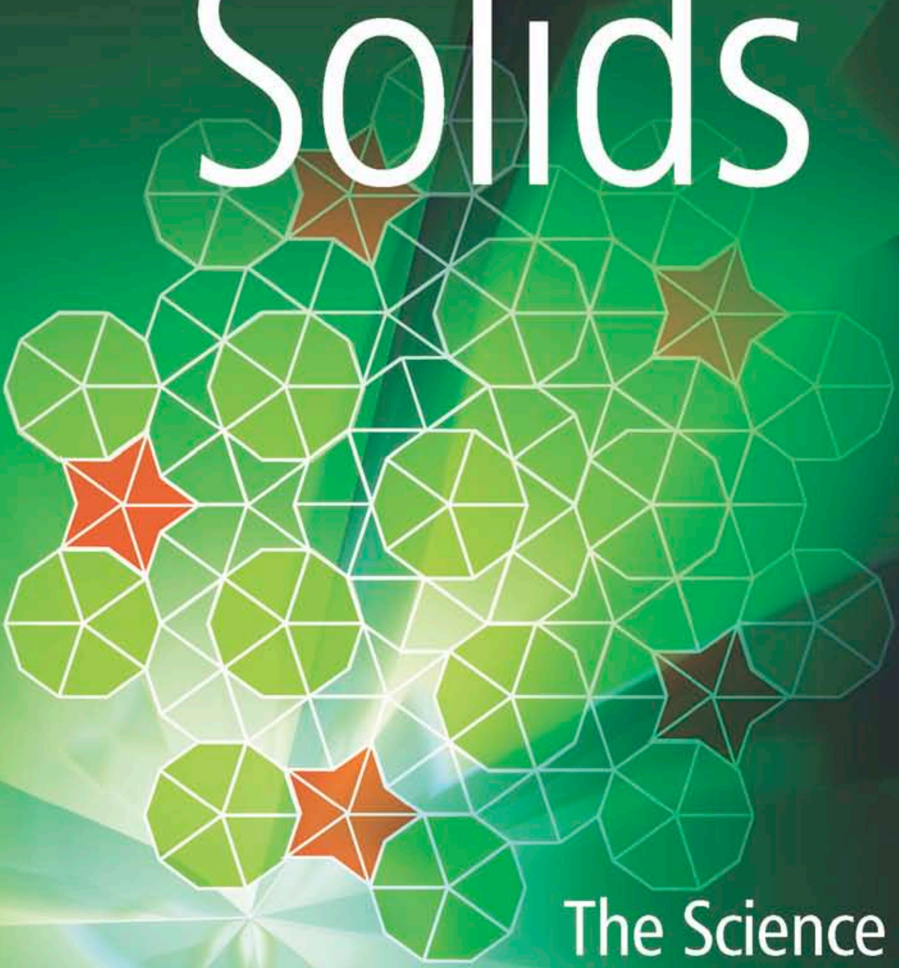
Richard J. D. Tilley

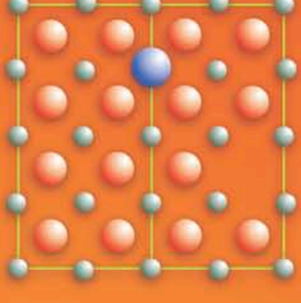
Second Edition

Understanding Solids



The Science of Materials

WILEY



Group

Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	IA	IIA	IIIB	IVB	VB	VIB	VIIB	VIIIB			IB	IIB	IIIA	IVA	VA	VIA	VIIA	VIII
1	1 H $1s^1$																	2 He $1s^2$
2	3 Li $2s^1$	4 Be $2s^2$											5 B $2s^2 2p^1$	6 C $2s^2 2p^2$	7 N $2s^2 2p^3$	8 O $2s^2 2p^4$	9 F $2s^2 2p^5$	10 Ne $2s^2 2p^6$
3	11 Na $3s^1$	12 Mg $3s^2$											13 Al $3s^2 3p^1$	14 Si $3s^2 3p^2$	15 P $3s^2 3p^3$	16 S $3s^2 3p^4$	17 Cl $3s^2 3p^5$	18 Ar $3s^2 3p^6$
4	19 K $4s^1$	20 Ca $4s^2$	21 Sc $4s^2 3d^1$	22 Ti $4s^2 3d^2$	23 V $4s^2 3d^3$	24 Cr $4s^1 3d^5$	25 Mn $4s^2 3d^5$	26 Fe $4s^2 3d^6$	27 Co $4s^2 3d^7$	28 Ni $4s^2 3d^8$	29 Cu $4s^1 3d^{10}$	30 Zn $4s^2 3d^{10}$	31 Ga $4s^2 4p^1$	32 Ge $4s^2 4p^2$	33 As $4s^2 4p^3$	34 Se $4s^2 4p^4$	35 Br $4s^2 4p^5$	36 Kr $4s^2 4p^6$
5	37 Rb $5s^1$	38 Sr $5s^2$	39 Y $5s^2 4d^1$	40 Zr $5s^2 4d^2$	41 Nb $5s^1 4d^4$	42 Mo $5s^1 4d^5$	43 Tc $5s^2 4d^5$	44 Ru $5s^1 4d^7$	45 Rh $5s^1 4d^8$	46 Pd $5s^0 4d^{10}$	47 Ag $5s^1 4d^{10}$	48 Cd $5s^2 4d^{10}$	49 In $5s^2 5p^1$	50 Sn $5s^2 5p^2$	51 Sb $5s^2 5p^3$	52 Te $5s^2 5p^4$	53 I $5s^2 5p^5$	54 Xe $5s^2 5p^6$
6	55 Cs $6s^1$	56 Ba $6s^2$	71 Lu $6s^2 5d^1$	72 Hf $6s^2 5d^2$	73 Ta $6s^2 5d^3$	74 W $6s^2 5d^4$	75 Re $6s^2 5d^5$	76 Os $6s^2 5d^6$	77 Ir $6s^2 5d^7$	78 Pt $6s^1 5d^9$	79 Au $6s^1 5d^{10}$	80 Hg $6s^2 5d^{10}$	81 Tl $6s^2 6p^1$	82 Pb $6s^2 6p^2$	83 Bi $6s^2 6p^3$	84 Po $6s^2 6p^4$	85 At $6s^2 6p^5$	86 Rn $6s^2 6p^6$
7	87 Fr $7s^1$	88 Ra $7s^2$	103 Lr $7s^2 6d^1$	104 Rf $7s^2 6d^2$	105 Db $7s^2 6d^3$	106 Sg $7s^2 6d^4$	107 Bh $7s^2 6d^5$	108 Hs $7s^2 6d^6$	109 Mt $7s^2 6d^7$									

57 La $6s^2 5d^1$	58 Ce $6s^2 4f^2$	59 Pr $6s^2 4f^3$	60 Nd $6s^2 4f^4$	61 Pm $6s^2 4f^5$	62 Sm $6s^2 4f^6$	63 Eu $6s^2 4f^7$	64 Gd $6s^2 5d^1 4f^7$	65 Tb $6s^2 4f^9$	66 Dy $6s^2 4f^{10}$	67 Ho $6s^2 4f^{11}$	68 Er $6s^2 4f^{12}$	69 Tm $6s^2 4f^{13}$	70 Yb $6s^2 4f^{14}$
89 Ac $7s^2 6d^1$	90 Th $7s^2 6d^2$	91 Pa $7s^2 6d^1 5f^2$	92 U $7s^2 6d^1 5f^3$	93 Np $7s^2 6d^1 5f^4$	94 Pu $7s^2 5f^6$	95 Am $7s^2 5f^7$	96 Cm $7s^2 6d^1 5f^7$	97 Bk $7s^2 5f^9$	98 Cf $7s^2 5f^{10}$	99 Es $7s^2 5f^{11}$	100 Fm $7s^2 5f^{12}$	101 Md $7s^2 5f^{13}$	102 No $7s^2 5f^{14}$

Understanding Solids

Understanding Solids

The Science of Materials

2nd edition

Richard J.D. Tilley
Professor Emeritus, University of Cardiff

WILEY
A John Wiley & Sons, Ltd., Publication

This edition first published 2013

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

The publisher and the author 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 fitness for a particular purpose. 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 every situation. 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. The fact that an organization or Website is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Website may provide or recommendations it may make. Further, readers should be aware that Internet Websites listed in this work may have changed or disappeared between when this work was written and when it is read. No warranty may be created or extended by any promotional statements for this work. Neither the publisher nor the author shall be liable for any damages arising herefrom.

Library of Congress Cataloging-in-Publication Data

Tilley, R. J. D.
Understanding solids : the science of materials / Richard J. D. Tilley. – 2nd edition.
pages cm
Includes index.
ISBN 978-1-118-42328-8 (hardback) – ISBN 978-1-118-42346-2 (paper) 1. Materials science. 2. Solids. I. Title.
TA403.T63 2013
620.1'1–dc23
2012047718

A catalogue record for this book is available from the British Library.

Cloth ISBN: 978-1-118-42328-8 (13 digits)
Paper ISBN: 978-1-118-42346-2

Set in 10/12 pt Times by Thomson Digital, Noida, India

For Anne

Contents

Preface to the Second Edition **xvii**
Preface to the First Edition **xix**

PART 1 STRUCTURES AND MICROSTRUCTURES 1

1 The electron structure of atoms **3**

1.1 The hydrogen atom 3
1.1.1 The quantum mechanical description 3
1.1.2 The energy of the electron 4
1.1.3 Electron orbitals 5
1.1.4 Orbital shapes 5
1.2 Many-electron atoms 7
1.2.1 The orbital approximation 7
1.2.2 Electron spin and electron configuration 7
1.2.3 The periodic table 9
1.3 Atomic energy levels 11
1.3.1 Spectra and energy levels 11
1.3.2 Terms and term symbols 11
1.3.3 Levels 13
1.3.4 Electronic energy level calculations 14
Further reading 15
Problems and exercises 16

2 Chemical bonding **19**

2.1 Ionic bonding 19
2.1.1 Ions 19
2.1.2 Ionic size and shape 20
2.1.3 Lattice energies 21
2.1.4 Atomistic simulation 23
2.2 Covalent bonding 24
2.2.1 Valence bond theory 24
2.2.2 Molecular orbital theory 30
2.3 Metallic bonding and energy bands 35
2.3.1 Molecular orbitals and energy bands 36
2.3.2 The free electron gas 37
2.3.3 Energy bands 40

2.3.4 Properties of metals 41
2.3.5 Bands in ionic and covalent solids 43
2.3.6 Computation of properties 44
Further reading 45
Problems and exercises 46

3 States of aggregation 49

3.1 Weak chemical bonds 49
3.2 Macrostructures, microstructures and nanostructures 52
3.2.1 Structures and scale 52
3.2.2 Crystalline solids 52
3.2.3 Quasicrystals 53
3.2.4 Non-crystalline solids 54
3.2.5 Partly crystalline solids 55
3.2.6 Nanoparticles and nanostructures 55
3.3 The development of microstructures 57
3.3.1 Solidification 58
3.3.2 Processing 58
3.4 Point defects 60
3.4.1 Point defects in crystals of elements 60
3.4.2 Solid solutions 61
3.4.3 Schottky defects 62
3.4.4 Frenkel defects 63
3.4.5 Non-stoichiometric compounds 64
3.4.6 Point defect notation 66
3.5 Linear, planar and volume defects 68
3.5.1 Edge dislocations 68
3.5.2 Screw dislocations 69
3.5.3 Partial and mixed dislocations 69
3.5.4 Planar defects 69
3.5.5 Volume defects: precipitates 70
Further reading 73
Problems and exercises 73

4 Phase diagrams 77

4.1 Phases and phase diagrams 77
4.1.1 One-component (unary) systems 77
4.1.2 The phase rule for one-component (unary) systems 79
4.2 Binary phase diagrams 80
4.2.1 Two-component (binary) systems 80
4.2.2 The phase rule for two-component (binary) systems 81
4.2.3 Simple binary diagrams: nickel–copper as an example 81
4.2.4 Binary systems containing a eutectic point: tin–lead as an example 83
4.2.5 Intermediate phases and melting 87
4.3 The iron–carbon system near to iron 88
4.3.1 The iron–carbon phase diagram 88
4.3.2 Steels and cast irons 89
4.3.3 Invariant points 89
4.4 Ternary systems 90

4.5 Calculation of phase diagrams: CALPHAD 93
Further reading 94
Problems and exercises 94

5 Crystallography and crystal structures 101
5.1 Crystallography 101
5.1.1 Crystal lattices 101
5.1.2 Crystal systems and crystal structures 102
5.1.3 Symmetry and crystal classes 104
5.1.4 Crystal planes and Miller indices 106
5.1.5 Hexagonal crystals and Miller-Bravais indices 109
5.1.6 Directions 110
5.1.7 Crystal geometry and the reciprocal lattice 112
5.2 The determination of crystal structures 114
5.2.1 Single crystal X-ray diffraction 114
5.2.2 Powder X-ray diffraction and crystal identification 115
5.2.3 Neutron diffraction 118
5.2.4 Electron diffraction 118
5.3 Crystal structures 118
5.3.1 Unit cells, atomic coordinates and nomenclature 118
5.3.2 The density of a crystal 119
5.3.3 The cubic close-packed (A1) structure 121
5.3.4 The body-centred cubic (A2) structure 121
5.3.5 The hexagonal (A3) structure 122
5.3.6 The diamond (A4) structure 122
5.3.7 The graphite (A9) structure 123
5.3.8 The halite (rock salt, sodium chloride, B1) structure 123
5.3.9 The spinel ($H1_1$) structure 125
5.4 Structural relationships 126
5.4.1 Sphere packing 126
5.4.2 Ionic structures in terms of anion packing 128
5.4.3 Polyhedral representations 129
Further reading 131
Problems and exercises 131

PART 2 CLASSES OF MATERIALS 137

6 Metals, ceramics, polymers and composites 139
6.1 Metals 139
6.1.1 The crystal structures of pure metals 140
6.1.2 Metallic radii 141
6.1.3 Alloy solid solutions 142
6.1.4 Metallic glasses 145
6.1.5 The principal properties of metals 146
6.2 Ceramics 147
6.2.1 Bonding and structure of silicate ceramics 147
6.2.2 Some non-silicate ceramics 149
6.2.3 The preparation and processing of ceramics 152
6.2.4 The principal properties of ceramics 154

6.3 Silicate glasses 154
6.3.1 Bonding and structure of silicate glasses 155
6.3.2 Glass deformation 157
6.3.3 Strengthened glass 159
6.3.4 Glass-ceramics 160
6.4 Polymers 161
6.4.1 Polymer formation 162
6.4.2 Microstructures of polymers 165
6.4.3 Production of polymers 170
6.4.4 Elastomers 173
6.4.5 The principal properties of polymers 175
6.5 Composite materials 177
6.5.1 Fibre-reinforced plastics 177
6.5.2 Metal-matrix composites 177
6.5.3 Ceramic-matrix composites 178
6.5.4 Cement and concrete 178
Further reading 181
Problems and exercises 182

PART 3 REACTIONS AND TRANSFORMATIONS 189

7 Diffusion and ionic conductivity 191

7.1 Self-diffusion, tracer diffusion and tracer impurity diffusion 191
7.2 Non-steady-state diffusion 194
7.3 Steady-state diffusion 195
7.4 Temperature variation of diffusion coefficient 195
7.5 The effect of impurities 196
7.6 Random walk diffusion 197
7.7 Diffusion in solids 198
7.8 Self-diffusion in one dimension 199
7.9 Self-diffusion in crystals 201
7.10 The Arrhenius equation and point defects 202
7.11 Correlation factors for self-diffusion 204
7.12 Ionic conductivity 205
7.12.1 Ionic conductivity in solids 205
7.12.2 The relationship between ionic conductivity and diffusion coefficient 208
Further reading 209
Problems and exercises 209

8 Phase transformations and reactions 213

8.1 Sintering 213
8.1.1 Sintering and reaction 213
8.1.2 The driving force for sintering 215
8.1.3 The kinetics of neck growth 216
8.2 First-order and second-order phase transitions 216
8.2.1 First-order phase transitions 217
8.2.2 Second-order transitions 217

8.3 Displacive and reconstructive transitions 218
8.3.1 Displacive transitions 218
8.3.2 Reconstructive transitions 219
8.4 Order–disorder transitions 221
8.4.1 Positional ordering 221
8.4.2 Orientational ordering 222
8.5 Martensitic transformations 223
8.5.1 The austenite–martensite transition 223
8.5.2 Martensitic transformations in zirconia 226
8.5.3 Martensitic transitions in Ni–Ti alloys 227
8.5.4 Shape-memory alloys 228
8.6 Phase diagrams and microstructures 230
8.6.1 Equilibrium solidification of simple binary alloys 230
8.6.2 Non-equilibrium solidification and coring 230
8.6.3 Solidification in systems containing a eutectic point 231
8.6.4 Equilibrium heat treatment of steel in the Fe–C phase diagram 233
8.7 High-temperature oxidation of metals 236
8.7.1 Direct corrosion 236
8.7.2 The rate of oxidation 236
8.7.3 Oxide film microstructure 237
8.7.4 Film growth via diffusion 238
8.7.5 Alloys 239
8.8 Solid-state reactions 240
8.8.1 Spinel formation 240
8.8.2 The kinetics of spinel formation 241
Further reading 242
Problems and exercises 242

9 Oxidation and reduction 247

9.1 Galvanic cells 247
9.1.1 Cell basics 247
9.1.2 Standard electrode potentials 249
9.1.3 Cell potential and Gibbs energy 250
9.1.4 Concentration dependence 251
9.2 Chemical analysis using galvanic cells 251
9.2.1 pH meters 251
9.2.2 Ion selective electrodes 253
9.2.3 Oxygen sensors 254
9.3 Batteries 255
9.3.1 ‘Dry’ and alkaline primary batteries 255
9.3.2 Lithium-ion primary batteries 256
9.3.3 The lead–acid secondary battery 257
9.3.4 Lithium-ion secondary batteries 257
9.3.5 Lithium–air batteries 259
9.3.6 Fuel cells 260
9.4 Corrosion 262
9.4.1 The reaction of metals with water and aqueous acids 262
9.4.2 Dissimilar metal corrosion 264
9.4.3 Single metal electrochemical corrosion 265
9.5 Electrolysis 266
9.5.1 Electrolytic cells 267
9.5.2 Electroplating 267

9.5.3 The amount of product produced during electrolysis 268
9.5.4 The electrolytic preparation of titanium by the *FFC Cambridge Process* 269
9.6 Pourbaix diagrams 270
9.6.1 Passivation, corrosion and leaching 270
9.6.2 The stability field of water 270
9.6.3 Pourbaix diagram for a metal showing two valence states, M^{2+} and M^{3+} 271
9.6.4 Pourbaix diagram displaying tendency for corrosion 273
Further reading 274
Problems and exercises 275

