reactors, nuclear energy 561–562
fatigue 304–305
  corrosion fatigue 305
  endurance limit 304
  fatigue failures 304–305
  fatigue limit 304
  stress cycles 304–305
feldspars 120
Fermi–Dirac statistics 24
Fermi energy 407
fermions 516
Fermi surface 24, 407–412, 415–416, *417*, 419, *420*
ferrimagnetic materials 371, 387–391
  ceramic magnets 389–390
  cubic spinel ferrites 387–388
  garnet structure ferrites 388–389
  hexagonal ferrites 389–390
ferroelectric nanoparticles 354–356

ferroelectrics 334–335, 342–357
antiferroelectric crystals 343–345
Curie temperature 347–349, *349*
Curie–Weiss behaviour 347, *348*, *349*
extrinsic ferroelectricity 356
ferroelectric crystals 343–345
hysteresis and domain growth 345–346, *347*
ferroelectric nanoparticles 354–356
ferroelectric transitions 343
ferromagnetic transitions 343
flexoelectricity 356–357
hydrogen bonds 347–349, *350*
improper ferroelectricity 356
modification of properties 352–354
polar groups 349–350, *351*
relaxor ferroelectrics 354, *355*
superlattices 354–356
temperature dependence 347
thin films 354–356
transition-metal cations 350–352

ferromagnetic materials 379–385
domains 382–383, *384*
exchange energy 380–382
ferromagnetism 379–380
hard and soft magnetic materials 382–383, 385
hysteresis 384–385
magnetic domains 382–383, *384*
saturation magnetisaation 379
Weiss domains 382–383, *384*
FFC Cambridge Process 263–264
fibre Bragg gratings 490
fibre optics 493–496
attenuation in glass fibres 493–494
dispersion and optical fibre design 494–496
graded index fibre 495
modal (or intermodal) dispersion 494
monomode fibres 495
optical amplification 495–496
stepped index multimode fibres 494–495
fibre-reinforced materials 152–153
Fick's first law 174
Fick's second law 168–169
finite element analysis 304
first-order phase transitions 200–201, 517–518
flexoelectricity 356–357
flexural strength, solids 277, *278*
fluorescent lamps 463–464

fluxon lattice (flux lattice, or vortex lattice) 433, 434
forward bias 421
fracture *see* deformation and fracture
Fraunhofer diffraction 486
free electron gas, metallic bonding 21, 22–24
free electron g-value 368
free radical polymerisation 145–146
Frenkel defects
  crystal structures 90–91
  diffusion in crystals 179, **180**
  diffusion in solids 176
  ionic conductivity 183
Fresnel diffraction 486
fuel cells 247, 250–252
  alkaline fuel cells (AFC) 250, *251*
  direct methanol conversion fuel cells (DMCFC) 251
  molten carbonate fuel cells (MCFC) 250, *251*
  phosphoric acid fuel cells (PAFC) 250, *251*
  proton exchange membrane fuel cells (PEMFC) 250, *251*
fuel electrode 251
fuel for space exploration, nuclear energy 561
Fukushima Daiichi nuclear accident 563–564
fullerenes 41, *42*
fusion and solar cycles, nuclear energy 562–563

g

galvanic cells 239–247
  see also batteries
  anodes 239–241
  cathodes 239–241
  cell potential 241, 243
  chemical analysis 243–247
    ion selective electrodes 244–246
    oxygen sensors 246–247
    pH meters 243–245
  electron flow 239–241
  standard electrode potentials 241–242
galvanizing 258
gamma (γ) rays 545
garnet structure ferrites 388–389
gauge length 279
gels 134
giant magnetocaloric effects 534–535
Gibbs energy 88–91, 191, 197, 200–201, 221, 243
Gibbs Phase Rule 47–48
Gladstone–Dale formula 477
glass-ceramics 132–133
glass deformation 129–131

glasses 37–38
crystal glasses 68
metallic glasses 115–116
silicate glasses 126–133
graded index fibre, fibre optics 495
graded refractive index (GRIN) coating, nanostructures 505
grain boundaries, crystal structures 97–99
grain growth, sintering 198
graphene 41
bi-layer graphene 444–445
graphite fluoride 309–310
graphite structure 71, 309–310
Grätzel cells 497–499
Griffith equation 294
Griffith's flaws 131, 293–294
Griffith theory of brittle fracture 293–294
gutta-percha *143*, 144
gypsum 154, 155, 156