PART 4 PHYSICAL PROPERTIES 279

10 Mechanical properties of solids 281

10.1 Strength and hardness 281
10.1.1 Strength 281
10.1.2 Stress and strain 282
10.1.3 Stress–strain curves 283
10.1.4 Toughness and stiffness 286
10.1.5 Superelasticity 286
10.1.6 Hardness 287
10.2 Elastic moduli 289
10.2.1 Young's modulus (the modulus of elasticity) (E or Y) 289
10.2.2 Poisson's ratio (ν) 291
10.2.3 The longitudinal or axial modulus (L or M) 292
10.2.4 The shear modulus or modulus of rigidity (G or μ) 292
10.2.5 The bulk modulus, K or B 293
10.2.6 The Lamé modulus (λ) 293
10.2.7 Relationships between the elastic moduli 293
10.2.8 Ultrasonic waves in elastic solids 293
10.3 Deformation and fracture 295
10.3.1 Brittle fracture 295
10.3.2 Plastic deformation of metals 298
10.3.3 Dislocation movement and plastic deformation 298
10.3.4 Brittle and ductile materials 301
10.3.5 Plastic deformation of polymers 302
10.3.6 Fracture following plastic deformation 302
10.3.7 Strengthening 304
10.3.8 Computation of deformation and fracture 306
10.4 Time-dependent properties 307
10.4.1 Fatigue 307
10.4.2 Creep 308
10.5 Nanoscale properties 312
10.5.1 Solid lubricants 312
10.5.2 Auxetic materials 313
10.5.3 Thin films and nanowires 315
10.6 Composite materials 317
10.6.1 Young's modulus of large particle composites 317
10.6.2 Young's modulus of fibre-reinforced composites 318
10.6.3 Young's modulus of a two-phase system 319
Further reading 320
Problems and exercises 321

11 Insulating solids 327

11.1 Dielectrics 327
11.1.1 Relative permittivity and polarisation 327
11.1.2 Polarisability 329
11.1.3 Polarisability and relative permittivity 330
11.1.4 The frequency dependence of polarisability and relative permittivity 331
11.1.5 The relative permittivity of crystals 332
11.2 Piezoelectrics, pyroelectrics and ferroelectrics 333
11.2.1 The piezoelectric and pyroelectric effects 333
11.2.2 Crystal symmetry and the piezoelectric and pyroelectric effects 335
11.2.3 Piezoelectric mechanisms 336
11.2.4 Quartz oscillators 337
11.2.5 Piezoelectric polymers 338
11.3 Ferroelectrics 340
11.3.1 Ferroelectric crystals 340
11.3.2 Hysteresis and domain growth in ferroelectric crystals 341
11.3.3 Antiferroelectrics 344
11.3.4 The temperature dependence of ferroelectricity and antiferroelectricity 344
11.3.5 Ferroelectricity due to hydrogen bonds 345
11.3.6 Ferroelectricity due to polar groups 347
11.3.7 Ferroelectricity due to medium-sized transition-metal cations 348
11.3.8 Poling and polycrystalline ferroelectric solids 349
11.3.9 Doping and modification of properties 349
11.3.10 Relaxor ferroelectrics 351
11.3.11 Ferroelectric nanoparticles, thin films and superlattices 352
11.3.12 Flexoelectricity in ferroelectrics 353
Further reading 354
Problems and exercises 355

12 Magnetic solids 361

12.1 Magnetic materials 361
12.1.1 Characterisation of magnetic materials 361
12.1.2 Magnetic dipoles and magnetic flux 362
12.1.3 Atomic magnetism 363
12.1.4 Overview of magnetic materials 365
12.2 Paramagnetic materials 368
12.2.1 The magnetic moment of paramagnetic atoms and ions 368
12.2.2 High and low spin: crystal field effects 369
12.2.3 Temperature dependence of paramagnetic susceptibility 371
12.2.4 Pauli paramagnetism 373
12.3 Ferromagnetic materials 374
12.3.1 Ferromagnetism 374
12.3.2 Exchange energy 376
12.3.3 Domains 378
12.3.4 Hysteresis 380
12.3.5 Hard and soft magnetic materials 380
12.4 Antiferromagnetic materials and superexchange 381
12.5 Ferrimagnetic materials 382
12.5.1 Cubic spinel ferrites 382
12.5.2 Garnet structure ferrites 383

12.5.3 Hexagonal ferrites 383
12.5.4 Double exchange 384
12.6 Nanostructures 385
12.6.1 Small particles and data recording 385
12.6.2 Superparamagnetism and thin films 386
12.6.3 Superlattices 386
12.6.4 Photoinduced magnetism 387
12.7 Magnetic defects 389
12.7.1 Magnetic defects in semiconductors 389
12.7.2 Charge and spin states in cobaltites and manganites 390
Further reading 393
Problems and exercises 393

13 Electronic conductivity in solids 399

13.1 Metals 399
13.1.1 Metals, semiconductors and insulators 399
13.1.2 Electron drift in an electric field 401
13.1.3 Electronic conductivity 402
13.1.4 Resistivity 404
13.2 Semiconductors 405
13.2.1 Intrinsic semiconductors 405
13.2.2 Band gap measurement 407
13.2.3 Extrinsic semiconductors 408
13.2.4 Carrier concentrations in extrinsic semiconductors 409
13.2.5 Characterisation 411
13.2.6 The p-n junction diode 413
13.3 Metal–insulator transitions 416
13.3.1 Metals and insulators 416
13.3.2 Electron–electron repulsion 417
13.3.3 Modification of insulators 418
13.3.4 Transparent conducting oxides 419
13.4 Conducting polymers 420
13.5 Nanostructures and quantum confinement of electrons 423
13.5.1 Quantum wells 424
13.5.2 Quantum wires and quantum dots 425
13.6 Superconductivity 426
13.6.1 Superconductors 426
13.6.2 The effect of magnetic fields 427
13.6.3 The effect of current 428
13.6.4 The nature of superconductivity 428
13.6.5 Josephson junctions 430
13.6.6 Cuprate high-temperature superconductors 430
Further reading 438
Problems and exercises 438

14 Optical aspects of solids 445

14.1 Light 445
14.1.1 Light waves 445
14.1.2 Photons 447
14.2 Sources of light 449

14.2.1 Incandescence 449
14.2.2 Luminescence and phosphors 450
14.2.3 Light-emitting diodes (LEDs) 453
14.2.4 Solid-state lasers 454
14.3 Colour and appearance 460
14.3.1 Luminous solids 460
14.3.2 Non-luminous solids 460
14.3.3 Attenuation 461
14.4 Refraction and dispersion 462
14.4.1 Refraction 462
14.4.2 Refractive index and structure 464
14.4.3 The refractive index of metals and semiconductors 465
14.4.4 Dispersion 465
14.5 Reflection 466
14.5.1 Reflection from a surface 466
14.5.2 Reflection from a single thin film 467
14.5.3 The reflectivity of a single thin film in air 469
14.5.4 The colour of a single thin film in air 469
14.5.5 The colour of a single thin film on a substrate 470
14.5.6 Low-reflectivity (antireflection) and high-reflectivity coatings 471
14.5.7 Multiple thin films and dielectric mirrors 471
14.6 Scattering 472
14.6.1 Rayleigh scattering 472
14.6.2 Mie scattering 475
14.7 Diffraction 475
14.7.1 Diffraction by an aperture 475
14.7.2 Diffraction gratings 476
14.7.3 Diffraction from crystal-like structures 477
14.7.4 Photonic crystals 478
14.8 Fibre optics 479
14.8.1 Optical communications 479
14.8.2 Attenuation in glass fibres 479
14.8.3 Dispersion and optical fibre design 480
14.8.4 Optical amplification 482
14.9 Energy conversion 483
14.9.1 Photoconductivity and photovoltaic solar cells 483
14.9.2 Dye sensitized solar cells 485
14.10 Nanostructures 486
14.10.1 The optical properties of quantum wells 486
14.10.2 The optical properties of nanoparticles 487
Further reading 489
Problems and exercises 489

15 Thermal properties 495

15.1 Heat capacity 495
15.1.1 The heat capacity of a solid 495
15.1.2 Classical theory of heat capacity 496
15.1.3 Quantum theory of heat capacity 496
15.1.4 Heat capacity at phase transitions 497
15.2 Thermal conductivity 498
15.2.1 Heat transfer 498
15.2.2 Thermal conductivity of solids 498

15.2.3 Thermal conductivity and microstructure 500
15.3 Expansion and contraction 501
15.3.1 Thermal expansion 501
15.3.2 Thermal expansion and interatomic potentials 502
15.3.3 Thermal contraction 503
15.3.4 Zero thermal contraction materials 505
15.4 Thermoelectric effects 506
15.4.1 Thermoelectric coefficients 506
15.4.2 Thermoelectric effects and charge carriers 508
15.4.3 The Seebeck coefficient of solids containing point defect populations 509
15.4.4 Thermocouples, power generation and refrigeration 509
15.5 The magnetocaloric effect 512
15.5.1 The magnetocaloric effect and adiabatic cooling 512
15.5.2 The giant magnetocaloric effect 513
Further reading 514
Problems and exercises 514

PART 5 NUCLEAR PROPERTIES OF SOLIDS 517

16 Radioactivity and nuclear reactions 519

16.1 Radioactivity 519
16.1.1 Naturally occurring radioactive elements 519
16.1.2 Isotopes and nuclides 520
16.1.3 Nuclear equations 520
16.1.4 Radioactive series 521
16.1.5 Nuclear stability 523
16.2 Artificial radioactive atoms 524
16.2.1 Transuranic elements 524
16.2.2 Artificial radioactivity in light elements 527
16.3 Nuclear decay 527
16.3.1 The rate of nuclear decay 527
16.3.2 Radioactive dating 529
16.4 Nuclear energy 531
16.4.1 The binding energy of nuclides 531
16.4.2 Nuclear fission 532
16.4.3 Thermal reactors for power generation 533
16.4.4 Fuel for space exploration 535
16.4.5 Fast breeder reactors 535
16.4.6 Fusion 535
16.4.7 Solar cycles 536
16.5 Nuclear waste 536
16.5.1 Nuclear accidents 537
16.5.2 The storage of nuclear waste 537
Further reading 538
Problems and exercises 539

Subject Index 543

Preface to the Second Edition

In this edition the structure of the First Edition has been retained. However, in the intervening years an enormous number of new experimental and theoretical studies have appeared in the literature, and although this is an introductory text, many of these new studies merit reporting. The additions are spread throughout the text, but can be roughly divided into four groups. The first area is computational science and engineering. Computer simulations of the behaviour of solids, from the engineering of large-scale structures using finite element analysis, via atomistic simulation and molecular dynamics, used to study features such as elastic behaviour and dislocation movement, through to the calculation of the electronic properties of solids via density functional theory, are used routinely in almost all disciplines that are related to the subject matter of this book. The computer calculation of phase diagrams (CALPHAD) also comes into this category. An indication of the principles underlying these programmes of work is given to underpin the computational "black boxes" available (Chapters 1, 2, 4 and 10).

Another area of change is that concerned with nanoscale properties. Here, not only is the primary literature growing rapidly but there are a rapidly increasing number of specialised books that deal with the subject. Some of the properties that are affected by small scales are described, including the magnetic and ferroelectric properties of thin films and superlattices (Chapters 11 and 12) and nanoparticle colours (Chapter 14).

Point defects are key components for the manipulation of the physical properties of solids, and this area has been supplemented. The universally used Kroger-Vink notation for point defects has been defined (Chapter 3). The interplay between defect populations and ionic conductivity (Chapter 7), relaxor ferroelectrics (Chapter 11), the electronic and magnetic properties of cobaltites and manganites, including colossal magnetoresistance (Chapter 12, 13), cuprate superconductors (Chapter 13) and thermoelectric properties (Chapter 15) are described.

New text has also been added to cover a number of important crystallographically related topics. These include quasicrystals (Chapter 3), phase transformations including first- and second-order transitions, displacive versus reconstructive transitions and order–disorder transitions, and martensitic transitions (Chapter 9), and coverage of crystal symmetry with respect to piezoelectricity (Chapter 11).

Other new topics include lithium-air batteries (Chapter 8), increased discussion of elastic moduli and their measurement using ultrasonic waves (Chapter 10), the newly appreciated physical property of flexoelectricity (Chapter 11), crystal field effects, the magnetic properties of garnets and photoinduced magnetism (Chapter 12), dye-sensitized solar cells (Chapter 14), zero thermal expanding solids (Chapter 15), increased discussion of nuclear stability and radioisotope dating (Chapter 16).

The text on all these topics is necessarily compact, but can be supplemented by reference to extensive additional sources listed in the Further Reading sections.

Because of this expansion of material, of necessity some sections of the original have had to be excised. For the same reason, much of the supplementary information in the first edition has been removed or incorporated into the relevant chapters. In making these changes, all the material has been rewritten and rearranged to the extent that no single page of this edition is identical to any in the first edition. In addition, all figures have been redrawn in colour. These, together with the solutions to the Introductory Questions and Problems and Exercises, are to be found on a companion web site (http://www.wiley.com/go/tilleysolids2e).

It is a pleasure to acknowledge the help from the staff at John Wiley, including Rebecca Stubbs, Emma Strickland and Sarah Tilley, all of whom were enthusiastic about this project and fielded my many queries efficiently. The staff of the Trevithick library in Cardiff University provided, as always, help in tracking down obscure references. Professor F.S. Stone and Dr D.F. Klemperer were generous in their encouragement. Finally I must acknowledge my wife Anne, who has continually supported and helped in more ways that it is possible to record.

R.J.D. Tilley
August 2012

Preface to the First Edition

This book originated in lectures to undergraduate students in materials science, which were later extended to geology, physics and engineering students. The subject matter is concerned with the structures and properties of solids. The material is presented with a science bias, and is aimed not only at students taking traditional materials science and engineering courses, but also courses in the rapidly expanding fields of materials chemistry and physics. The coverage aims to be complementary to established books in materials science and engineering. The level is designed to be introductory in nature, and as far as is practical, the book is self-contained. The chapters are provided with questions designed to reinforce the concepts presented. These are in two parts. A multiple choice 'Quick Quiz' is designed to be tackled rapidly, and aims to uncover weaknesses in a student's grasp of the fundamental concepts described. The 'Calculations and Questions' are more traditional, containing numerical examples to test the understanding of formulae, and derivations that are not carried out in the main body of the text. Many chapters contain one or more appendices that bear directly upon the material, but which would disrupt the flow of the subject matter if included within the chapter itself. These are meant to provide more depth than is possible otherwise. Further Reading allows students to take matters a little further. With only one exception, the references are to printed information. In general, it would be expected that a student would initially turn to the Internet for information. Sources here are rapidly located and this avenue of exploration has been left to the student.

The subject matter is divided into five sections. Part 1 covers the building blocks of solids. Here the topics centre upon atoms and bonding, and the patterns of structure that result. In this section, the important concepts of microstructure and macrostructure are developed, leading naturally to an understanding of why nanostructures possess unique properties. Defects that are of importance are also described here. Part 2 is concerned with the traditional triumvirate of metals, ceramics and polymers, together with a brief introduction to composite materials. The subject is condensed into a single chapter. It provides an overview of a comparative nature, focused upon giving a broad appreciation of why the fundamental groups of materials appear to differ so much, and laying the foundations for why some, such as ceramic superconductors, seem to behave so differently from their congeners. Part 3 has a more chemical bias, and describes reactions and transformations. The principles of diffusion are outlined in Chapter 7; electrochemical ideas, which lead naturally to batteries, corrosion and electroplating, are described in Chapter 8. Solid-state transformations, which impinge upon areas as diverse as shape memory alloys, semiconductor doping and sintering, are introduced in Chapter 9. Part 4 is a description of the physical properties of solids, and complements the chemical aspects detailed in Part 3. The topics covered are those of importance to both science and technology, mechanical, Chapter 10, insulators, Chapter 11, magnetic, Chapter 12, electronic, Chapter 13, optical, Chapter 14, and thermal, Chapter 15. Part 5 is concerned with radioactivity. This topic is of enormous importance, and in particular the disposal of nuclear waste in solid form is of pressing concern.

The material in all of the later sections is founded upon the concepts presented in Part 1, that is, properties are explained as arising naturally from the atomic constituents, the chemical bonding, the microstructures and defects present in the solid. This leads naturally to an understanding of why nanostructures have seemingly different properties from bulk solids. Because of this, nanostructures are not gathered together in one section, but considered throughout the book, in the context of the better-known macroscopic properties of the material.

It is a pleasure to acknowledge the help of Dr A. Slade and Mrs Celia Carden of John Wiley, who have given continual encouragement and assistance to this venture. Ms Rachael Catt read the complete manuscript with meticulous care, exposed ambiguities and inconsistencies in both text and figures, and added materially to the final version. Mr Allan Coughlin read large parts of earlier drafts, clarified many obscurities and suggested many improvements. Mr Rolfe Jones has provided information and micrographs of solids whenever called upon. As always, my family has been ever-supportive during the writing of this book, and my wife Anne has endured the hours of being a computer widow without complaint. To all of these, my heartfelt thanks.

PART 1

Structures and microstructures

1

The electron structure of atoms

- What is a wavefunction?
- What is an atomic term?
- How are the energy levels of atoms labelled?

An atom of any element 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*. The proton number is an integer specifying the number of protons in the nucleus, each of which carries one unit of positive charge. 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 fewer electrons than are required for charge neutrality are called *ions*; those which have lost electrons have an overall positive charge and those that have gained electrons have an overall negative charge. Positively charged ions are sometimes called *cations* and negatively charged ions are sometimes called *anions*.

The electrons associated with the chemical elements in a material (whether in the form of a gas, liquid or solid) control the chemical and physical properties of the atoms. The energies and regions of space occupied by electrons in an atom may be calculated using *quantum theory*.

1.1 The hydrogen atom

1.1.1 The quantum mechanical description

A hydrogen atom is the simplest of atoms. It 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. *Hydrogenic* or *hydrogen-like* atoms or ions are very similar, in that they can be analysed in terms of a single electron bound to a nucleus with an apparent charge different from unity. Information about the electron can be obtained by solving the *Schrödinger equation*, in which the electron is represented as a wave. The permitted solutions to this equation, called *wavefunctions*, describe the energy and probability of location of the electron in any region around the nucleus. Each of the solutions contains three integer terms called *quantum numbers*. They are *n*, the *principal quantum number*, *l*, the *orbital angular momentum quantum number*, and m_l, the *magnetic quantum number*. The names of the last two quantum numbers pre-date modern

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

quantum chemistry. They are best regarded as labels rather than representing classical concepts such as the angular momentum of a solid body. The quantum numbers define the *state* of a system.

1.1.2 The energy of the electron

The principal quantum number, n, defines the energy of the electron. It can take integer 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 simple formula:

$$E = \frac{-A}{n^2} \tag{1.1}$$

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 for the lowest energy, $n = 1$, state. The states of higher energy each have n^2 different wavefunctions, all of which have the same energy, that is, there are four different wavefunctions corresponding to $n = 2$, nine different wavefunctions for $n = 3$, and so on. These wavefunctions are differentiated from each other by different values of the quantum numbers l and m_l, as explained below. 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). It is important to be aware of the fact that the electron can only take the exact energy values given by equation (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 (Chapter 14). 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:

$$\begin{aligned} E = E_1 - E_2 &= -2.179 \times 10^{-18}\left[\frac{1}{n_1^2} - \frac{1}{n_2^2}\right] \text{J} \\ &= -13.6\left[\frac{1}{n_1^2} - \frac{1}{n_2^2}\right] \text{eV} \end{aligned} \tag{1.2}$$

The energy of the photon emitted when the electron falls back from E_2 to E_1 is the same. 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:

$$E = h\nu = \frac{hc}{\lambda} \tag{1.3}$$

where h is the Planck constant. (Note that this equation applies to the transition between any two

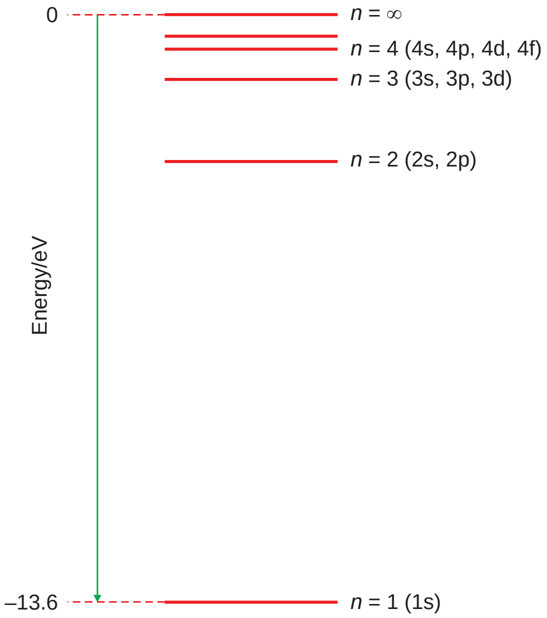


Figure 1.1 The energy levels available to an electron in a hydrogen atom.

[1] The unit of energy, *electron volt*, eV, is frequently used for atomic processes. 1 eV = 1.602×10^{-19} J.

energy levels on any atom, not just between energy levels on hydrogen or a hydrogenic atom.) 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 equation (1.2). 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:

$$E = \frac{-AZ^2}{n^2} \qquad (1.4)$$

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

1.1.3 Electron orbitals

The principal quantum number is not sufficient to determine the location of the electron in a hydrogen atom. In addition, the two other interdependent quantum numbers, l and m_l, are needed.

- l takes values of 0, 1, 2 . . . $(n - 1)$
- m_l takes values of 0, ±1, ±2 . . . $\pm l$

Each set of quantum numbers defines the *state* of the system and is associated with a wavefunction. For a value of $n = 1$, there is only one wavefunction, 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 $l = 2$. These states are referred to as *orbitals* and for historical reasons they 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*, and those with $l = 3$ are called *f orbitals* (Table 1.1).

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). For example, the L shell corresponds to the four orbitals associated with $n = 2$.

Table 1.1 Quantum numbers and orbitals for the hydrogen atom

n	l	m_l	Orbital	Shell
1	0	0	1s	K
2	0	0	2s	L
	1	−1, 0 +1	2p (3 orbitals)	
3	0	0	3s	M
	1	−1, 0 +1	3p (3 orbitals)	
	2	−2, −1, 0 +1, +2	3d (5 orbitals)	
4	0	0	4s	N
	1	−1, 0 +1	4p (3 orbitals)	
	2	−2, −1, 0 +1, +2	4d (5 orbitals)	
	3	−3, −2, −1,0, +1, +2, +3	4f (7 orbitals)	

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.

1.1.4 Orbital shapes

The *probability* of encountering the electron in a certain small volume of space surrounding a point with coordinates x, y and z is proportional to the square of the wavefunction at that point. With this information, it is possible to map out regions around the nucleus where the electron density is greatest.

The probability of encountering an electron in an s orbital does not depend upon direction but does vary with distance from the nucleus (Figure 1.2a,b,c). This probability peaks at a radial distance of 0.05292 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 called the *Bohr radius*. As the electron is promoted to the 2s, 3s, 4s orbitals, the maximum probability peaks further and further from the nucleus. Thus a high-energy electron is most likely to be found far 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 1.2d,e).

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 1.2, multiplied by an angular part. The maximum probability of finding the electron depends upon 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.

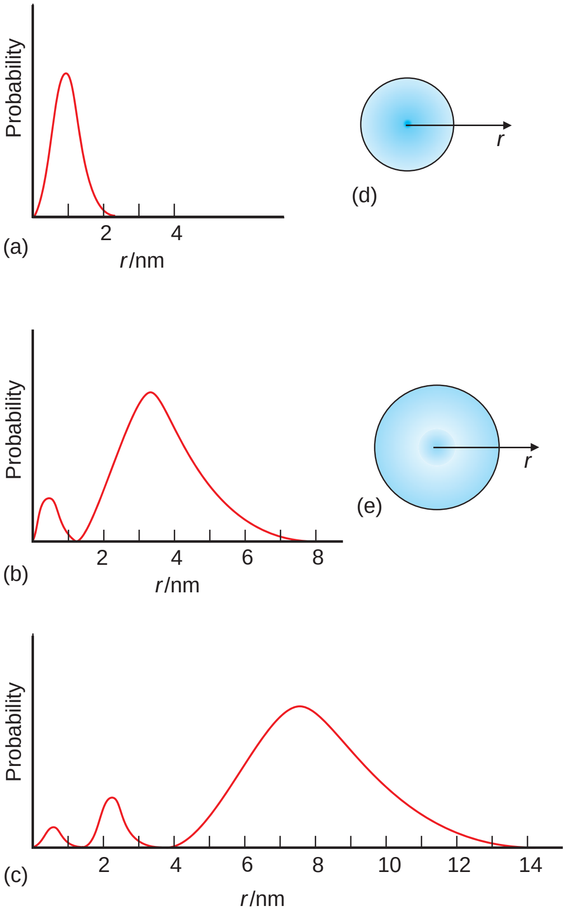


Figure 1.2 The probability of finding an electron at a distance r from the nucleus: (a) 1s; (b) 2s; (c) 3s. The boundary surfaces of the orbitals: (d) 1s; (e) 2s.

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 1.3). The corresponding orbitals are labelled $n\mathrm{p}_x$, $n\mathrm{p}_y$ and $n\mathrm{p}_z$, for

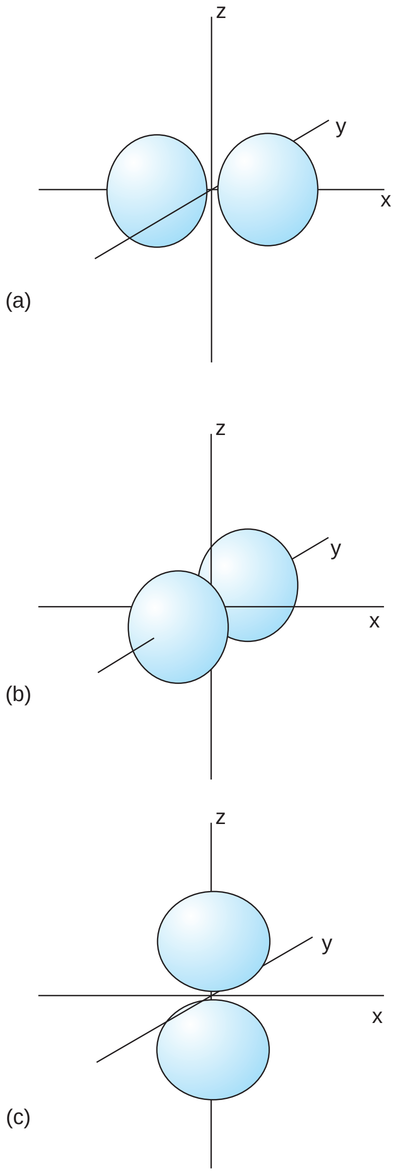


Figure 1.3 The boundary surfaces of the p orbitals: (a) p_x; (b) p_y; (c) p_z.

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.

The distribution of the electron in either the d or f orbitals is more complicated than those of the p orbitals. There are five d orbitals, and seven f orbitals. Three of the 3d set of wavefunctions 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 1.4). Except for the d_{z^2} orbital, two perpendicular planar nodes separate the lobes and intersect at the nucleus. In the d_{z^2} orbital, the nodes are conical surfaces.

1.2 Many-electron atoms

1.2.1 The orbital approximation

If we want to know the energy levels and electron distribution of an atom with a nuclear charge of $+Z$ surrounded by Z electrons, it is necessary to write out a more extended form of the Schrödinger equation that takes into account not only the attraction of the nucleus for each electron, but also the repulsive interactions between the electrons themselves. The resulting equation has proved impossible to solve analytically, but increasingly accurate numerical solutions have been available for many years.

The simplest level of approximation, called the *orbital approximation*, supposes that an electron moves in a potential due to the nucleus and the average field of all the other electrons present in the atom. This means that 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. In this approximation the orbital shapes are the same as for hydrogen, but the energy levels of all of the orbitals drop sharply as Z_{eff} increases (Figure 1.5). 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 significance. Moreover, the change of energy as Z increases justifies the approximation that the valence electrons of all atoms are at similar energies.

Although *shapes* of the orbitals are not changed from the shapes found for hydrogen, the radial part of the wave function is altered, and the *extension* of the 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 and *n*f 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 1.6). However, the energy differences between the higher energy orbitals are very small, and this simple ordering is not followed exactly for heavier atoms.

1.2.2 Electron spin and electron configuration

The results presented so far, derived from solutions to the simplest form of the Schrödinger equation, do not explain the observed properties of atoms exactly. In order to account for the discrepancy the electron is allocated a fourth quantum number called the *spin quantum number*, s. The spin quantum number has a value of $^1/_2$. The spin of an electron on an atom

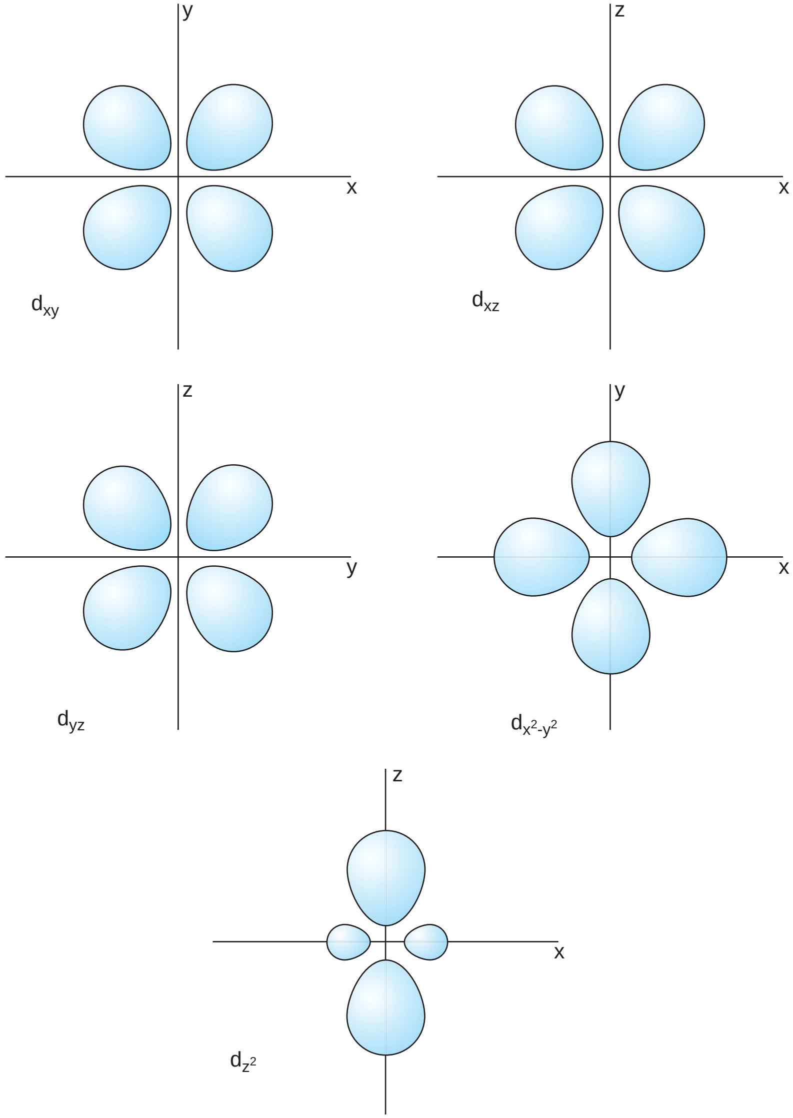


Figure 1.4 The boundary surfaces of the d orbitals: (a) d_{xy}; (b) d_{xz}; (c) d_{yz}; (d) $d_{x^2-y^2}$; (e) d_{z^2}.

can adopt one of two different directions, represented by a quantum number, m_s, which takes values of $+\frac{1}{2}$ or $-\frac{1}{2}$. These two spin directions have considerable significance in chemistry and physics and are frequently represented by ↑, *spin up*, or α, and ↓, *spin down* or β. Although the spin quantum number was originally postulated to account for certain experimental observations, it arises naturally in more sophisticated formulations of the Schrödinger equation that take into account the effects of relativity.

The *electron configuration* of an atom is the description of the number of electrons in each orbital, based upon the orbital model. This is usually given for the lowest energy possible, called the *ground state*. To obtain the electron configuration

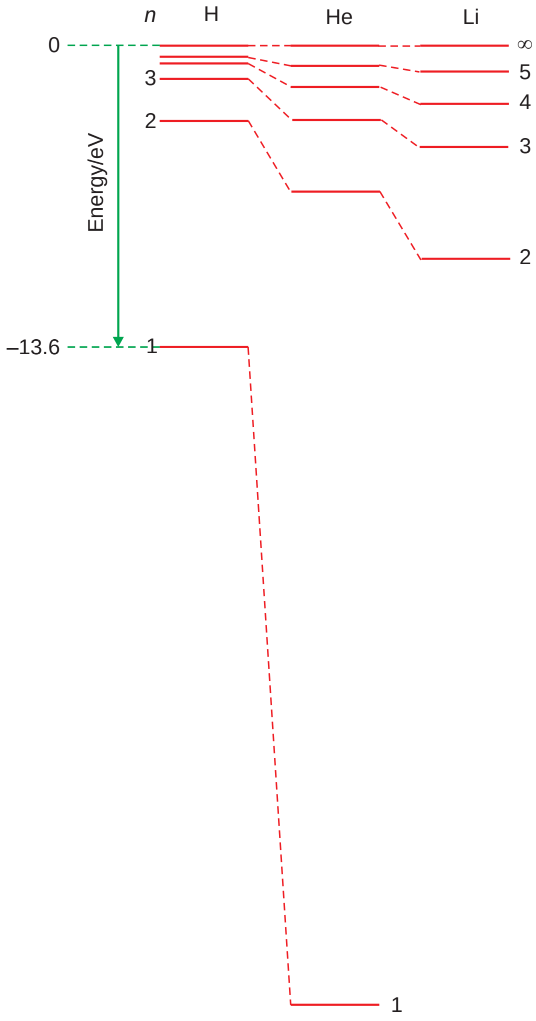


Figure 1.5 The decrease in energy of the orbitals of the first three elements in the periodic table, hydrogen, helium and lithium, as the charge on the nucleus increases.

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 systematically from the bottom up (Figure 1.6). This is called the *Aufbau* (or *building up*) *principle*. Before the configurations can be constructed, it is vital to know that each orbital can hold a maximum of two electrons, which must have opposite values of m_s, either $+^1/_2$ or $-^1/_2$. This fundamental feature of quantum mechanics is due to the *Pauli Exclusion Principle*: *no more than two electrons can occupy a single orbital, and if they do, the spins must be different, that is, spin up and spin down*. Two electrons in a single orbital are said to be *spin paired*.

The electron configurations of the elements can now be described. Each orbital can hold a maximum of two electrons, so that s orbitals can hold two electrons, the three p orbitals can hold a total of six electrons, the five d orbitals can hold a total of ten 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.*

Hydrogen has only one electron, and it will go into the orbital of lowest energy, the 1s orbital. The electron configuration is written as $1s^1$. Helium has two electrons and both can be placed in the 1s orbital to give an electron configuration $1s^2$. There is only one orbital associated with the $n = 1$ quantum number, hence the corresponding shell (K) is now filled. Further electrons must now be added to the L shell, corresponding to the 2s and 2p orbitals. Proceeding as before, the electron configuration of the next few elements are Li, $1s^2$ $2s^1$ or, in a compact notation [He] $2s^1$; Be, [He] $2s^2$; B, [He] $2s^2$ $2p^1$; C, [He] $2s^2$ $2p^2$ and so on. Note that it is normal practice to replace the configuration of filled inner shells corresponding to a noble gas by a contraction: [He] for helium, [Ne] for neon, [Ar] for argon, [Kr] for krypton, [Xe] for xenon and [Rn] for radon. Thus the electron configuration of Rb is written [Kr] $5s^1$, signifying a single electron outside of the K, L, M and N closed shells that make up the configuration of the noble gas krypton.

1.2.3 The periodic table

The periodic table (Figure 1.7 and front endpaper), originally an empirical arrangement of the elements in terms of chemical properties, is understandable in

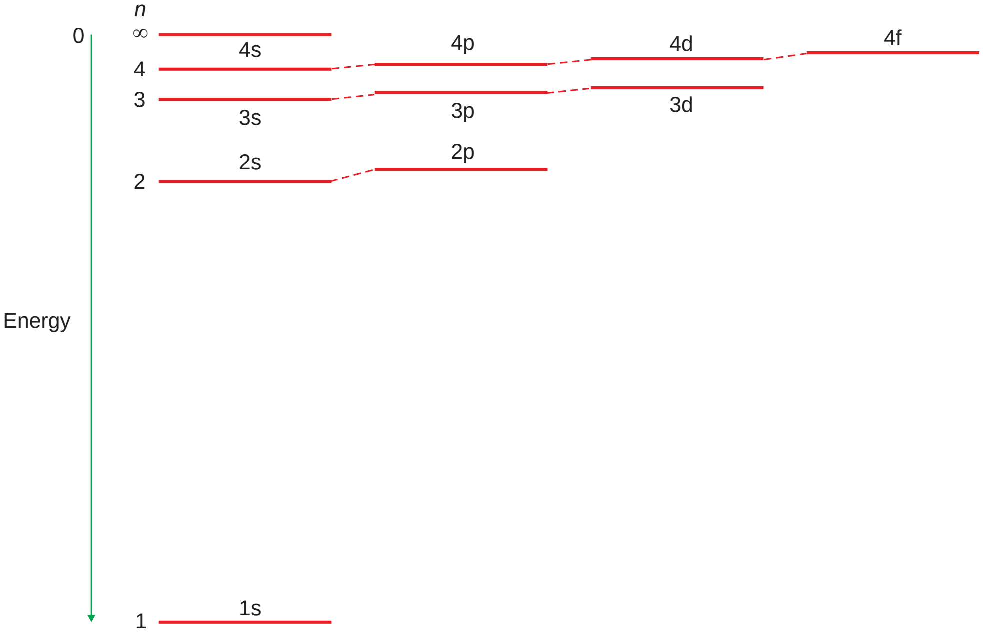


Figure 1.6 The schematic energy levels for a light, many-electron atom.

terms of the electron configurations just discussed. The chemical and many physical properties of the elements are simply controlled by the outer (valence) electrons. The valence electron configuration varies in a systematic and repetitive way as the various shells are filled. This leads naturally to the periodicity displayed in the periodic table. For example, the filled shells are very stable

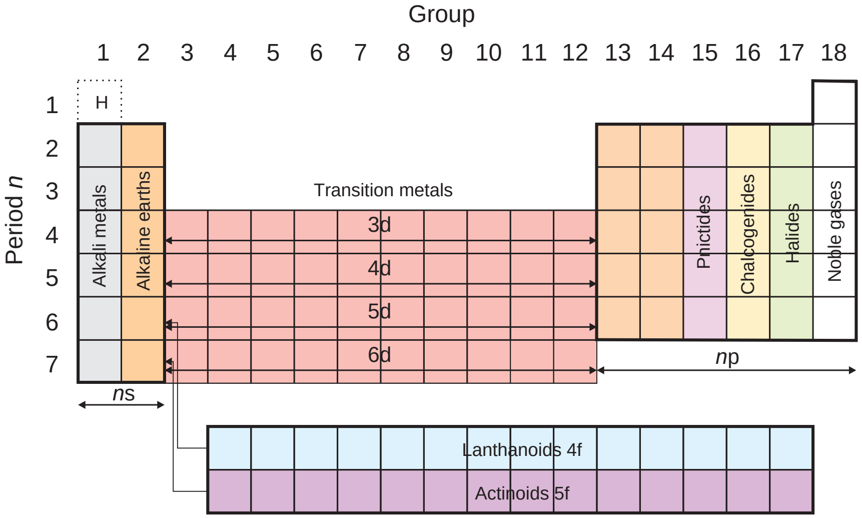


Figure 1.7 The relationship between electronic configuration and the periodic table arrangement.

configurations and only take part in chemical reactions under extreme conditions. The atoms with this configuration, the *noble gases*, are placed in group 18 of the periodic table. A new noble gas appears each time a shell is filled. Following any noble gas is an element with one electron in the outermost s orbital, lithium, sodium, potassium, and so on. These are the *alkali metals*, found in Group 1, and once again, a new alkali metal is found after each filled shell. Similarly, the *alkaline earth* elements, typified by magnesium, calcium and strontium, listed in Group 2 of the periodic table, all have two valence electrons, both in the outermost s orbital. Thus, the periodic table simply expresses the Aufbau principle in a chart format.