h

half-life
  nuclear decay 554–555
  radioactivity **552**
half-reactions, galvanic cells 240, 241–244
halite structure 71–72
Hall coefficient 418–419
Hall effect 417–418
Hall mobility 419
Hall voltage 418
halophosphate lamps 463
hardness numbers 283
hardness of solids *see* strength and hardness of solids
Haven ratio 184
HDPE (high-density polyethylene) 140
heat capacity 515–518
  phase transitions 516–518
  theories 515–517
heat equation 518
Heikes equation 530–531
hexagonal close-packed (hcp, A3) structure 70
  crystal structures 82–84
hexagonal ferrites 389–390
high-carbon steel 54
high-density polyethylene (HDPE) 140
high thermal conductivity materials 519, **520**

holograms 490–493
formation 490–491
hologram recording media 492
rainbow holograms 490, *491*
Hooke's law 280, 286, 287, 289
hot pressing, sintering 196
Hubbard subbands 423
Hume-Rothery solubility rules 112
Hund's first rule 8–9, 10, 381
Hund's second rule 10
Hund's third rule 10, 575
hydrogen, atomic structure 3–4
hydrogen bonds 29
ferroelectricity 347–349, *350*
hysteresis 209, 282–283
ferromagnetic materials 384–385
hysteresis and domain growth, ferroelectric crystals 345–346, *347*

i

impact strength, solids 277, *278*

improper ferroelectricity 356

incandescence 460–461

indentation hardness 283–284, *285*

indium tin oxide (ITO) 426–427

infrared 456

insulating bands 423–424

interference

- light waves 457, *458*
- optical interference filters 483
- SQUID (superconducting quantum interference device) 436–437

intergranular fracture 291

internal oxidation, alloys 225

interplanar spacings 79, *80*

interpolation 92

interstices 84–85, 88–91, **94**, 95–96, *98*

interstitialcy diffusion 175, 176

interstitial diffusion 175, 176, 181

interstitial positions 84–85, 88–91, **94**, 95–96, *98*

interstitial sites 84–85, 88–91, **94**, 95–96, *98*

interstitial solid solutions 112

intrinsic semiconductors 405–406, 411–412

ion–dipole interaction energies 30

ionic bonding 12–15
atomistic simulation 14–15
ionic radii 12–13
ionic size 12–13
lattice energies 13–14
ionic conductivity 181–184
Arrhenius equation 183
diffusion coefficients 183–184
Frenkel defects 183
Schottky defects 183
superionic conductors 184–189
ionic polarisability 330
ionic radii 12–13
ionic size and bonding 12–13
ionic structures in terms of anion packing, crystal structures 84–86
ion selective electrodes 244–246
iso-electronic crystals, hardness 284
isoprene 143
iso-strain condition 315
isotopes and nuclides, radioactivity 546
isotropic materials, magnetic flux 367

j

Jahn–Teller effect 202
jellium model 31
Josephson junctions, superconductivity 435–437
Joule heating 410–411

k

kaolinite 120, *123*
Kevlar 152
Knoop hardness number (KHN) 284
Kröger–Vink notation, point defects 93–94

l

Lambert's law 485
lamellae, polymers 140, *141*
Lamé modulus, elastic moduli 289–290
laminates 152, *153*
Langevin function 377
lanthanum cuprate, La_2CuO_4, superconductivity 437–438
lasers
  neodymium (Nd^{3+}) solid state laser: four-level lasers 473–474
  ruby laser: three-level lasers 471–472
  selective laser sintering (SLS) 196
  semiconductor lasers 474
  solid-state lasers 469–474
latex 144
lath martensite 208
lattice energies, ionic bonding 13–14
lattice parameters, crystal structures 74
lattices, crystal 66–68
LDPE (low-density polyethylene) 140
leaching, Pourbaix diagrams 264–265
LEDS (light emitting diodes) 464–467
Lennard-Jones potential 15, 30, *31*

light 455–460
colour and appearance 459–460
light waves 455–457
photons 457–459
sources of light 460–474
fluorescent lamps 463–464
incandescence 460–461
light emitting diodes (LEDS) 464–467
luminescence 461–463
neodymium (Nd^{3+}) solid state laser: four-level lasers 473–474
organic light emitting devices/diodes (OLEDs) 467–469
ruby laser: three-level lasers 471–472
semiconductor lasers 474
solid-state lasers 469–474
light emitting diodes (LEDS) 464–467
light waves 455–457, *458*
linear absorption coefficient 485
linear attenuation coefficient (extinction coefficient) 485
linear defects, crystal structures 95–96
linear expansivity of a solid 522–523
liquid crystal display thermometers, thermochromic effects 535–537
liquid-phase sintering 196
lithium–air batteries 249–250
lithium-ion batteries 253–254
lithium-ion primary batteries 249

locons 522
longitudinal or axial modulus, elastic moduli 289
longitudinal waves 290–291
longpass filters, reflection 483
Lorentz coefficient 521
Lorentz–Lorenz equation 334
low-carbon steel 54
low-density polyethylene (LDPE) 140
lower bound 316
lower yield point 298
lubricants, solid 309–310
luminescence 461–463

m

Madelung constant 13–14
magnetic data storage 391
magnetic defects 395–399
  cobaltites 396–399
  colossal magnetoresistance (CMR) 398–399
  manganites 396–399
  semiconductors 395–396
  spin states 396–399

magnetic materials 365–372
antiferromagnetic compounds 370–371
atomic magnetism 368–369
characteristics 365–366, *367*
diamagnetic materials 365
Gouy balance 365, *366*
magnetic field strength 366
magnetic permeability 366
magnetic susceptibility 366
magnetisation 366
permeability of free space 366
vacuum permeability 366
ferrimagnetic materials 371
magnetic dipoles 366–372
magnetic flux 366–368
isotropic materials 367
magnetic flux density 367
magnetic induction 367
magnetic polarisation 367
overview 369–372
magnetocaloric effect 533–535
adiabatic cooling 533–534
giant magnetocaloric effects 534–535
magnetostatic energy 383
manganites, magnetic defects 396–399

many electron atoms 4–6
derivation of atomic terms 573–574
energy levels 575
ground state term of an atom 574
martensite finish temperature 209
martensite start temperature 209
martensitic transformations 206–214
austenite–martensite transition 207–210
Ni–Ti alloys 211–212
Nitinol 211, 213–214
shape-memory alloys 212–214
thermoelastic martensitic transformation 212
zirconia 210–211
Matthiessen's rule 410
Maxwell–Boltzmann statistics 177, 515–516
mechanochemistry, solid-state reactions 228–230
medium-carbon steel 54
Meissner effect 432, *433*
mercury lamps 464
metal–insulator transitions 422–427
electron–electron repulsion 423–425
electronic conductivity 422–427
metals and insulators 422–423
modification of insulators 425–426
transparent conducting oxides 426–427

metallic bonding 21–28
bands in ionic and covalent solids 27–28
energy bands 21–22, 24–28
free electron gas 21, 22–24
molecular orbitals 21–22
metallic glasses 115–116
metallic radii 110–112
metallocene catalyst 145
metallocenes, polymers 146–147
metal-organic frameworks (MOFs) 147–149, *150*, 151–152

metals 109–118
see also corrosion
alloy solid solutions 112–115
ductility 301
high temperature oxidation 220–225
alloys 225
direct corrosion 220–222
film growth via diffusion 223–225
oxide film microstructure 222–223
rate of oxidation 222
metallic glasses 115–116
metallic radii 110–112
plastic deformation 294–297
principal properties 116–118
quasicrystals 115–116
structures 109–110
substitutional alloy formation 112–114
valence states 265–268
microdomains 354
micro-hardness 284
microstructures
phase diagrams 214–219, *220*
thermal conductivity 520–522
microstructures development 43–45
processing 44–45
solidification 43–44
Mie scattering 484, *485*

mild carbon steel 54
Miller–Bravais Indices, crystal planes and directions 76–77, 78
Miller Indices, crystal planes and directions 74–77, 78
modal (or intermodal) dispersion, fibre optics 494
MOFs (metal-organic frameworks) 147–149, 150, 151–152
Mohs scale of hardness
  silicate structures **124**
  solids 283, 284
molar (decadic) attenuation coefficient 486
molar heat capacity 515
molar polarisability 331
molecular dynamics simulations 303–304
molecular orbitals 16
  bond energies 18–21
  energy bands 21–22
  metallic bonding 21–22
molybdenum disulfidee 309–310
monomers 133
monomode fibres, fibre optics 495
morphotropic phase boundaries (MPBs) 353
Morse potential 15
mortar 153
moth-eye AR coatings, nanostructures 505
Mott–Hubbard splitting 423–424
Mott insulators 422–423
MPBs (morphotropic phase boundaries) 353
multiple quantum well structures 42, 43