The outermost electrons take part in chemical bonding. The main group elements are those with electrons in outer s and p orbitals giving rise to strong chemical bonds (Chapter 2). The valence electron configuration of all the elements in any group is identical, indicating that the chemical and physical properties of these elements will be very similar. The d and f orbitals are shielded by s and p orbitals from strong interactions with surrounding atoms and do not take part in strong chemical bonding. Those elements with partly filled d orbitals are called *transition metals*, typified by iron and nickel, while those with partly filled f orbitals are called the *lanthanoids* (4f) or *actinoids* (5f). The electrons in these orbitals are responsible for many of the interesting electronic, magnetic and optical properties of solids.

1.3 Atomic energy levels

1.3.1 Spectra and energy levels

Spectra are a record of transitions between electron energy levels. Each spectral line can be related to the switch from one energy level to another. 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 equation (1.3):

$$\Delta E = h\nu = \frac{hc}{\lambda}$$

where h is Planck's constant.

The electron configurations described above, which essentially apply to a single electron moving under the combined electrostatic field of the nucleus and all of the other electrons, are not able to account for the observed transitions. A more complex model of the atom is required to derive the possible energy levels appropriate to any electron configuration. There are a number of ways of deriving these *many-electron* quantum numbers and the associated *many-electron states* of the atom.

The approach most frequently encountered, called *Russell-Saunders coupling*, makes the approximation that the electrostatic repulsion between electrons is the most important energy term. To obtain revised configurations, all of the individual s values of the electrons are summed to yield a total spin angular momentum quantum number S. (Note that one-electron quantum numbers are written in lower case, while many-electron quantum numbers are written in upper case.) Similarly, all of the individual l values for the electrons present are summed to give a total orbital angular momentum quantum number L. The values of S and L can also be summed to give a total angular momentum quantum number J.

An alternative approach to Russell-Saunders coupling is to assume that the interaction between the orbital angular momentum and the spin angular momentum is the most important. This interaction is called *spin–orbit coupling*. 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 *j-j coupling*.

Broadly speaking, Russell-Saunders coupling works well for lighter atoms and *j-j* coupling for heavier atoms. Other coupling schemes have also been worked out which find use for medium and heavy atoms.

1.3.2 Terms and term symbols

In the Russell-Saunders coupling scheme, the total spin angular momentum quantum number, $S(2)$, for two electrons is obtained by combining the individual quantum numbers s_1 and s_1 in the following way:

$$S(2) = (s_1 + s_2), (s_1 + s_2 - 1), \ldots |s_1 - s_2|$$

where $|s_1 - s_2|$ is the modulus (absolute value, taken as positive) of $s_1 - s_2$. As s_1 and s_2 are equal to $^1\!/_2$, then $S(2) = 1$ or 0.

In order to obtain the value of S for three electrons, $S(3)$, the value for two electrons, $S(2)$, is combined with the spin quantum number ($s_3 = {}^1\!/_2$) of the third electron, in the same way:

$$S(3) = (S(2) + {}^1\!/_2), (S(2) + {}^1\!/_2 - 1), \ldots$$
$$|S(2) - {}^1\!/_2|$$

Both of the values for $S(2)$, 1 and 0, are permitted, so we obtain:

- $S(3) = 1 + {}^1\!/_2, 1 + {}^1\!/_2 - 1 = {}^3\!/_2, {}^1\!/_2$
- $S(3) = 0 + {}^1\!/_2 = {}^1\!/_2$

Thus S(3) can take values of $^3\!/_2$ or $^1\!/_2$.

The same procedure, called the *Clebsch-Gordon rule*, is used to obtain the S values for four electrons, by combining s_4 with S(3), and so on. 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. As all electrons in filled shells are spin-paired, it is only necessary to count the spins in the outer unfilled orbitals to obtain values of S for the atom as a whole.

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-Gordon 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. As before, all closed shells have zero angular momentum, so in deciding L, only outer electrons in unfilled shells need to be counted.

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*. It represents a set of energy levels, 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.

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:

$$S = {}^1\!/_2 + {}^1\!/_2, {}^1\!/_2 - {}^1\!/_2 = 1, 0.$$
$$2S + 1 = 3 \text{ or } 1.$$
$$L = (1 + 1), (1 + 1 - 1), |1 - 1| = 2, 1, 0 (\mathrm{D, P, S})$$

The total number of possible terms for the two p electrons is given by combining these values:

$$^3\mathrm{D}, {}^3\mathrm{P}, {}^3\mathrm{S}, {}^1\mathrm{D}, {}^1\mathrm{P}, {}^1\mathrm{S}$$

Note that 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 method is straightforward but time-consuming and is not described here) the allowed terms are:

$$^3\mathrm{P}, {}^1\mathrm{D}, {}^1\mathrm{S}$$

In general the *energies* of the terms are difficult to obtain and must be calculated using quantum mechanical procedures. Fortunately the lowest energy (ground state) term is easily found, using Hund's *second rule*: *(a) The term with the lowest energy has the highest multiplicity; (b) 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 for 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 (Figure 1.8).
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 $-{}^1\!/_2$, to give 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 (Figure 1.8), the ground term of both the $2p^2$ and $2p^4$ configurations is ^{3}P.

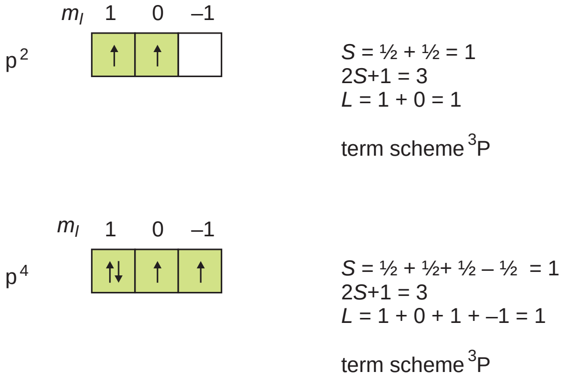


Figure 1.8 The derivation of ground state term symbols for p^2 and p^4 electron configurations.

1.3.3 Levels

The term symbol does not account for the true energy level complexity found in atoms. For example, 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*. A similar, but different, change, called the *Stark effect*, arises in the presence of a strong electric field.

These result from the interaction between the spin and the orbital momentum (spin–orbit coupling) that is ignored in Russell-Saunders coupling. A new quantum number, J, is needed. It is given by:

$$J = (L + S), (L + S - 1), \ldots |L - S|$$

where $|L - S|$ is the modulus (absolute value, taken as positive) of L minus S. Thus the term ^{3}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 level, a doublet, two, a triplet three, and so on. Thus, ground state term ^{3}P is composed of three levels: ${}^3\mathrm{P}_0$, ${}^3\mathrm{P}_1$ and ${}^3\mathrm{P}_2$. The magnitude of the energy difference between these levels depends upon the strength of the interaction between L and S.

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 $g_\mathrm{J}\ \mu_\mathrm{B}\mathbf{B}$, where g_J is the Landé g-value:

$$g_J = 1 + \frac{(\mathrm{J}(\mathrm{J}+1) - \mathrm{L}(\mathrm{L}+1) + \mathrm{S}(\mathrm{S}+1))}{2\mathrm{J}(\mathrm{J}+1)}$$

μ_B is a fundamental physical constant, the Bohr magneton, and $\mathbf{B}$ is the magnetic induction. (For more information, see section 12.1.3.) 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,*

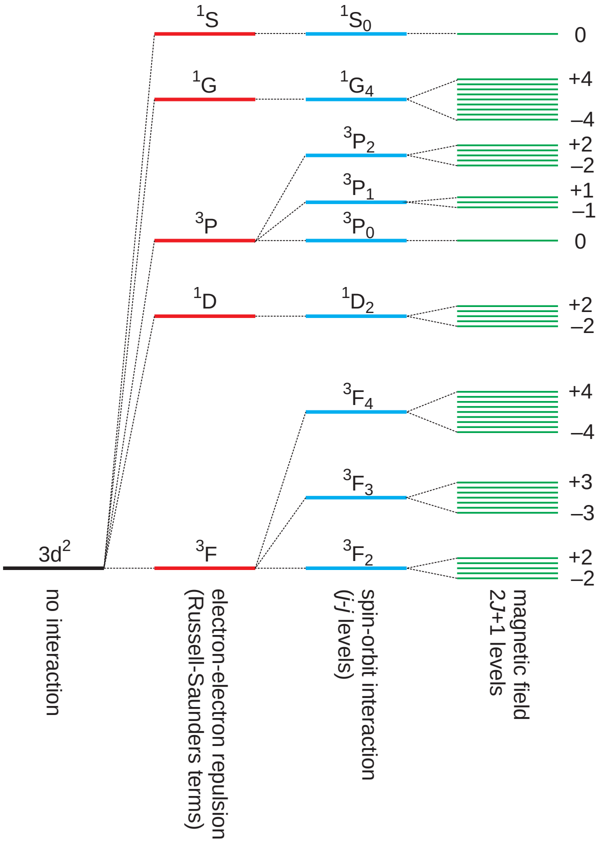


Figure 1.9 The evolution of the energy levels of an atom with a $3d^2$ electron configuration, taking into account increasing interactions. The energy scales are schematic.

and that with the highest J value if the valence shell is more than half full.

These increasing degrees of complexity are illustrated for a $3d^2$ transition metal atom in Figure 1.9. At the far left of the figure, the electron configuration is shown. This is useful chemically, but is unable to account for the spectra of the atom. The Russell-Saunders terms that arise from this arrangement are given to the right of the configuration. In Russell-Saunders coupling the electron–electron repulsion is considered to dominate the interactions. The terms are spilt further if spin–orbit coupling (j-j coupling), is introduced. The number of levels that arise is the same as the multiplicity of the term, $2S + 1$. Finally, the levels are split further in a magnetic field to give $2J + 1$ levels. The magnitude of the splitting is proportional to the magnetic field, and the separation of each of the new energy levels is the same.

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.

1.3.4 Electronic energy level calculations

The allowed energies of the electrons in an atom are found by solving the Schrödinger equation. The solution of this equation is possible for hydrogen, but is impossible even for the next atom, helium, with two electrons attached to a single nucleus. The reason for this is that each electron is attracted to the nucleus and repelled by the other electron. Thus the electrons do not move independently of one another, but their motion is *correlated*. This correlation term, which must be included in the calculation, is the central problem.

There are a number of ways of approximately calculating the electron energy levels. One will be described here, the *Hartree-Fock* procedure, because it provides a simple picture of atomic structure, and leads naturally to estimates of the electronic energies of molecules and non-molecular solids (Chapter 2). Take helium as the simplest example. In this case the wavefunction describing the two-electron atom, $\psi(\mathbf{r}_1, \mathbf{r}_2)$, which is a function of the position of the electrons, $\mathbf{r}_1$ and $\mathbf{r}_2$, is the product of (say), two hydrogenic 1s orbitals:

$$\psi(\mathbf{r}_1, \mathbf{r}_2) = \phi 1s(\mathbf{r}_1)\alpha\, \phi 1s(\mathbf{r}_2)\beta - \phi 1s(\mathbf{r}_1)\beta\, \phi 1s(\mathbf{r}_2)\alpha$$

where $\phi 1s(\mathbf{r}_1)\alpha$ means that electron 1 is in the 1s orbital with spin α, and $\phi 1s(\mathbf{r}_2)\beta$ means that electron 2 is in the 1s orbital with spin β, and so on. (Note that any orbital functions can be chosen, not just hydrogenic orbitals, if they make computation easier.) Now electrons are indistinguishable and they have a spin, ↑ or ↓. Moreover, two electrons with the same spin cannot occupy the same orbital. To take this into account, the wavefunction of the system must change sign when any two electrons are exchanged, called *exchange symmetry*. This restriction on swapping electron positions lowers the energy by an amount called the *exchange energy*.

To solve the equation, electron 1 is supposed to experience an *effective* potential due to the nucleus and the charge density contributed by electron 2. The function $\phi(\mathbf{r}_2)$ is chosen and the effective potential is calculated. The Schrödinger equation written using the effective potentials is the *Hartree-Fock equation* for He. This approximate Schrödinger equation is used to calculate $\phi(\mathbf{r}_1)$ and the energy ε_1. In general the new $\phi(\mathbf{r}_1)$ will be different than the original choice because that was a hydrogenic function that ignored the potential due to the other electron. The process is now repeated for the orbital of electron 2 using the revised orbital of electron 1. This is continued until the revised input does not lead to any further change in the output orbitals, at which point the orbitals are *self-consistent*. The orbitals and energies are called *Hartree-Fock self-consistent field* (HF-SCF) orbitals and energies.

The main shortcoming of the method is that electronic correlation has been completely ignored and the results lack an important energy term. When energies are known experimentally from, for example, spectra, the correlation energy can be derived by subtracting the Hartree-Fock energy:

$$\text{Correlation energy} = E_{\text{exp}} - E_{\text{HF}}$$

Further reading

Elementary chemical concepts and an introduction to the periodic table are clearly explained in the early chapters of:

Atkins, P. and Jones, L. (1997) *Chemistry*, 3rd edn. W.H. Freeman, New York.

McQuarrie, D.A. and Rock, P.A. (1991) *General Chemistry*, 3rd edn. W.H. Freeman, New York.

The outer electron structure of atoms is described in the same books, and in greater detail in:

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

Shriver, D.F., Atkins, P.W. and Langford, C.H. (1994) Chapter 1, *Inorganic Chemistry*, 2nd edn. Oxford University Press, Oxford.

The quantum mechanics of atoms is described lucidly by:

McQuarrie, D.A. (1983) *Quantum Chemistry*. University Science Books, Mill Valley, CA.

An invaluable dictionary of quantum mechanical language and expressions is:

Atkins, P.W. (1991) *Quanta*, 2nd edn. Oxford University Press, Oxford.

Problems and exercises

Quick quiz

1 A wavefunction is
(a) A description of an electron.
(b) An atomic energy level.
(c) A solution to the Schrödinger equation

2 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.

3 The Pauli Exclusion principle leads to the conclusion that
(a) The position of an electron cannot be specified with limitless precision.
(b) Only two electrons of opposite spin can occupy a single orbital
(c) No two electrons can occupy the same orbital.

4 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.

5 The outer electron configuration of the noble gases is
(a) $n\mathrm{s}^2n\mathrm{p}^6$.
(b) $n\mathrm{s}^2n\mathrm{p}^6(n+1)\mathrm{s}^1$.
(c) $n\mathrm{s}^2n\mathrm{p}^5$.

6 The valence electron configuration of the alkali metals is
(a) $n\mathrm{s}^2$.
(b) $n\mathrm{p}^1$.
(c) $n\mathrm{s}^1$.

7 The valence electron configuration of carbon is
(a) $1\mathrm{s}^2 2\mathrm{p}^2$.
(b) $2\mathrm{s}^2 2\mathrm{p}^2$.
(c) $2\mathrm{s}^2 2\mathrm{p}^4$.

8 The valence electron configuration of calcium, strontium and barium is
(a) $n\mathrm{s}^2n\mathrm{p}^2$.
(b) $n\mathrm{s}^2$.
(c) $(n-1)\mathrm{d}^1n\mathrm{s}^2$.

9 What atom has filled K, L, M and N shells?
(a) Argon.
(b) Krypton.
(c) Xenon.

10 How many electrons can occupy orbitals with $n=3$, $l=2$?
(a) 6 electrons.
(b) 10 electrons.
(c) 14 electrons.

11 How many permitted l values are there for $n=4$?
(a) One.
(b) Two.
(c) Three.

12 How many electrons can occupy the 4f orbitals?
(a) 14.
(b) 10.
(c) 7.

13 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.

14 A term symbol is
(a) A label for an atomic energy level.
(b) A label for an orbital.
(c) A description of a configuration.

15 The many-electron quantum number symbol D represents
(a) $L=1$.
(b) $L=2$.
(c) $L=3$.

16 An atom has a term ^{1}S. What is the value of the spin quantum number, S?
(a) $\frac{1}{2}$.
(b) 0.
(c) 1.

17 An atom has a term ^{1}S. What is the value of the orbital quantum number, L?
(a) 2.
(b) 1.
(c) 0.

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 ^{4}S. What are the possible values of J? What is the ground state level?

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

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

1.15 What is the splitting g_J, for sulphur, 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 1.9, for the ground state of a chlorine atom, with a ground state $^2P_{3/2}$.

2

Chemical bonding

- How big is an ion?
- What is covalent bonding?
- What are energy bands?

Theories of chemical bonds have several roles. Firstly, they must explain the *cohesion* between atoms. In addition, they must account for the concept of chemical *valence*. Valence is the notion of the *combining power* of atoms and accounts for chemical formulae and geometry. Ionic and covalent bonds are essentially theories of chemical valence. Any theory of metallic bonding must explain not only the distinctive *physical properties* of metals but should also account for the fact that the majority of the elements in the periodic table are metals.

Chemical bonds arise in the interactions between the outer electrons, the *valence electrons*, on the combining of atoms, and chemical bonding is essentially an aspect of quantum theory. For the purposes of explanation, three approximations can be used. In ionic bonding the valence electrons are firmly attached to atomic nuclei and electrostatic interactions provide the interatomic glue. Covalent bonding accounts for molecules, and in this model valence electrons are shared in orbitals that extend over several of the atomic nuclei present to form (covalent) bonds between them. This concept is taken to its greatest extent in metallic bonding. Here the valence electrons are more or less free to occupy all of the material, whether solid or liquid. It must be remembered, though, that these three extremes form parts of a continuum, and a true image of bonding, especially with respect to structure–property relations, often needs a combination of more than one approach.

2.1 Ionic bonding

2.1.1 Ions

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 on the cations, written as a superscript, is equal to the number of electrons lost. For example, the alkali metals with an outer electron configuration of ns^1 form M^+ cations, while the alkaline earths, with an outer electron configuration ns^2, form M^{2+} cations. The transition metal ions generally have a number of d electrons in their outer shell, and because the energy difference between the various configurations is small, are able to lose a variable number of

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

electrons. For example, vanadium may exist as V^{2+}, V^{3+}, V^{4+} or V^{5+} cations.