n

Nabarro–Herring creep 309

nanodomains 354

nanoparticle arrays, nanostructures 504–505

nanoparticles 40–42, *43*

  nanostructures 502–504

nanoscale properties 309–314

  auxetic materials 310–312, *313*

  nanoindentation 312–314

  nanowires 312–314

  solid lubricants 309–310

  thin films 312–314

nanostructures 40–42, *43*, 391–395, 501–505
data recording 391
optical aspects
graded refractive index (GRIN) coating 505
moth-eye AR coatings 505
nanoparticle arrays 504–505
nanoparticles 502–504
quantum dot light emitting device (QLED) 504–505
quantum dots 502–505
quantum wells 502
ultrahigh spatial-frequency surface relief grating 505
perovskite superlattices 392–393
photoinduced magnetism 393–395
quantum confinement of electrons 445–447
superlattices 445, *446*
superparamagnetism 391–392
thin films 391–392
(Napierian) linear absorption coefficient 485
(Napierian) linear attenuation coefficient 485
(Napierian) linear scattering coefficient 485
NASICON-related crystals, superionic conductors 188–189
neck growth, sintering 198
negative photoresists, solid-state reactions 228, *229*
negative thermal expansion (NTE) materials 524
neodymium (Nd^{3+}) solid state laser: four-level lasers 473–474
neodymium cuprate, Nd_2CuO_4, superconductivity 438–439

neoprene 144, *145*
Nernst–Einstein equation 183–184
Nernst equation 243, 244, 246, 256, 257
  stability field of water 265
Ni–Ti alloys, martensitic transformations 211–212
Nitinol, martensitic transformations 211, 213–214
non-crystalline solids 37–40
  aerogels 40
  amorphous solids 40
  glasses 37–38
  polymers 38–40
non-equilibrium solidification and coring 214–216
non-steady-state diffusion 169–171
NTE (negative thermal expansion) materials 524
n-type semiconductors *406*, 407, 415-422
nuclear accidents 563–564
nuclear decay 554–557
  half-life 554–555
  radioactive dating 546, 555–557
  rate of nuclear decay 554–555
nuclear energy 557–563
  binding energy of nuclides 557–558
  fast breeder reactors 561–562
  fuel for space exploration 561
  fusion and solar cycles 562–563
  nuclear fission 558–560
  thermal reactors for power generation 560–561

nuclear equations, radioactivity 546–547
nuclear fission, nuclear energy 558–560
nuclear stability, radioactivity 550–551
nuclear waste 563–565
  nuclear accidents 563–564
  storage 564–565
nucleation 43
nylons 136–137

o

offset yield strength 294
Ohm's Law 405
OLEDs (organic light emitting devices/diodes) 467–469
oligomers 133
one-way shape-memory effect 213
optical amplification, fibre optics 495–496
optical interference filters 483
optical pumping 471–472
order–disorder transitions 204–206
  inverse spinel structure 205
  normal spinel structure 205
  orientational ordering 205–206
  positional ordering 205
organic framework structures
  covalent organic frameworks (COFs) 147–149, *150*, 151
  metal-organic frameworks (MOFs) 147–149, *150*, 151–152
organic light emitting devices/diodes (OLEDs) 467–469
organic materials *see* polymers
orientational polarisability 330

oxidation
galvanic cells 239–241
high temperature oxidation of metals 220–225
alloys 225
direct corrosion 220–222
film growth via diffusion 223–225
oxide film microstructure 222–223
rate of oxidation 222
water 265
oxygen sensors 246–247

p

parabolic rate constant
  rate of oxidation 222
  solid-state reactions 227
paramagnetic materials 372–379
  crystal field effects 373–376
  crystal field splitting 374–376
  crystal field theory 373–376
  magnetic moments 372–373
  Pauli paramagnetism 378–379
  spin states 373–376
  temperature dependence of paramagnetic susceptibility 376–378
partial solid solutions 49–50
partly crystalline solids 40
passivation, Pourbaix diagrams 264, 268
Pauli paramagnetism 378–379
PBG (photonic band gap) 490
pearlite *218*, 219, *220*, *221*
Peltier coefficient 528–530
peptide bonds 342, *343*
Periodic Table 571
peritectic point 54
permanent dipoles 29–30
permittivity of free space 327–328

perovskite-related structures and series, superconductivity 440–444
perovskite solar cells 499–501
perovskite structure 72–73
perovskite superlattices, nanostructures 392–393
PET (polyethylene terephthalate) 136
phase diagrams 45–54
  equilibrium heat treatment of steel in the Fe–C phase diagram 218–219, *220*
  equilibrium solidification of simple binary alloys 214
  microstructures 214–219, *220*
  non-equilibrium solidification and coring 214–216
  one-component (unary) systems 45–47
  solidification in systems containing a eutectic point 216–218
  ternary systems 54–57
  two-component (binary) systems 48–54
    binary systems containing a eutectic point 49–51, 54
    intermediate phases 52
    iron–carbon system close to iron 52–54
    simple binary diagrams 48–49
Phase Rule 47–48
phase transitions 199–201
  continuous transitions 201
  first-order phase transitions 200–201, 517–518
  heat capacity 516–518
  second-order transitions 201, 518
pH meters 243–245

phonon absorption 494
phonon assisted transition 471
phonons 410, 516
phonon spectrum 516
phosphors 462
photoconductivity 496–497
photoinduced magnetism, nanostructures 393–395
photonic band gap (PBG) 490
photons 457–459
photoresists, solid-state reactions 227–228, 229
photovoltaic solar cells 496–497
physical adsorption (physisorption), metal oxide film microstructure 222
piezoelectrics 334–342
  biomolecular materials 342
  piezoelectric effect 334–342
    crystal symmetry 335–337
  piezoelectric mechanisms 337–338
  piezoelectric polymers 339–342
  pyroelectric effect 334–337
    crystal symmetry 335–337
  quartz oscillators 338–339
Pilling–Bedworth ratio 223, 268
planar defects, crystal structures 96–99

plastic deformation 280
crystal structures 95
metals 294–297
polymers 299–301
plastic strain 294
plate martensite 208
PMRs (polar microregions) 354
p–n junction diode 419–422, 464, 474, 497, *498*
point defects 87–95
Arrhenius equation 178–180
crystals of the elements 88–89
Frenkel defects 90–91
Kröger–Vink notation 93–94
non-stoichiometric compounds 91–93
point defect notation 93–95
Schottky defects 90–91
solid solutions 89–90
Poisson's ratio **287**, 288-289
polar groups, ferroelectricity 349–350, *351*
polarisation
dielectrics 329–330
magnetic polarisation 367
remanent polarisation 346
residual polarisation 346
spontaneous polarisation 334, 346
polarised light 456–457
polar microregions (PMRs) 354

polar molecules 20–21
polycrystalline solids 43–44
polyethylene *see* polythene
polyethylene terephthalate (PET) 136
polyhedral representations, crystal structures 86–87, *88*
polyisoprene 143

polymers 38–40, 133–152
addition polymers 134–135
classification, thermosetting materials 137
condensation polymers 135
conducting polymers 427–431
covalent organic frameworks (COFs) 147–149, *150*, 151
cross-linking 137–138, 144
divisions 133–134
elastomers 134, 142–144
elongation 299–300, *301*
epoxy resins 137–138
gels 134
lamellae 140, *141*
metal-organic frameworks (MOFs) 147–149, *150*, 151–152
microstructures of polymers 138–142
neoprene 144, *145*
nylons 136–137
piezoelectric polymers 339–342
plastic deformation 299
fracture following plastic deformation 299–301
polyethylene terephthalate (PET) 136
polymer formation 134–138
polythene 38–40, 135, 140, 141
principal properties 149–152

production of polymers 144–147
chain growth 144–147
free radical polymerisation 145–146
metallocenes 146–147
step growth 144–147
termination 146, *147*
properties 138, 141–142
thermoplastic materials 134
thermosetting materials 134
poly(methyl methacrylate) 141
polypeptides 342, *343*
polysynthetic twinning 208, *209*
polythene 38–40, 135, 140, 141
Portland cement 153–154, *155*
positive photoresists., solid-state reactions 228, *229*
Pourbaix diagrams 264–268
corrosion 264–268
leaching 264–265
passivation 264, 268
stability field of water 265
valence states 265–268
power law creep 307–308
practical reaction rate constant, solid-state reactions 227
precipitates, crystal structures 99
precipitation strengthening 302
pressure waves 290–291
primary creep 307