Atoms in at the lower part of groups 13, 14 and 15 are able to take two ionic states. For instance, tin has an outer electron configuration of [Kr] $4d^{10}$ $5s^2$ $5p^2$. Loss of the two p electrons will generate the Sn^{2+} state, with a configuration of [Kr] $4d^{10}$ $5s^2$. Further loss of the two s electrons will produce the stable configuration, [Kr] $6d^{10}$, of Sn^{4+}. The atoms that behave in this way are characterised by two valence states, separated by a charge difference of 2+. The examples are indium (1+, 3+), thallium (1+, 3+), tin (2+, 4+), lead (2+, 4+), antimony (3+, 5+) and bismuth (3+, 5+). When present, the pair of s electrons has important physical and chemical effects, and ions with this configuration are called *lone pair ions.*

Non-metals, with almost complete noble gas configurations, tend to gain electrons and form negatively charged *anions*. For example, the halogens, with an outer electron configuration of np^5, tend to pick up an electron to form X^- anions, while the chalcogens, with an outer electron configuration of np^4, tend to pick up two electrons to form X^{2-} anions. The charge on the anions, written as a superscript, is equal to the number of electrons gained. 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.

Ionic bonding is due to the fact that 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. The formula of an ionic compound follows directly from the idea that cations have integer positive charges, anions have integer negative charges, and ionic compounds are neutral. Thus a monovalent cation can combine with a monovalent anion to give neutral compounds MX, such as potassium chloride, KCl. Similarly, a divalent cation will combine in a 1:1 ratio with a divalent anion, as in magnesium oxide, MgO; and a trivalent cation will combine with a trivalent anion in the same way, as in aluminium nitride, AlN. It is necessary for two monovalent cations to combine with a divalent anion to form a neutral unit M_2X, for example, sodium oxide, Na_2O. Similarly, a divalent cation will need to combine with two monovalent anions to give neutral MX_2, such as magnesium chloride, $MgCl_2$. Trivalent cations need three monovalent anions, as in aluminium chloride, $AlCl_3$, and two trivalent cations need to combine with three divalent anions to give a neutral unit, for example, aluminium oxide, Al_2O_3. Other more complex formulae can be derived in the same way; spinel, $MgAl_2O_4$ can be thought of as $MgO + Al_2O_3$, and perovskite, $CaTiO_3$, can be rationalised as $CaO + TiO_2$.

2.1.2 Ionic size and shape

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. This technique only gives 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 2.1).

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 sulphur ions in octahedral coordination. Ideally, tables of cationic radii should apply to a specific anion and coordination geometry.

+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 2.1 Ionic radii for ions commonly found in solids. The cation radii given are for ions octahedrally coordinated to oxygen, except for those marked (t), which are in tetrahedral coordination, and for the anions. For the definition of *high spin configuration,* see Section 12.2.2.

Several trends in ionic radius are apparent:

- Cations are usually smaller than anions, the main exceptions being the largest alkali metal and alkaline earth metal cations, all larger than F$^-$. The reason for this is that removal of electrons to form cations leads to a contraction of the electron orbital clouds due to the relative increase in nuclear charge. Similarly, addition of electrons to form anions leads to an expansion of the charge clouds due to a relative decrease in the nuclear charge.
- The radius of an ion increases with atomic number.
- The 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].
- Successive valence increases will decrease the radius. For example, Fe^{3+} is smaller than Fe^{2+}.
- Increase in negative charge has a smaller effect than increase in positive charge. For example, F$^-$ is similar in size to O^{2-} and Cl$^-$ is similar in size to S^{2-}.

While the majority of the ions of elements can be considered to be spherical, the lone pair ions In$^+$, Tl$^+$, Sn $^{2+}$, Pb^{2+}, Sb^{3+} and Bi^{3+} are definitely not so. These ions tend to adopt a distorted trigonal bipyramidal coordination, and it is hard to assign a unique radius to such ions.

Complex ions, such as CO_3^{2-} and NO_3^-, are not spherical, although at high temperatures, rotation often makes them appear so.

2.1.3 Lattice energies

A considerable advantage of the ionic bonding model is that ionic interaction energies can be fairly easily assessed. In pre-computer days often the only significant way of estimating the stability and thermodynamic properties of structures was via the calculation of lattice energies.

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:

$$E_e = \frac{(+e)(-e)}{4\pi\varepsilon_0 r} = \frac{-e^2}{4\pi\varepsilon_0 r} \qquad (2.1)$$

where the point charges on the interacting species are $\pm e$, and the distance separating the charges is r. The energy is zero when the ions are infinitely far from each other and a negative overall energy implies stability.

The electrostatic energy of an ionic crystal, called the *Madelung energy*, has a form identical to equation (2.1) 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.

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

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

$$E_e = N_A \left[\frac{-e^2}{4\pi\varepsilon_0 r}\right] \alpha Z^2$$

where N_A is Avogadro's constant and r is the nearest 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.

Table 2.1 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	$M^{3+}{}_2X_3^{2-}$	Al_2O_3	1.669

In such cases it is simplest to separate out the charge contribution and stoichiometry to give a *reduced Madelung constant*, which is a purely geometric term. The reduced Madelung constant of a compound M_mX_n, is defined by:

$$E_e = -N_A \frac{e^2}{4\pi\varepsilon_0 r}\alpha(Z_M Z_X)\frac{m+n}{2} \qquad (2.2)$$

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 2.1). 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:

$$\alpha Z_M Z_X (m+n)/2 = \mathrm{A}$$
$$\alpha (m+n)/2 = \mathrm{M}$$

Electrostatic attraction will increase as cations and anions 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 as a function of the form:

$$E_r = \frac{B}{r^n}$$

where B and n are empirical constants.

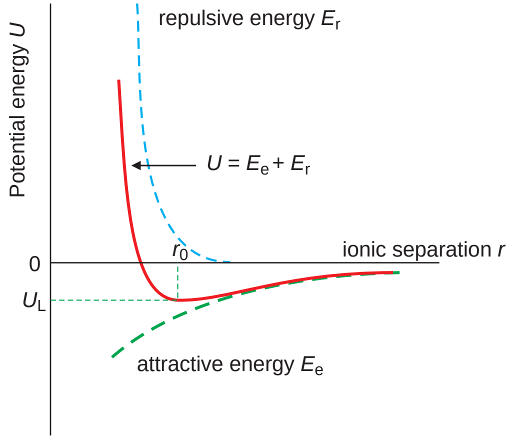


Figure 2.2 The total potential energy between monovalent ions, U, as a function of the ionic separation, r. The total energy is the sum of attractive and repulsive potential energy terms. The lattice energy, U_L, corresponds to the minimum in the total energy curve, reached at an interionic separation of r_0.

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

$$U = E_e + E_r = \frac{-N_A\,\alpha\,Z^2e^2}{4\pi\varepsilon_0 r} + \frac{N_A B}{r^n}$$

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 2.2). Thus, we can write:

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

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

$$U_L = \left(\frac{-N_A\alpha Z^2e^2}{4\pi\,\varepsilon_0 r_0}\right)\left(1 - \frac{1}{n}\right)$$

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 above, gives the *Born-Mayer equation* for the lattice energy:

$$U_L = \left(\frac{N_A\alpha Z^2e^2}{4\pi\varepsilon_0 r_0}\right)(1 - (r^*/r_0))$$

2.1.4 Atomistic simulation

The analysis in the previous section 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:

$$E_e = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

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 2.2).

(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 that vary from one ion pair to another.

Many ions are polarizable (sections 11.1 and 14.4.2). 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:

$$E_\mathrm{p} = \tfrac{1}{2}k\,\mathrm{d}r^2$$

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

$$E_\mathrm{p} = \tfrac{1}{2}k_2\,\mathrm{d}r^2 + (1/24)\mathrm{k}_4\,\mathrm{d}r^4$$

where dr 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 are 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 atomic array which are easily compared with 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, the atomic trajectories either following Newton's laws of motion (*molecular* or *lattice dynamics*) or else are 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. Recently, molecular dynamics simulations have been used to study the enormously complex problem of protein folding.

2.2 Covalent bonding

In the previous section, electrons were essentially ignored, apart from their peripheral role in providing electrostatic charges. This approach is unable to explain the properties of most molecules, in which the atoms are described as being linked by *covalent bonds*. To understand molecular properties, calculation of electron energies is needed. The first quantum mechanical rationalisation of this was *valence bond theory*, and this is outlined because it gives a clear idea of the origin of basic molecular shape. However, the energies of the electrons in a molecular assembly are more easily calculated by *molecular orbital theory*, described later.

2.2.1 Valence bond theory

2.2.1.1 Single bond formation

Valence bond theory starts with the idea that a covalent bond consists of a pair of electrons shared between the bound atoms. The electrons, one on each atom, are in half-filled orbitals. In addition the *direction* of a bond will be such as to make the orbitals of the bonding electrons overlap as much as possible, and the strongest bonds are formed when the overlapping of the orbitals is at a maximum.

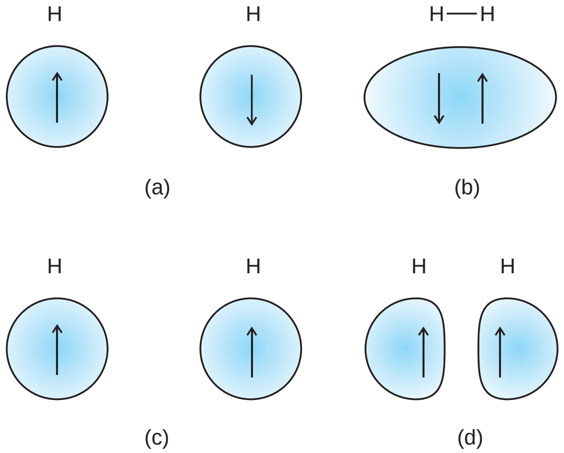


Figure 2.3 Isolated hydrogen atoms can have electrons in an (a) antiparallel, or (c) parallel, spin arrangement. (b) When orbitals (a) overlap, the electron density accumulates between the nuclei to form a covalent bond. (d) When orbitals (c) overlap, the electron density is low between the nuclei and no bond forms.

The simplest covalent bond is, perhaps, the single bond between two hydrogen atoms that make up the hydrogen (H_2) molecule. The bond can be envisaged to form in the following way. 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 as the interatomic distance is reduced, both electrons begin to experience attraction from both nuclei. There is also electrostatic repulsion between the two electrons, but the attraction preponderates. 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 2.3a,b). If the spins of the two electrons are parallel, the Pauli Exclusion Principle stipulates that it is energetically unfavourable for the electron clouds to overlap. The electron density avoids the region between the nuclei and bonding does not occur (Figure 2.3c,d).

Any partly filled orbitals can overlap to form bonds in this way, but the most commonly occurring examples are s and p orbitals. A bond formed by s orbitals, end-on p orbitals, or by s and p orbitals, has rotational symmetry about the bond axis, which is the line joining the nuclei contributing the electrons (Figure 2.4). As a result, a cross-section through the bond looks like an s orbital and, in recognition of this symmetry relationship, they are termed σ *bonds*.

A different type of bonding orbital can be formed between two p orbitals, each with a single electron and with opposed spins, approaching each other sideways on (Figure 2.5). In this case, the pile-up of the electron density occurs on either side of a nodal plane in which the two nuclei are situated. Each individual bond has two lobes, one lobe to one side of the internuclear axis and one lobe to the other. In this configuration the bond looks like a p orbital in cross-section and are termed π *bonds*.

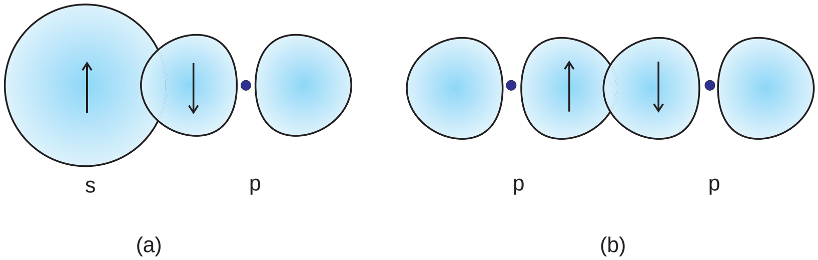


Figure 2.4 (a) A covalent σ bond formed by the overlap of an s orbital and an end-on p orbital when the two electrons have antiparallel spins. (b) A covalent σ bond formed by the overlap of two end-on p orbitals when the electrons have antiparallel spins.

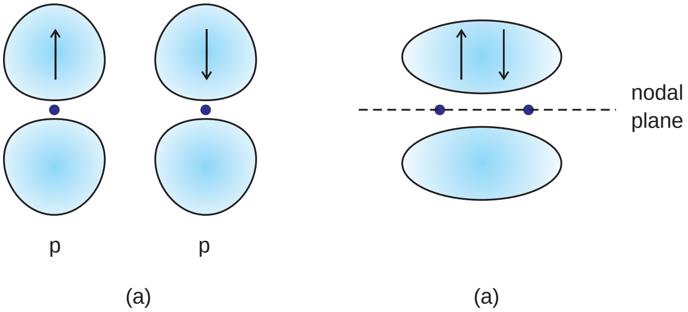


Figure 2.5 (a) Two sideways-on p orbitals containing electrons with antiparallel spins. (b) A π bond formed by the sideways-on overlap of p orbitals.

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.

2.2.1.2 *Orbital hybridisation*

From what has been said so far, one would expect carbon, with electron configuration $1s^2\ 2s^2\ 2p^2$, to form compounds with two bonds at 90° to one another. This is because there are two p orbitals, each containing one electron that can overlap with other partly filled orbitals on other atoms. Thus, in reaction with hydrogen, an angular molecule of formula CH_2 should form. Now the common valence of carbon is 4, and in the latter half of the 19th century, organic chemists established beyond doubt that in small molecules formed by carbon, the four bonds are directed away from the C atom towards the corners of a tetrahedron. This discrepancy was resolved by introducing the concept of *orbital hybridisation*.

Hybridisation involves combining orbitals in such a way that they can make stronger bonds (with greater overlap) than the atomic orbitals depicted earlier. To illustrate this, suppose that we have one s and one p orbital available on an atom (Figure 2.6a). These could form two bonds, but neither orbital can utilise all of its overlapping ability when another atom approaches. However, under the influence of bonding atoms, the electron waves that in shorthand are designated as an s and a p orbital can combine, or *hybridise*, to produce *two* new orbitals pointing in opposite directions (Figure 2.6b). Each resulting

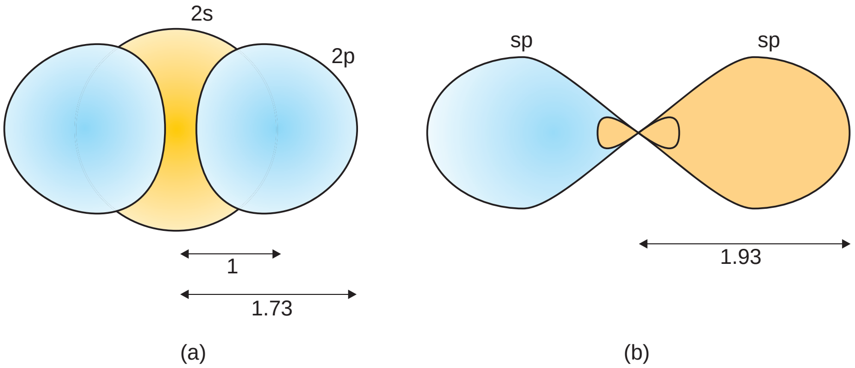


Figure 2.6 (a) The 2s and $2p_x$ orbitals on an atom. (b) Two sp hybrid orbitals formed by combining the 2s and $2p_x$ orbitals. The orbitals point directly away from each other.

hybrid orbital is composed of one large lobe and one very small lobe, which can be thought of as the positive s orbital adding to the positive lobe of the p orbital to produce a large lobe, and the positive s orbital adding to the negative lobe of the p orbital to give a small lobe. The overlapping power of the new combination is found to be significantly larger than that of s or p orbitals, because the extension of the hybrid orbitals is 1.93, compared with 1.0 for an s orbital and 1.73 for a p orbital. Although it requires energy to form the hybrid configuration, this is more than recouped by the stronger bonding that results. Since the hybrid orbitals are a combination of one s and one p orbital, they are called *sp-hybrid orbitals*. The large lobe on each of the hybrid orbitals can be used for bond formation, and bond angles of 180° are expected. Mercury makes use of sp-hybrid bonds in the molecule $HgCl_2$.

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. 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° (Figure 2.7). The *sp^2-hybrid orbitals* have an overlapping power of about twice that of s orbitals. This type of bonding is found in boron trichloride. It is also commonly encountered in borosilicate glasses, where the boron atoms are linked to three oxygen atoms at the corners of an equilateral triangle by sp^2-hybrid bonding orbitals.

It is now possible to return to the case of carbon. 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, *sp^3-hybrid orbitals* are possible. Calculation shows that the resulting four bonds will point towards the corners of a tetrahedron, at angles of 109.5° to each other (Figure 2.8). These angles are just the tetrahedral angles found for methane, carbon tetrachloride and many carbon compounds. Moreover, the hybrid 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 kJmol^{-1}.

Hybridization explains the geometry of ammonia and water. Nitrogen has an outer electron configuration of $2s^2\ 2p^3$, and oxygen has an outer electron configuration of $2s^2\ 2p^4$. Although bonding to three or two atoms respectively is possible, using the available p orbitals, stronger bonds result if

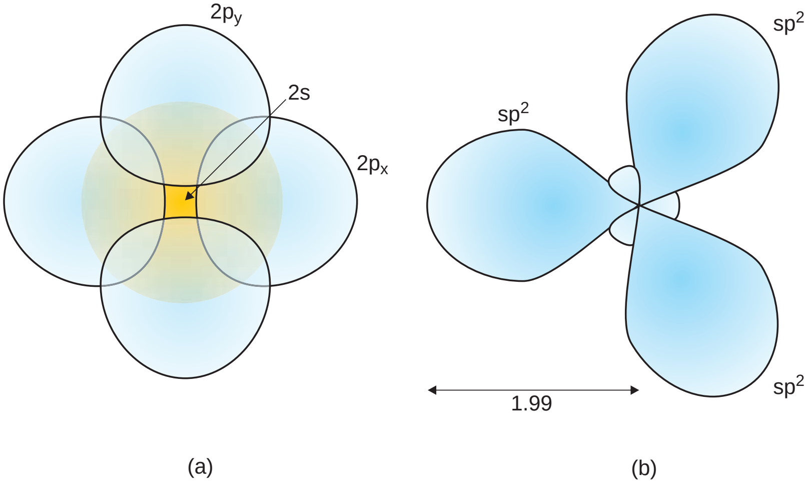


Figure 2.7 (a) The 2s, $2p_x$ and $2p_y$ orbitals on an atom. (b) Three sp^2 hybrid orbitals formed by combining the three original orbitals. The orbitals are arranged at an angle of 120° to each other and point towards the vertices of an equilateral triangle.