propagons 522
protective kinetics, rate of oxidation 222
protective scale, metal oxide film microstructure 223
proteins 342, *343*
p-type semiconductors *406*, 407, 415-422
P waves 290–291
pyroelectric effect 334–337
pyroxenes 120

q

QLED (quantum dot light emitting device), nanostructures 504–505
quantum confinement of electrons, nanostructures 445–447
quantum dot light emitting device (QLED), nanostructures 504–505
quantum dots 42, 43
  nanostructures 502–505
quantum wells, nanostructures 502
quantum wire 42, 43
quartz oscillators, piezoelectrics 338–339
quasicrystals, metals 115–116
quicklime 153

r

radioactive dating 546, 555–557
radioactivity 545–554
  artificial radioactive atoms 551–554
    heavy elements 551–553
    light elements 553–554
  half-life **552**
  isotopes and nuclides 546
  naturally occurring radioactive elements 545–546
  nuclear equations 546–547
  nuclear stability 550–551
  radioactive dating 546, 555–557
  radioactive series 547–550
    actinium series 548–550
    neptunium/plutonium series 549–550
    thorium series 548
    uranium series 547–548
rainbow holograms 490, *491*
random walk diffusion 171–175
rapid prototyping, sintering 196, 199
rapid sintering 198–199
Rayleigh scattering 483–484, *485*, 493–494
reaction couple, spinel formation 226
reaction sintering 195–197
reconstructive transitions 203–204
redox reactions, galvanic cells 239–241

reduced Young's modulus 314
reduction
  galvanic cells 239–241
  water 265
reduction in grain size 301
reflection 477–483
  bandpass filters 483
  dielectric mirrors 483
  high-reflectivity coatings 482–483
  light 459
  longpass filters 483
  low-reflectivity (antireflection) coatings 482–483
  optical interference filters 483
  reflection from a surface 477–478
  reflection from a transparent thin film 478–482
    multiple thin films 483
  shortpass filters 483
refraction 474–477
  refractive coefficient 474–477
  refractive index 474–477
refrigeration, thermoelectric effects 531–533
relative permittivity
  crystals 332–334
  dielectrics 327–330, 332–334
relaxor ferroelectrics 354, *355*
remanent or residual flux density (remanence) 384–385
remanent polarisation 346

residual polarisation 346
Reuss bound/limit 316
reverse bias 421
ring diffusion 175–176
Rockwell hardness number (RHN) 283
ruby laser: three-level lasers 471–472
Russell–Saunders coupling 11, 368

S

scattering 483–485
  Mie scattering 484, *485*
  Rayleigh scattering 483–484, *485*, 493–494
Schottky defects
  crystal structures 90–91
  diffusion in crystals 179, **180**
  diffusion in solids 176
  ionic conductivity 183
screw dislocations, crystal structures 95–96, *97*
secant modulus 280, *282*
secondary creep 307
second-order transitions 201, 518
Seebeck coefficient 527–533
Seebeck effect 527–533
selective laser sintering (SLS) 196
self-diffusion 167–169, **172**, 174–178
  correlation factors 180–181, 184
  in crystals 178
  interstitial diffusion 181
  in one dimension 176–178
  in solids 175–176
semiconductor lasers 474
semiconductor nanostructures 42, *43*

semiconductors 411–422
band gap 406
band gap measurement 412–413
characterisation 416–419
diluted magnetic semiconductors 395
extrinsic semiconductors 407, 413–416
intrinsic semiconductors 405–406, 411–412
magnetic defects 395–396
n-type semiconductors *406*, 407, 415-422
p–n junction diode 419–422
p-type semiconductors *406*, 407, 415-422
semimetals *406*, 407
shape-memory alloys
martensitic transformations 212–214
one-way shape-memory effect 213
shear modulus
elastic moduli 289–290
solids 284, *286*
shear strength, solids 277, *278*
Shockley equation 421
shortpass filters, reflection 483

silicate glasses 126–133
bonding 127–129
glass-ceramics 132–133
glass deformation 129–131
non-Arrhenian behaviour 130
strengthened glass 131–132
structure 127–129
network model 127–129
viscosity 129–131
working range 129
silicates, crystalline *see* crystalline silicates
silicate structures **124**
sintering 45, 195–199
3D printing 196, 199, 227
additive manufacturing 199
driving force 197–198
grain growth 198
hot pressing 196
liquid-phase sintering 196
neck growth 198
rapid prototyping 196, 199
rapid sintering 198–199
reaction sintering 195–197
selective laser sintering (SLS) 196
slaked lime 153
slip bands 295
slip directions 296

slip planes 296
slip systems 296–297
SLS (selective laser sintering) 196
sodalite units 121, *123*
sodium (Na) superionic conductor, superionic conductors 188–189
solar cells
  dye-sensitised solar cells (DSSCs) 497–499
  perovskite solar cells 499–501
  photoconductivity 496–497
  photovoltaic solar cells 496–497
  solar concentrators 497
solid lubricants 309–310
solid solutions, point defects 89–90
solid-state lasers 469–474
solid-state reactions 225–230
  mechanochemistry 228–230
  photoresists 227–228, *229*
  spinel formation 225–227, 230
solitons 430, *431*
space charge polarisability 330–331
spalling, metal oxide film microstructure 223
specific heat capacity 515
specific modulus 280
specific strength 280
specular reflection 459
sphere packing, crystal structures 82–84

spinel formation, solid-state reactions 225–227, 230
spinel structure 73–74
  inverse spinel structure 205
  normal spinel structure 205
spinodal decomposition 38
spin states
  magnetic defects 396–399
  paramagnetic materials 373–376
spintronics 395
spontaneous emission of photons 458–459, *458*
spontaneous polarisation 334, 346
SQUID (superconducting quantum interference device) 436–437
stabilised zirconia oxides, superionic conductors 188
stability field of water, Pourbaix diagrams 265
standard electrode potentials, galvanic cells 241–242
standard hydrogen electrode 242
standard reaction Gibbs energy 243
steady-state creep 307
steady-state diffusion 169–171
step growth, polymers 144–147
stepped index multimode fibres, fibre optics 494–495
stiffness, solids 280–282, 286
strain hardening 302

strength and hardness of solids 277–284, *285*
hardness 283–284, *285*
Brinell hardness number (BHN) 283
compliance 286
diamond 284
hardness numbers 283
indentation hardness 283–284, *285*
iso-electronic crystals 284
Knoop hardness number (KHN) 284
micro-hardness 284
Mohs scale 283
Rockwell hardness number (RHN) 283
stiffness 286
Vickers hardness number (VHN) 283–284

strength 277, 278
compressive strength 277, 278, 279
elastic moduli 277, 280, 282, 285–291
elastomers 280, 281
flexural strength 277, 278
impact strength 277, 278
shear modulus 284, 286
shear strength 277, 278
specific strength 280
stiffness 280–282
stress-strain curves 281, 282, 283
tensile strength 277, 278, 279–280
torsional strength 277, 278
toughness 280–282
Young's modulus 277, 280, 282, 286–288
stress and strain 278–280
superelasticity 282
strengthened glass 131–132
Griffith's flaws 131
tempered glass 131
stress and strain, solids 278–280
stress concentration factor, 292
stress cycles, fatigue 304–305
stress-induced martensitic transformation 282, 283
stress-strain curves, solids 281, 282, 283
substitutional alloy formation 112–114
substitutional defect 413

subtractive colouration 459
sun tanning beds 464
superconductivity 431–445
Bardeen–Cooper–Schrieffer (BCS) theory 431–432, 434–435
bi-layer graphene 444–445
cuprate high temperature superconductors 437–444
effect of magnetic fields and current 432–434
Josephson junctions 435–437
lanthanum cuprate, La_2CuO_4 437–438
neodymium cuprate, Nd_2CuO_4 438–439
perovskite-related structures and series 440–444
SQUID (superconducting quantum interference device) 436–437
superconductors 431–432
transition temperature 431
unconventional superconductors 431–432
yttrium barium copper oxide, $YBa_2Cu_3O_7$ 439–440
supercooling 37–38, 46–47, 127
superelasticity 54, 212
superexchange, antiferromagnetic materials 386–387
superionic conductors 184–189
β-alumina oxides 185–188
disordered cation compounds 184–185
NASICON-related crystals 188–189
sodium (Na) superionic conductor 188–189
stabilised zirconia oxides 188

superlattices 42, 43
ferroelectrics 354–356
nanostructures 392–393, 445, *446*
perovskite superlattices 392–393
superparamagnetism, nanostructures 391–392
syndiotactic polymers 141