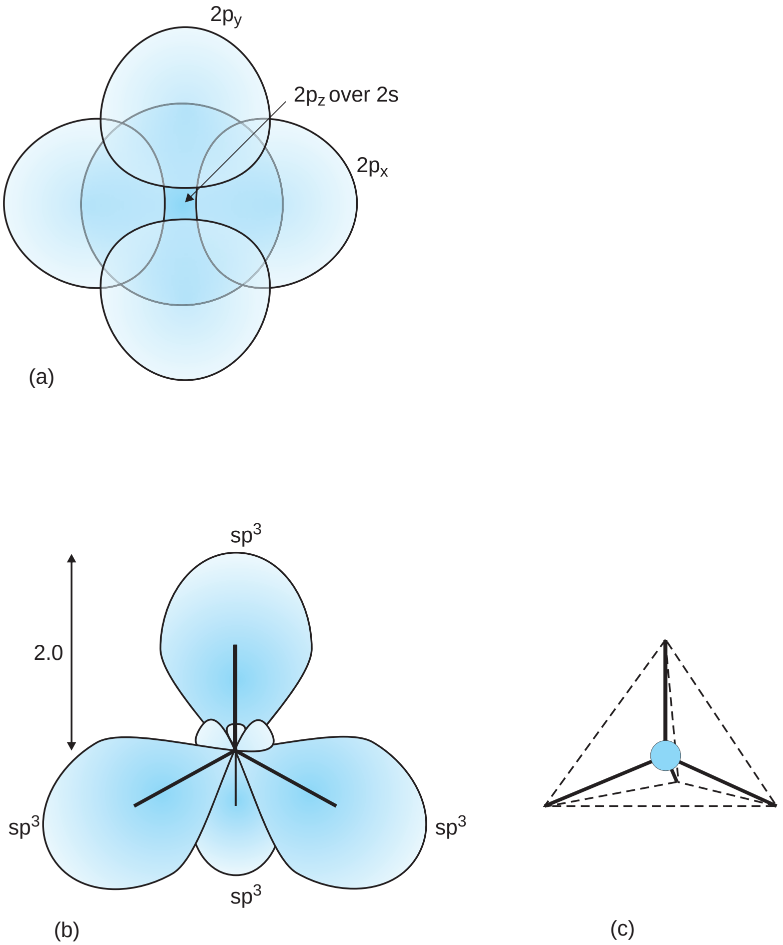


Figure 2.8 (a) The 2s, $2p_x$, $2p_y$ and $2p_z$ orbitals on an atom. (b) Four sp^3 hybrid orbitals formed by combining the four original orbitals. The orbitals are at an angle of 109.5° to each other and point towards the vertices of a tetrahedron.

hybridisation occurs. In both atoms, the s and p orbitals form sp^3 hybrids. In the case of nitrogen there are five electrons to be allocated. Three of these go into separate sp^3-hybrid orbitals that are available for bonding. The other two electrons fill the remaining orbital. This cannot now be used for bonding and is said to contain a *lone pair* of electrons. The predicted molecular geometry is tetrahedral. The lone pair of electrons add significant physical and chemical properties to the ammonia molecule.

A similar situation holds for water. There are now six electrons on the oxygen atom to allocate to the four sp^3 orbitals. In this case, two orbitals are filled and accommodate lone pairs of electrons, and two remain available for bonding. The predicted H—O—H bond angle is the tetrahedral angle, 109°. In fact the H—O—H angle is 104.5°. Qualitatively it is possible to say that the presence of the lone pairs distorts the perfect tetrahedral geometry of the hybrid orbitals. Quantitatively it indicates that the hybridisation model needs further modification. As in the case of ammonia, the lone pairs contribute significant physical and chemical properties to the molecules.

Hybridisation is no more than a convenient way of viewing the way the electron orbitals interact

Table 2.2 The geometry of some hybrid orbitals

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

*Other orbital combinations may also give the same geometry.

during chemical bonding and is not limited to just s and p orbitals. It can also occur with d and f orbitals provided that the energy requirements are covered by the resultant bond formation. The shapes of various hybrid orbitals are given in Table 2.2.

2.2.1.3 Multiple bonds

One of the characteristic features of covalent compounds is the presence of multiple bonds between atoms. In valence bond theory, single bonds, described above, 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 one σ and one π bond, whilst a triple bond consists of one σ and two π bonds. The traditional representation of these bonds as lines does not make it clear that two different bond types exist.

Triple bonding occurs in the nitrogen molecule. The outer electron configuration of nitrogen is $2s^2$ $2p^3$. As two nitrogen atoms approach each other, one pair of these *p* orbitals, say the p_x orbitals, combine in an end-on fashion, to form a σ bond. The other two p orbitals, p_y and p_z, overlap in a sideways manner to form two π bonds. The two π bonds comprise four lobes altogether, surrounding the σ bond. In the case of the oxygen molecule a similar state of affairs is found. Oxygen has an outer electron configuration of $2s^2$ $2p^4$. One p orbital will be filled with an electron pair and take no part in bonding. Only the two p orbitals, p_x and p_y, are available for bonding. Close approach of two oxygen atoms will allow the p_x orbitals to overlap end-on to form a σ bond, and the p_y orbitals to overlap in a sideways fashion to form a π bond.

Multiple bonding is of considerable importance in carbon compounds and figures prominently in the chemical and physical properties of polymers. Three compounds are illustrative: ethyne (acetylene, C_2H_2), ethene (ethylene, C_2H_4) and benzene (C_6H_6). In these, the framework of the molecule is formed by hybridisation and σ bonds while the multiple bond aspect is due to π bonds. Ethyne employs sp-hybrid orbitals that form from the 2s $2p_x$ orbitals on each carbon atom. One lobe bonds to hydrogen and the other to carbon to form the linear σ bonded H—C—C—H molecular skeleton (Figure 2.9a,b). The p_y and p_z orbitals on the carbon atoms each hold one electron and overlap sideways-on to form

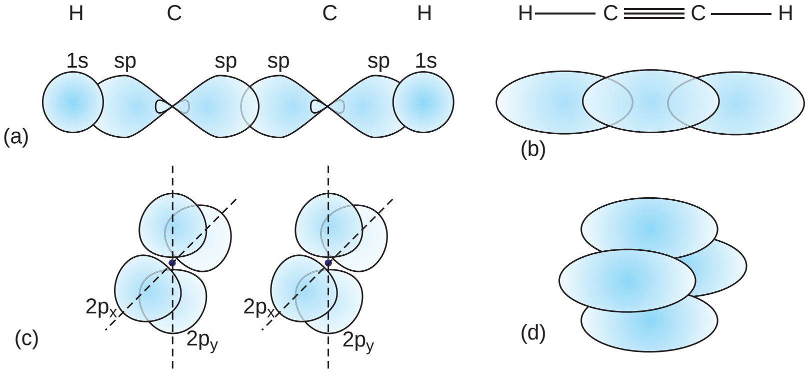


Figure 2.9 Bonding in ethyne (acetylene), C_2H_2. (a, b) Overlap of the 1s orbitals of H with the sp-hybrid orbitals on C results in a σ bonded linear molecule. (c, d) Overlap of the $2p_y$ and $2p_z$ orbitals on C results in the formation of two π bonds with lobes surrounding the C—C σ bond.

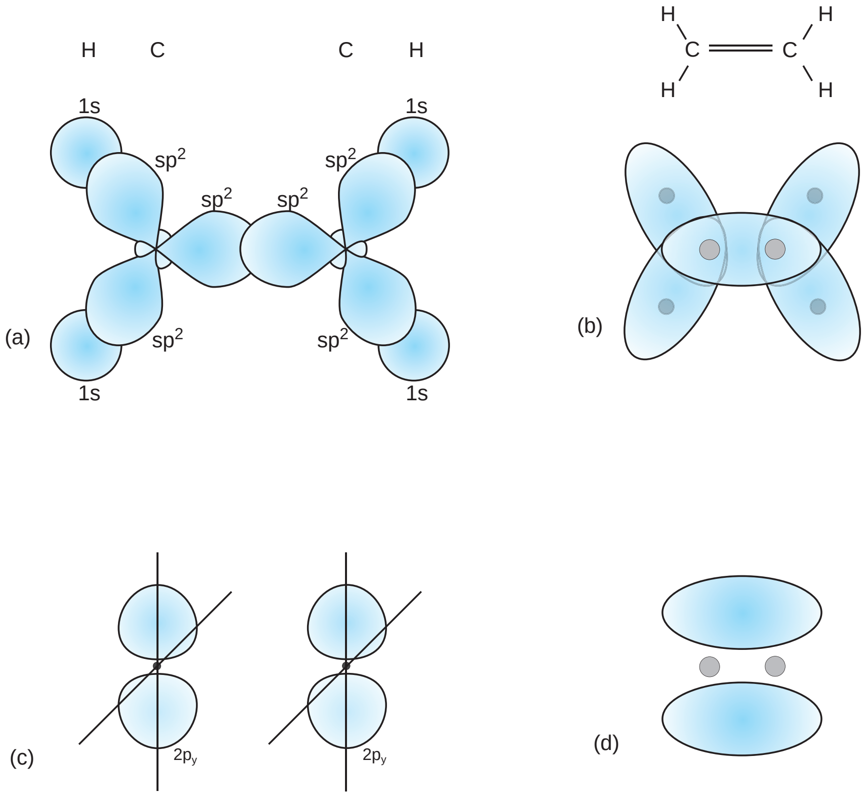


Figure 2.10 Bonding in ethene (ethylene), C_2H_4. (a, b) Overlap of the 1s orbitals of H with the sp^2-hybrid orbitals on C results in a σ-bonded molecule. (c, d) Overlap of the 2p$_y$ orbitals on C results in the formation of a π bond.

two π bonds (Figure 2.9c,d). In ethene, sp^2-hybrid orbitals form on each carbon atom, leaving one unpaired electron in the unaltered p$_z$ orbital. The three sp^2-hybrid orbitals on each carbon atom bond to two hydrogen atoms and one carbon atom in a triangular arrangement, to form the σ-bonded skeleton of the molecule (Figure 2.10a,b). The remaining p$_z$ orbitals on the two carbon atoms overlap sideways-on to form a π bond (Figure 2.10c,d).

A similar situation occurs in benzene. Each carbon atom forms sp^2 hybrids, and six carbon atoms link to each other and to six hydrogen atoms to produce a hexagonal σ bond skeleton (Figure 2.11a). The remaining p$_z$ orbitals, one on each carbon atom, overlap sideways-on to form π bonds with lobes that extend above and below the plane of the hexagon (Figure 2.11b). Although the σ bonded framework of these organic molecules is well explained, the π bonding orbitals are best treated as delocalised molecular orbitals, described below.

2.2.2 Molecular orbital theory

2.2.2.1 The energies of molecular orbitals in diatomic molecules

The molecular orbital approach to bonding parallels the approach used in building up the electron configuration of the elements (Chapter 1). As atoms come

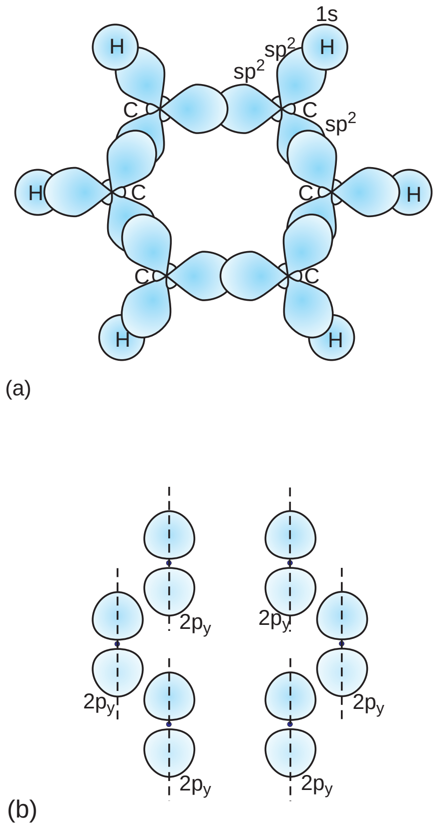


Figure 2.11 Bonding in benzene, C_6H_6. (a) Overlap of the 1s orbitals of H with the sp^2-hybrid orbitals on C results in a σ-bonded hexagonal molecule. (b) Overlap of the $2p_y$ orbitals on C results in the formation of π bonds with lobes above and below the plane of the C—H hexagon.

together to form a molecule, the atomic orbitals interact and spread over all of the nuclei to give *molecular* orbitals. The energies of these molecular orbitals are calculated and then the available electrons are fed into them, using the Pauli principle and Hund's rules (the Aufbau principle) to arrive at a molecular electron configuration. In order to be sure that a molecule will form, the total energy of the electrons in the molecular orbitals that are occupied must be less than the total energy when the electrons occupy separate atomic orbitals.

Commonly a set of molecular orbitals is obtained by adding together contributions from all of the atomic orbitals involved. This is called the *linear combination of atomic orbitals* or *LCAO* method. Thus, the wavefunction of a molecular orbital formed from two atomic orbitals, one centred on each atom of the pair forming the molecule, is, in its simplest form:

$$\psi_{\text{MO}} = c_1\phi_1 + c_2\phi_2$$

where ϕ_1 and ϕ_2 are atomic orbitals on the atoms 1 and 2, and the coefficients c_1 and c_2 give the contribution of each atomic orbital to the molecular orbital. If both atoms are identical, coefficients c_1 and c_2 would be expected to be equal, and if the atoms are different they would be expected to be dissimilar.

The calculations show that when two atomic orbitals interact, two molecular orbitals form, one with a higher energy 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. These molecular orbitals extend over the whole of the molecule. If they have cylindrical symmetry they are called σ orbitals, the 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^* (There are a number of conventions regarding the labelling of molecular orbitals; using $*$ to denote an antibonding orbital is the simplest).

The principle is easily 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, one bonding, σ1s, and one antibonding, σ^*1s (Figure 2.12). 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 $2E_{\text{bond}}$.

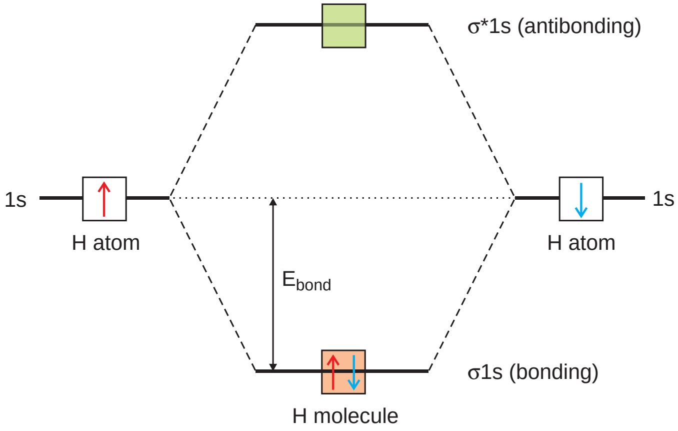


Figure 2.12 The close approach of two hydrogen atoms, each with an electron in a 1s orbital, leads to the formation of two molecular orbitals, a bonding σ1s molecular orbital and an antibonding σ^*1s orbital. In the H_2 molecule, both electrons occupy the bonding orbital, and a strong bond with energy $2E_{\text{bond}}$ results.

When two helium atoms interact there are four electrons to place 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. That is, electrons from the separate atomic orbitals are allocated to the molecular orbitals from the lowest energy upwards, remembering that the σ1s and σ^*1s orbitals are already filled and constitute an unreactive core. The interaction of the 2s outer orbitals will form σ2s and σ^*2s orbitals. In addition, the 2p orbitals can also overlap to form molecular orbitals. End-on overlap 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 2.13). The molecular configurations of the homonuclear diatomic molecules can now be obtained (Table 2.3).

An important verification of the molecular orbital approach was provided by the oxygen molecule, O_2.

Table 2.3 The electron configurations of some homonuclear diatomic molecules

Molecule	Ground state configuration	Bond length, nm	Bond energy, /kJmol^{-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 $(\sigma 1s)^2(\sigma^* 1s)^2(\sigma 2s)^2(\sigma^* 2s)^2$.

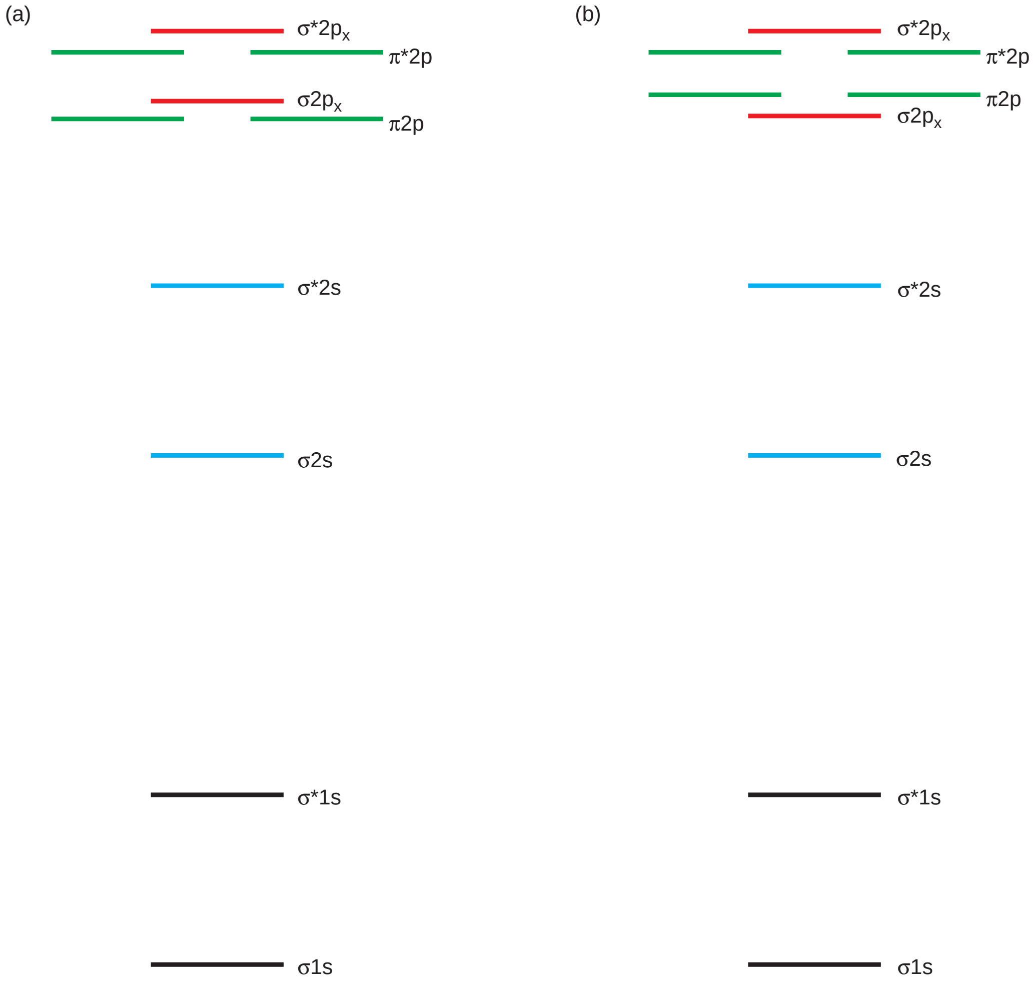


Figure 2.13 (a) Schematic molecular orbital energy level diagram for homonuclear diatomic molecules H_2 to N_2. (b) Schematic energy level diagram for the homonuclear diatomic molecules O_2 to Ne_2.

This molecule had long been known to be paramagnetic: a puzzling property. However, the electron configuration shows that the two electrons with highest energy have to be placed in separate orbitals. These unpaired electrons make the molecule paramagnetic (Section 12.1).

The molecular orbital energies of the molecules ethene, ethyne and benzene can be treated in an approximate way by using a simple strategy. The σ bonded skeleton is put to one side, and only the molecular orbital energies of the π orbitals are calculated. This is because these orbitals are of most relevance in determining the physical properties and chemical reactivity of the species. The approach is very useful when discussing more complex systems of double bonding in organic molecules and is referred to as the *Hückel method.*

2.2.2.2 Bonding between unlike atoms

When a molecular orbital, whether of σ or π type, 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. However, the molecular orbital diagram remains

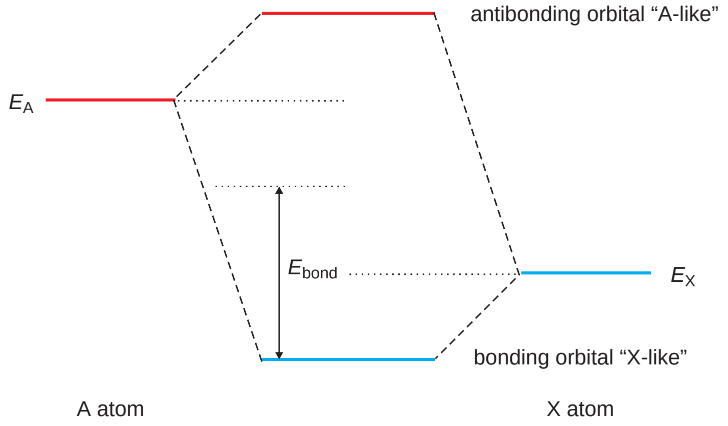


Figure 2.14 Molecular orbitals formed by a more metallic atom A and a less metallic atom X.

similar (Figure 2.14). 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 (or less electronegative, see below) 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.

A bonding molecular orbital concentrates electronic charge density in the region between the bonded nuclei (subject, in the case of π bonding, to the limitation set by the nodal plane). If the two nuclei are different, they will have different effective nuclear charges. This will cause the concentration of charge to shift to increase the screening of the higher effective charge and decrease that of lower, until both have become equalised. 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. Obviously with a very large difference in effective nuclear charge, one would have something approaching ions being formed, both electrons of the molecular orbital becoming almost completely associated with the X atom, giving it nearly unit negative charge, while the A atom would have almost unit positive charge.

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 $\delta+$, and the other end with a small negative charge, $\delta-$. The charge separation gives rise to an *internal electric dipole* and such molecules are called *polar molecules*.

A polyatomic molecule may contain a number of polar bonds. 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.

The idea of atoms possessing a tendency to attract electrons is rather useful, and the *electronegativity* of an element represents a measure of its power to attract electrons during chemical bonding. Atoms with a low electronegativity are called *electropositive* elements. These are the metals and when bonded do not have a strong tendency to attract electrons, tending to form cations. Atoms with a high electronegativity, called *electronegative* elements, tend to attract electrons in a chemical bond, and prefer to form anions. The magnitude of the partial charges ($\delta+$, $\delta-$) in a polar molecule is dependent upon the electronegativity difference between the two atoms involved.

2.2.2.3 Molecular orbital calculations

The calculation of the energies of the electrons that are part of a molecule is deceptively simple in principle but of extreme difficulty in practice. An initial simplification is to regard the nuclei as fixed, so that only the electrons need to be considered: the *Born-Oppenheimer approximation*. The task is then to solve the Schrödinger equation for the wavefunction that represents the many electrons in the molecule. An analytical solution is impossible and so efforts have been directed towards approximate solutions. The main method that has been used to successfully compute the electronic ground state of a molecule is the *Hartree-Fock method* (also see Section 1.3.4).

The first step, when working to determine the electronic properties of a molecule, is to construct the molecular orbitals using the LCAO method. The molecular orbitals described above were as simple as possible, being constructed from just one atomic orbital on each of two atoms. To obtain realistic molecular orbitals, many more atomic orbitals can be combined:

$$\psi_{\mathrm{MO}} = c_1\phi_1 + c_2\phi_2 + c_3\phi_3 + c_4\phi_4 \ldots$$

The set of atomic orbitals used is called the *basis set*. The choice of the basis set is determined in practice by the complexity of the computing problem in hand. The *minimum basis set* is that which uses the least possible number of atomic orbitals. For example, to calculate the properties of ammonia the minimal basis set would contain three 1s orbitals from the three H atoms, and the N orbitals 1s, 2s, $2p_x$, $2p_y$ and $2p_z$. Minimum basis sets rarely give results in close accord with experiment, and ideally many more orbitals are included.

Having constructed the molecular orbitals, it is then necessary to solve the Schrödinger equation for the many-electron molecular orbital. As with atoms, the molecular orbital, which depends upon the positions and interactions of all of the electrons, $\psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \ldots)$, is replaced with the orbital approximation:

$$\psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3 \ldots) = \psi(\mathbf{r}_1)\psi(\mathbf{r}_2)\psi(\mathbf{r}_3) \ldots$$

There are several factors to take into account when writing this in a form that can be computed. Firstly, electrons are indistinguishable, and secondly, they have a spin, ↑ or ↓. To take this into account the wavefunction of the system must change sign when any two electrons are exchanged. This *exchange symmetry* leads to a lowering of energy of the system by the *exchange energy*. A second factor is electron–electron repulsion. For example, with just two electrons, the motion of one depends upon the repulsion from the other, which in turn depends upon the position and motion of the other. That is to say, motion of one electron is not independent of the motion of the other – they are *correlated*. This correlation, which is vastly more complicated in a many-electron orbital, also leads to a reduction in energy, called the *correlation energy*.

The Hartree-Fock self-consistent field method is now used to construct the best molecular orbitals for the entity. One electron is selected and its molecular orbital is computed by allocating an estimated form to all of the other functions $\psi(\mathbf{r}_{n-1})$ and allowing the electron to experience a potential which is the average electron density of all of the other electrons and nuclei. (Note that the use of an averaged potential means that electron correlation energy is *omitted*.) This new orbital $\psi(\mathbf{r}_1)$ is now used and the procedure repeated for all other electron orbitals until no further change is detected between cycles. This is the self-consistent field limit and the end molecular orbital is the Hartree-Fock self-consistent field (HF-SCF) orbital. Each iteration requires the evaluation of complex integrals, and the success of computer routines to reach a meaningful end result is largely due to clever mathematical operations that make the task feasible within a reasonably short computation time. Computer software is now readily available for the calculation of orbital energies and molecular geometry via the Hartree-Fock method. Much use is made of this method for the estimation of the conformation of molecules of potential pharmaceutical use, as molecular shape is one of the most important factors governing effective drug action.

2.3 Metallic bonding and energy bands

Metallic bonds must possess the following characteristics. The bonds 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 bonds must permit easy electron transfer throughout the structure.

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 and are not really associated with any particular atom. Formally, this is called *free-electron theory*. A second approach, more chemical in nature, considers metallic bonds as a delocalisation of covalent bonds throughout the solid. Formally, this is called the *tight-binding theory*. To understand metallic materials in the broadest sense, a combination of both approaches must be used. Moreover, accurate calculations use more sophisticated models than either the free-electron or tight-binding approaches. The application of these ideas, called *band theory*, successfully explains the detailed physical properties of metals.

2.3.1 Molecular orbitals and energy bands

The molecular orbital model of chemical bonding suggests that when two atoms approach, the outer atomic orbitals on each atom interact to form molecular orbitals. Each pair of atomic orbitals generates two molecular orbitals, one of lower energy, the bonding orbital, and one of higher energy, the antibonding orbital (Figure 2.15). Suppose 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 highest energy (antibonding) and lowest 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 2.15). From the point of view of properties, the highest energy (outermost) band is the most important. 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*.

Although the 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

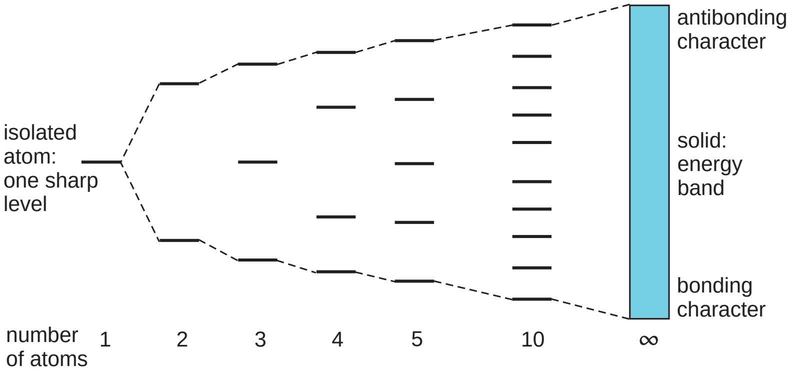


Figure 2.15 The schematic development of an energy band from delocalised molecular orbitals. Each atom in the molecule contributes one molecular orbital. At the left, an isolated atom has sharp energy levels. As the number of atoms increases, the number of discrete orbitals merges into a band of effectively continuous energy levels.

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 of a free atom. On the other hand, the 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 2.16). 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.

2.3.2 The free electron gas

The free-electron theory of metals started with a gross approximation: the idea that a metal contained a 'gas' of electrons, uninfluenced by other electrons, or anything else. The model, an adaptation of the kinetic theory of gases, was very successful. The electrons were supposed to be flying about in the metal, in every direction, at random. 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 the metal should increase slightly with temperature, as it does (Chapter 12).

However, it could not explain some important properties. In particular, the theory was unable to

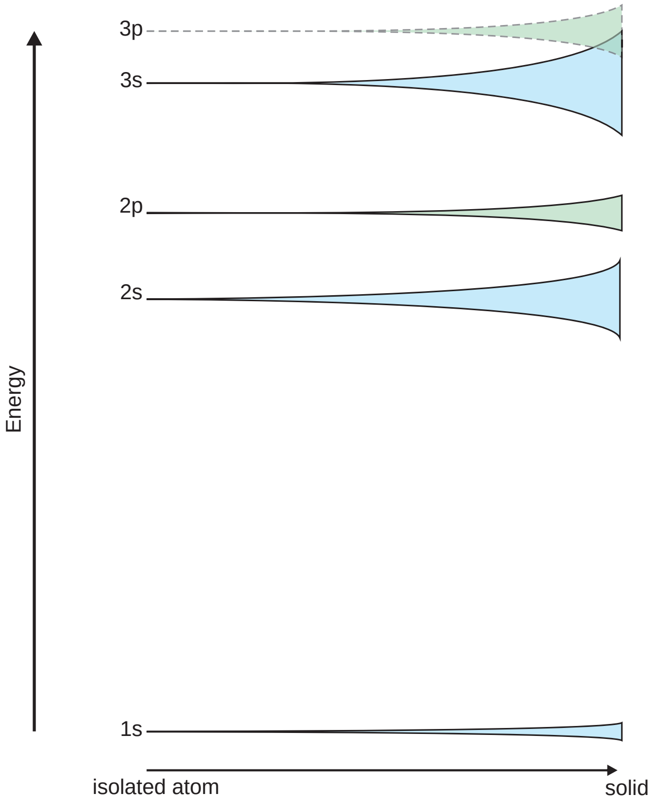


Figure 2.16 The schematic development of energy bands from atomic orbitals for magnesium metal as the atoms approach each other. The outermost filled 3s and empty 3p bands overlap in the metal.

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 4.5 R, while insulators would have a molar specific heat of only 3 R, where R is the gas constant (Section 15.1).

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 2.17). The allowed wavelengths, λ, are given by:

$$\lambda = \frac{2a}{n}$$

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

$$\psi = \sqrt{2/a}\ \sin kx$$

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

$$k = \frac{2\pi}{\lambda} = \frac{n\pi}{a} \tag{2.3}$$

(In three dimensions, **k** is a vector quantity, called the *wave vector*.)

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

$$E_\mathrm{n} = \frac{k^2h^2}{8\pi^2 m}$$

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 versus k graph is parabolic (Figure 2.18).

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:

$$E(n_x, n_y, n_z) = \frac{h^2}{8m}\left(\frac{n_x^2}{a^2} + \frac{n_y^2}{b^2} + \frac{n_z^2}{c^2}\right)$$

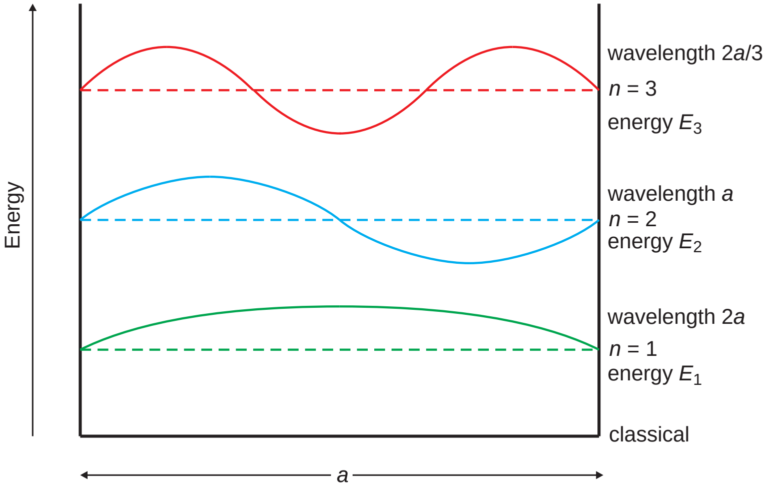


Figure 2.17 The energy levels and allowed wavelengths of an electron wave confined to a line of length a.

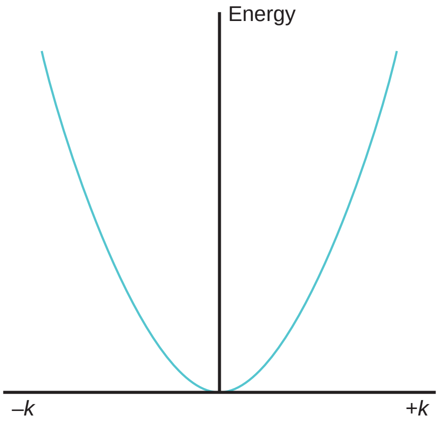


Figure 2.18 The relationship between the energy of an electron, E, confined to a line, and the wave number, k. The curve is a parabola as $E \propto k^2$.

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 length a, this reduces to:

$$E(n_x, n_y, n_z) = \frac{h^2}{8ma^2}\left(n_x^2 + n_y^2 + n_z^2\right)$$

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

$$N(E) = 2\pi\, V\left(\frac{8m}{h^2}\right)^{3/2} E^{1/2}$$

This equation is called the *density of states function*. This curve, like the E versus 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. The energy levels calculated above, 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 2.19). In three dimensions

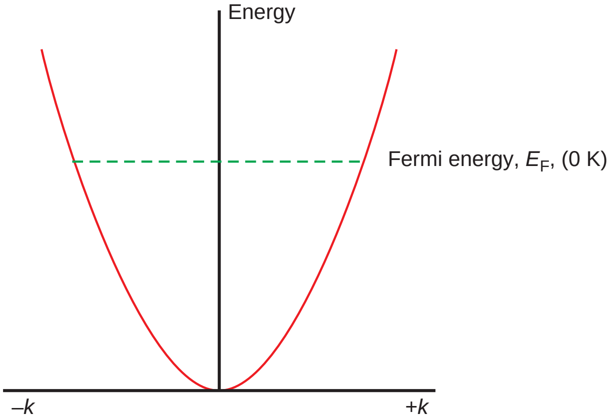


Figure 2.19 The Fermi energy, E_F, is the uppermost energy level filled at 0 K. At higher temperatures, the electrons are distributed over nearby energy levels and the boundary becomes less sharp.

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 in 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, and a major drawback of the classical theory has been corrected.

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 that the probability, P_i, that an energy level, E_i, will be occupied is given by:

$$P_i = \frac{1}{\exp\left(\frac{E_i - \mu}{k_B T}\right) + 1}$$

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:

$$P_i = \frac{1}{\exp\left(\frac{E_i - E_F}{k_B T}\right) + 1}$$

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

$$P(\mathrm{E}) = \frac{1}{\exp\left(\frac{E - E_F}{k_B T}\right) + 1}$$

where P(E) is the probability than an energy level E is occupied by an electron at absolute temperature T (Figure 2.20). At $T = 0$ K, $P(\mathrm{E}) = 1$ for $E < E_F$ and 0 for $E > E_F$. We also note from the equation that at $E = E_F$, $P(\mathrm{E}) = {}^1\!/_2$. Thus the Fermi level in metals above 0 K can be defined as the energy for which $P(\mathrm{E}) = {}^1\!/_2$.

2.3.3 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

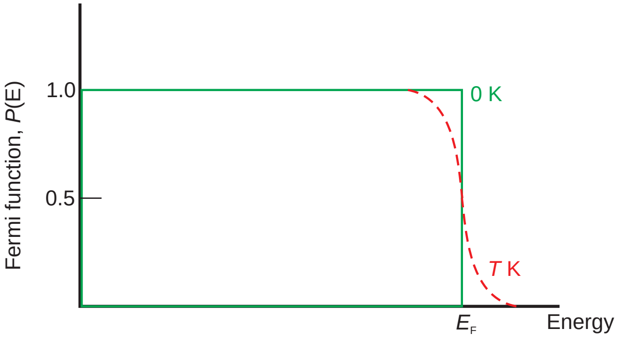


Figure 2.20 The Fermi function, P(E), giving the probability of the occupation of an energy level as a function of the energy, E. At 0 K, the probability is 1.0 up to the Fermi energy, E_F, and thereafter 0. At higher temperatures, the curve changes smoothly from 1.0 to 0.

free-electron waves multiplied by a function with the same periodicity as the crystal structure. These wave equations are called *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 above. The electron is 'not aware' of the atoms in the solid.

As the wavelength approaches the same dimensions as the atomic 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* (sections 5.2 and 14.7). Although diffraction is a complicated process, in the present situation, the conditions for diffraction to occur are given by Bragg's law (Section 5.2). For an electron wave normally incident upon planes of atoms, diffraction of the wave occurs at values of the wave number:

$$k = \pm\frac{n\pi}{a}$$

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 wave numbers will not be able to pass through the crystal and the one-dimensional E versus k curve is broken at these values (Figure 2.21a). Propagation will occur again when the wave number 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 2.21b).

The discontinuities in the energy versus k curve due to electron diffraction mark the boundaries of *Brillouin zones*. The first Brillouin zone in the one-dimensional case (Figure 2.21) 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 -2π/a on the negative side. Further Brillouin zones can be similarly located. The concept of a Brillouin zone is an abstract concept, as the zones exist in a space defined by the wave vector and the energy of the electron. The wave vector is proportional to the velocity and the momentum of the electron, and the zones are sometimes described as existing in *velocity* or *momentum space*.

2.3.4 Properties of metals

The models described account for the cohesive energy of a metal. The upper energy levels of a band are antibonding in character, whilst the lower energy levels are bonding in character. The

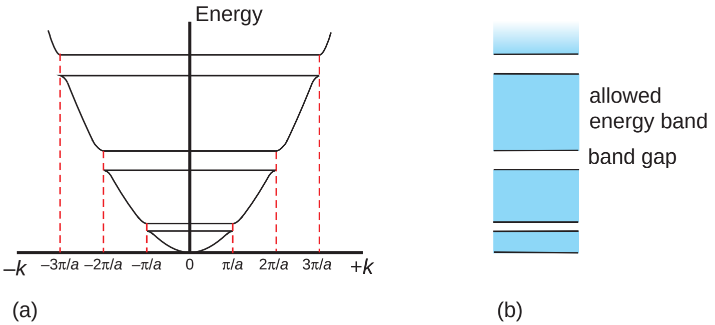


Figure 2.21 (a) The one-dimensional free-electron E versus k curve broken up into energy bands, due to the periodic potential of atomic nuclei of separation a. (b) Schematic representation of energy bands.