t

talc 120, *123*
Tamman temperature 297
tangent modulus 280, *282*
tempered glass 131
tempering, martensitic transformations 210
tensile strength 277, *278*, 279–280
terminal solid solutions 49–50
ternary systems, phase diagrams 54–57
tertiary creep 309
thermal conductivity 518–522
  heat transfer 518–520
  microstructure 520–522
thermal contraction 524–527
thermal diffusivity 518
thermal expansion 522–524, 525
  interatomic potentials 523–524
thermal reactors for power generation, nuclear energy 560–561
thermal scattering 410
thermochromic effects 535–537
  liquid crystal display thermometers 535–537
  vanadium dioxide 537
thermocouples, thermoelectric effects 531–533
thermoelastic martensitic transformation 212

thermoelectric effects 527–533
charge carriers 529–530
power generation 531–533
refrigeration 531–533
Seebeck coefficient 527–533
Seebeck effect 527–528, *529*
thermocouples 531–533
thermoelectric coefficients 527–533
thermopiles 531–533
thermometers, liquid crystal display thermometers 535–537
thermopiles, thermoelectric effects 531–533
thermoplastic materials 134
thermosetting materials 134
thin films
ferroelectrics 354–356
nanoscale properties 312–314
nanostructures 391–392
reflection from a transparent thin film 478–482
multiple thin films 483
superparamagnetism 391–392
three-point bend test 279
time-dependent properties 304–309
creep 305–309
fatigue 304–305
titanium, electrolytic preparation of titanium, using FFC Cambridge Process 263–264
torsional strength, solids 277, *278*

toughness, solids 280–282
tracer diffusion 167–169
transcrystalline fracture 291
transgranular fracture 291
transition-metal cations, ferroelectricity 350–352
transition temperature 347
transparent conducting oxides (TCOs) 426–427
transparent metals 426–427
transverse waves 290–291
trichromatic (colour 80) lamps 463
twin boundaries, crystal structures 98, 99
two-component (binary) systems 48–54
  binary systems containing a eutectic point 49–51, 54
  intermediate phases 52
  iron–carbon system close to iron 52–54
  simple binary diagrams 48–49

u

ultrahigh-carbon steel 54
ultra-high molecular weight polyethylene, UHMWPE 140
ultrahigh spatial-frequency surface relief grating, nanostructures 595
ultramarines 121
ultrasonic waves in elastic solids 290–291
ultraviolet 456
unit cells, crystal structures 65–68, 69–71
upper bound 315
upper yield point 298

v

vacancy diffusion 175, 176
valence band (VB) 406
valence induction 429
valence states, Pourbaix diagrams 265–268
vanadium dioxide, thermochromic effects 537
van der Waals bonding 30
van der Waals gap 310
van der Waals materials 41
Vegard's law
  crystal structures 74
  polymers 141
Vickers hardness number (VHN) 283–284, 286
visible spectrum 456, **457**
vitrification 126
Vogel–Fulcher–Tammann equation 130
Voigt bound/limit 315
volume defects: precipitates, crystal structures 99
vortex lattice (fluxon lattice or flux lattice) 433, 434
vulcanisation 144

w

wavefunctions 3–4, 7, 23–24, 25, 31
wavelength dispersion, fibre optics 494
weak chemical bonds 28–31
  hydrogen bonds 29
  ion–dipole interaction energies 30
  permanent dipoles 29–30
Weiss domains, ferromagnetic materials 382–383, *384*
Wiedemann–Franz law 116–117
work hardening 301
working range, silicate glasses 129

y

yield point 280
yield stress 280
Young's modulus 277, 280, 282, 286–288
  composite materials 315–317
  reduced Young's modulus 314
yttria-stabilised zirconia (YSZ) 250–251, 261
yttrium barium copper oxide, $YBa_2Cu_3O_7$, superconductivity 439–440

z

Zeeman effect 10–11
zeolites 121
zero thermal expansion (ZTE) materials 525, 526, 527
Ziegler–Natta catalyst 145
zirconia 92, 123–125
  calcia-stabilised 246–247
  martensitic transformations 210–211
  stabilised zirconia oxides 188
  yttria-stabilised zirconia (YSZ) 250–251, 261
ZTE (zero thermal expansion) materials 525, 526, 527

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.