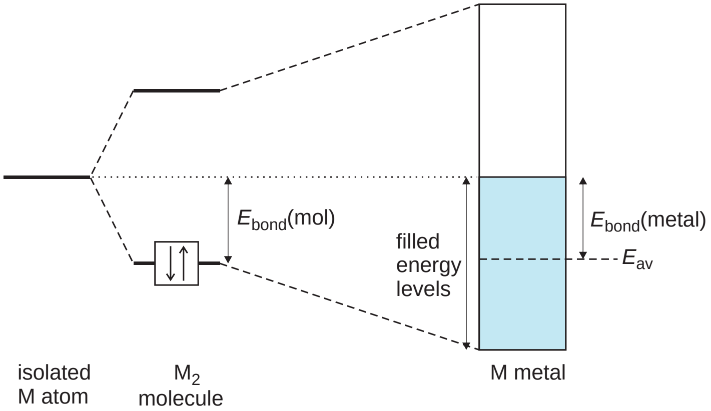


Figure 2.22 The bonding energy in a molecule, M_2, compared with an isolated M atom, is due to filling the bonding molecular orbital while the antibonding orbital remains empty. In a solid metal, bonding energy is due to filling the lower bonding character energy levels, while leaving the upper antibonding character energy levels empty.

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 will 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 (Figure 2.22). Moreover, as the electrons are delocalised, metallic bonds are expected to be *non-directional*.

This simple picture must be treated with care. For example, magnesium has an electron configuration [Ne] $3s^2$. A simple 3s band would be filled, suggesting that Mg metal would be unstable. This is not so because the 3p orbitals, although empty on an isolated atom, broaden into a band that overlaps the 3s band (Figure 2.16). The combined s and p band can hold eight electrons per atom. As each Mg atom contributes two, these will mainly occupy bonding levels and a solid will be stable.

The same proviso, that the uppermost band is only partly occupied, 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 of nearest neighbours (8 to 12), 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 with 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.

2.3.5 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 in 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 they 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 2.23). Thus the ionic model is still a good model for the bonding in NaCl.

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 2.24). 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 closer, 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

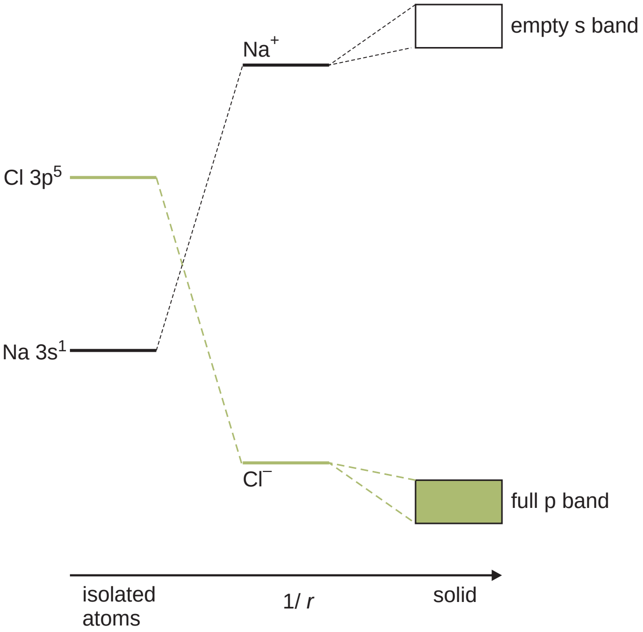


Figure 2.23 Schematic development of energy bands from isolated atoms in sodium chloride, NaCl. As the spacing between the atoms decreases, firstly ionisation occurs, and then energy bands develop. (For clarity, the core energy levels of both atoms have been ignored and only relative changes in energy are included.)

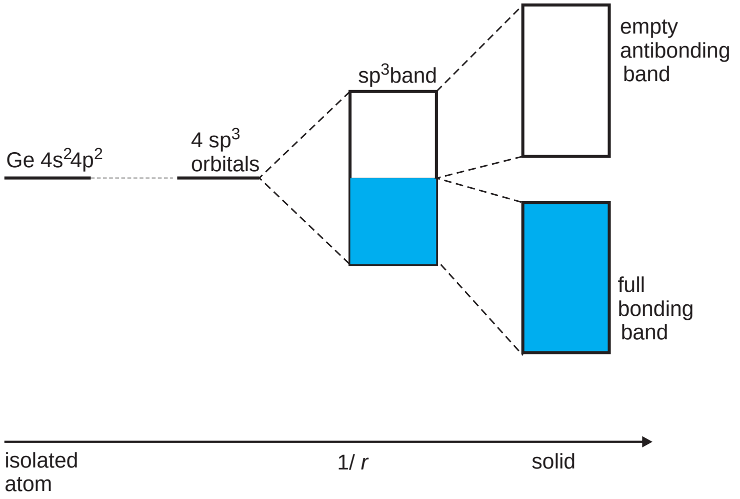


Figure 2.24 Schematic development of energy bands from isolated atoms in germanium, Ge. As the spacing between the atoms decreases, firstly orbital hybridisation occurs, and then energy bands develop. (For clarity, the core energy levels of the atoms have been ignored and only relative changes in energy are included.)

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.

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, 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, bands widen, the energy gap narrows, and the cohesive energy falls. Unlike diamond, α-Sn is a weak solid, easily crushed.

While this description is quite different from 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.

2.3.6 Computation of properties

The calculation of material properties using the HF-SCF method does not take into account the important energy lowering due to electron correlation (Section 2.2.2.3). This shortcoming is 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. The LCAO method can be used if appropriate, but once again, for many materials a simpler approach is satisfactory, by using plane waves similar to those derived as solutions to the wave equation for a free electron gas (Section 2.3.2). 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 *pseudo-potential*, which reduces the number of plane waves needed, the plane wave basis set, to manageable proportions.

Further reading

The following references greatly expand the material in this chapter:

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

Atkins, P.W. (1991) *Quanta*, 2nd edn. Oxford University Press, Oxford.

Cottrell, A. (1988) *Introduction to the Modern Theory of Metals*. Institute of Metals, London.

Harrison, W.A. (1980) *Electronic Structure and the Properties of Solids*. W.H. Freeman, San Francisco.

Pauling, L. (1960) *The Nature of the Chemical Bond*, 3rd edn. Cornell University Press. (The classic textbook mostly concerned with valence bond theory.)

Pearson, W.B. (1972) *The Crystal Chemistry and Physics of Metals and Alloys*. Wiley-Interscience, New York.

Ionic radii are discussed and tabulated by:

Shannon, R.D. and Prewitt, C.T. (1969) Acta Crystallographica, **B25**: 925–46.

Shannon, R.D. and Prewitt, C.T. (1970) Acta Crystallographica, **B26**: 1046.

Shannon, R.D. (1976) Acta Crystallographica, **A32**: 751–6.

Computations are described in:

Catlow, C.R.A. (2006) Computer modelling of solids, in *Encyclopedia of Inorganic Chemistry*, 2nd edn (ed. R. B. King). John Wiley and Sons, Ltd., Hoboken, NJ.

Cramer, C.J. (2004) *Essentials of Computational Chemistry*, 2nd edn. John Wiley and Sons, Ltd., Chichester.

Gillan, M.J. (1997) The virtual matter laboratory. *Contemp. Phys.*, **38**: 115–30.

Jensen, F. (2007) *Introduction to Computational Chemistry*, 2nd edn. John Wiley and Sons, Ltd., Chichester.

Lewars, E.G. (2011) *Computational Chemistry*, 2nd edn. Springer, Heidelberg.

Lindorff-Larsen, K., Piana, S., Dror, R.O. and Shaw, D.E. (2011) *Science*, **334**: 517–20.

Various authors, *Materials Research Society Bulletin*, **37**, May 2012, is devoted to the simulation of materials properties using many-body potentials.

There are many academic and commercial software routines that are available for molecular and solid state modelling. Lists of these can be accessed via Wikipedia: *List of quantum chemistry and solid-state physics software* (accessed 2012).

For a widely used program for atomistic simulation (GULP) see:

Gale, J.D. (1997) *J. Chem. Soc., Faraday Trans.*, **93**: 629–37, and http://gulp.curtin.edu.au.

Problems and exercises

Quick quiz

1 The concept of the *valence* of an atom refers to:
(a) The charge on an ion.
(b) The strength of the chemical bonds formed.
(c) The combining ability of an atom.

2 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.

3 Which of the following is a trivalent ion:
(a) V^{3+}.
(b) V^{4+}.
(c) V^{5+}.

4 Ionic bonds are formed by which process:
(a) Electron sharing.
(b) Electron transfer.
(c) Electron delocalisation.

5 A key feature of an ionic bond is that it is:
(a) Strongly directional.
(b) Completely non-directional.
(c) Acts between identical atoms.

6 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.

7 The lattice energy of an ionic solid is:
(a) The sum of electrostatic and repulsive energies.
(b) The minimum of the electrostatic and repulsive energies.
(c) The difference between the electrostatic and repulsive energies.

8 Cations are generally:
(a) Smaller than anions.
(b) Bigger than anions.
(c) The same size as anions.

9 A cation gets:
(a) Larger as the charge on it increases.
(b) Smaller as the charge on it increases.
(c) Remains the same size whatever the charge.

10 Covalent bonds are formed by which process:
(a) Electron sharing.
(b) Electron transfer.
(c) Electron delocalisation.

11 Covalent bonds are:
(a) Strongly directional.
(b) Completely non-directional.
(c) Variable in direction.

12 Covalent bonds are also called:
(a) Molecular orbitals.
(b) Hybrid orbitals.
(c) Electron-pair bonds.

13 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.

14 A π bond is characterised by:
(a) Being formed by two p orbitals.
(b) Reflection symmetry about the bond axis.
(c) Being formed between different atoms.

15 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.

16 A polar covalent bond:
(a) Forms between different size atoms.
(b) Involves ions linked with molecular orbitals.
(c) Has the electron pair unevenly distributed in the molecular orbital.

17 The electronegativity of an atom is:
(a) A measure of its tendency to attract electrons.
(b) A measure its tendency to repel electrons.
(c) A measure its tendency to form a covalent bond.

18 Strongest covalent bonds form:
(a) When the orbitals are of the same type.
(b) When the orbitals overlap maximally.
(c) When the orbitals are symmetrically arranged.

19 A hybrid orbital is:
(a) An overlapping pair of orbitals.
(b) A combination of orbitals on adjacent atoms.
(c) A combination of orbitals on a single atom.

20 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.

21 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.

22 Metallic bonds are formed by which process:
(a) Electron sharing.
(b) Electron transfer.
(c) Electron delocalisation.

23 The cohesive energy of a metal is due to:
(a) A partly filled energy band.
(b) An empty energy band.
(c) An overlap of energy bands.

24 A metallic bond is predominantly:
(a) Strongly directional.
(b) Completely non-directional.
(c) Partly directional.

25 Wave functions are said to be degenerate if:
(a) They have different energies.
(b) They overlap.
(c) They have the same energy.

26 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.

27 Energy gaps in crystals can be thought of as due to:
(a) The diffraction of electron waves.
(b) The bonding of electrons.
(c) The localisation of electron waves.

28 Energy bands can form:
(a) Only in crystals.
(b) Only in metals.
(c) Only in electrical conductors.

Calculations and questions

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

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

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

2.4 Write the symbols of the ions formed by iron, chlorine, aluminium, sulphur, lanthanum, and tantalum.

2.5 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.

2.6 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.

2.7 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.

2.8 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 0.19697 kg $\mathrm{mol^{-1}}$, and the density is 19281 kg $\mathrm{m^{-3}}$.

2.9 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 0.05869 kg $\mathrm{mol^{-1}}$, and the density is 8907 kg $\mathrm{m^{-3}}$.

2.10 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 0.06355 kg $\mathrm{mol^{-1}}$, and the density is 8933 kg $\mathrm{m^{-3}}$.

2.11 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 0.02431 kg $\mathrm{mol^{-1}}$, and the density is 1738 kg $\mathrm{m^{-3}}$.

2.12 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 0.05585 kg $\mathrm{mol^{-1}}$, and the density is 7873 kg $\mathrm{m^{-3}}$.

2.13 Calculate the lowest energy level of a free electron in a cube of metal of 1 cm edge length, 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.

2.14 Estimate the Fermi energy of silver. The molar mass of silver is 0.1079 kg $\mathrm{mol^{-1}}$, the density is 10500 kg $\mathrm{m^{-3}}$, and each silver atom contributes one electron to the ‘electron gas’.

2.15 Estimate the Fermi energy of sodium. The molar mass of sodium is 0.02299 kg $\mathrm{mol^{-1}}$, the density is 966 kg $\mathrm{m^{-3}}$, and each sodium atom contributes one electron to the ‘electron gas’.

2.16 Estimate the Fermi energy of calcium. The molar mass of calcium is 0.04408 kg $\mathrm{mol^{-1}}$, the density is 1530 kg $\mathrm{m^{-3}}$, and each calcium atom contributes two electrons to the ‘electron gas’.

2.17 Estimate the Fermi energy of aluminium. The molar mass of aluminium is 0.02698 kg $\mathrm{mol^{-1}}$, the density is 2698 kg $\mathrm{m^{-3}}$, and each aluminium atom contributes three electrons to the ‘electron gas’.

2.18 Draw the wave functions 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.

2.19 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.

3

States of aggregation

- What is the difference between a crystal and a quasicrystal?
- What are point defects?
- What is the microstructure of a solid?

3.1 Weak chemical bonds

Aggregation is not solely due to the strong chemical bonds described in the previous chapter. Even noble gas atoms experience weak interatomic forces that lead to liquefaction and, except for helium, solidification at low temperatures. Although these interactions are weak in terms of bond energy (Table 3.1), they are of vital importance, especially in living organisms.

The strongest of the weak bonds involves dipoles. *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, $\delta+$) associated with the hydrogen atom, and a region of increased electron density (augmented negative charge, $\delta-$) 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 a $\delta+$ region residing on each hydrogen and a $\delta-$ region on the oxygen. These add together so that the resultant dipole 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 *van der*

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

Table 3.1 Forces between atoms, ions and molecules

Type of bond	Approximate energy/kJ mol^{-1}	Species involved
Covalent	350	Atoms with partly filled orbitals
Ionic	250	Ions only
Metal	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

Waal's 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 3.1). 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).

These weak interactions can be represented by potential energy curves similar to those described in the previous chapter as acting between atoms. 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 term. 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:

$$V_{LJ} = 4V_m\left[\left(\frac{r(0)}{r}\right)^6 - \left(\frac{r(0)}{r}\right)^{12}\right]$$

where $r(0)$ is the value of r when V_{LJ} is zero (Figure 3.2).

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.

A *hydrogen bond* is a weak bond formed when a hydrogen atom lies between two highly electronegative atoms, such as fluorine, oxygen, chlorine or nitrogen. The bond results from the interaction of the small positive charge, $\delta+$, found in dipolar molecules containing hydrogen, with the partial charge of $\delta-$ 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, hydrogen fluoride, and potassium hydrogen fluoride KHF_2. The existence of hydrogen bonds dramatically changes many of the properties of the material.

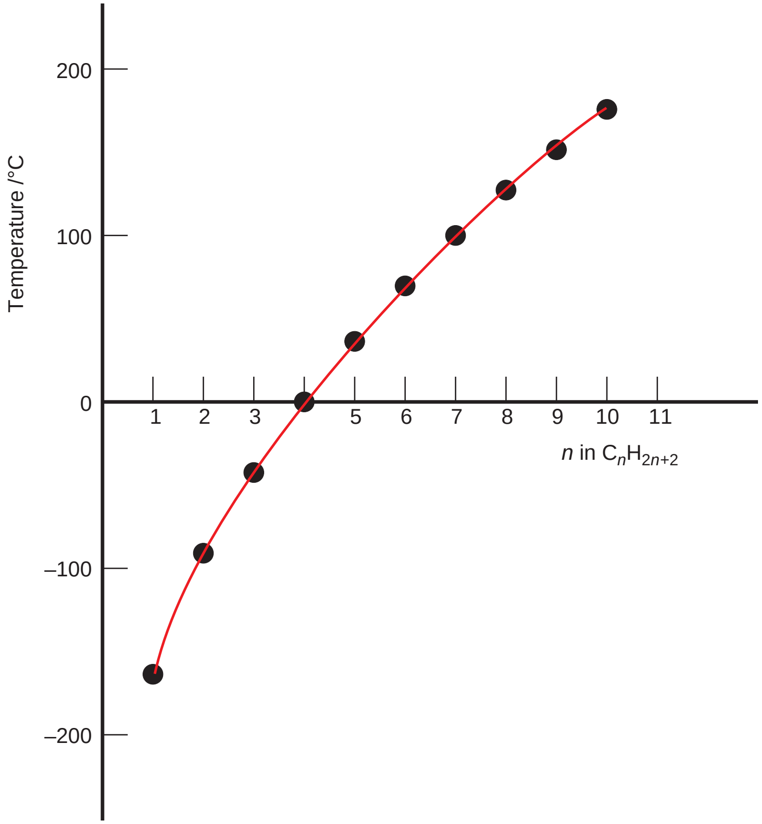


Figure 3.1 The boiling points of the alkanes (saturated hydrocarbons) of formula C_nH_{2n+2}, plotted against the value of n, which is proportional to the size of the molecule.

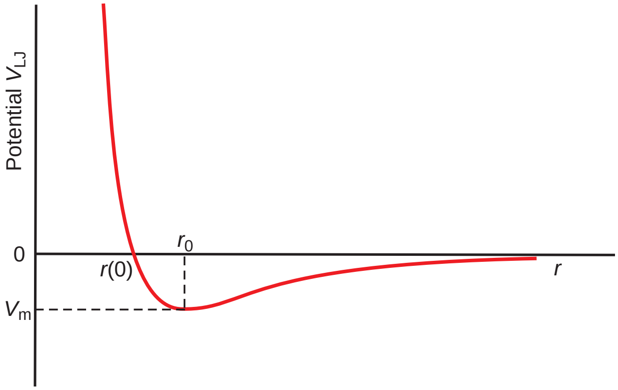


Figure 3.2 The Lennard-Jones potential between two atoms separated by a distance r.

For example, HF, H_2O and 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. This 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 11.3.5).

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 interionic 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 $1/r^6$.

3.2 Macrostructures, microstructures and nanostructures

3.2.1 Structures and scale

The shape and scale of an object reflect its function (Table 3.2). The gross shape of a container is different to the shape of a blade, and the purposes of the two objects are readily discriminated by eye. The attributes of an object that fit it to its functional use are at a smaller scale, often called the *macrostructure*. For example, a container may be glazed or porous, something still readily detected by eye.

Many of the properties of materials, though, are dominated by structures at a submillimetre scale: the *microstructure* of the material. As the name implies, the microstructure of an object is observed using optical microscopy. For example, good-quality ceramics have a microstructure that is a mixture of crystallites in glass. At the lower end of the microstructure scale, defects in materials play an important role. Electron microscopy replaces optical microscopy as the required tool here.

Even smaller dimensions, which characterise structure closer to an atomic scale, are referred to as the *nanostructure* of the object, and small particles consisting of few atoms in extent are often referred to as *nanoparticles*. For this degree of detail, high-resolution electron microscopy or allied techniques are needed.

At a more fundamental level again, the *crystal structure* of a material provides detail of the chemical atoms present, their positions, and often any disorder existing. The techniques of X-ray, neutron and electron diffraction are required to obtain this information, supplemented by a number of spectroscopic techniques and, at surfaces, atomic force microscopy.

3.2.2 Crystalline solids

Crystals are solids in which all of the atoms occupy well-defined locations, ordered across the whole of the material. They are defined in terms of a unit cell, a brick-like shape that contains all of the

Table 3.2 Scales of structure

Description	Dimensions	Example	Analytical tools
Gross shape	10^{-2} m upwards	Ceramic vessel	Visual examination
Macrostructure	$\sim 10^{-2}$ m	Surface glaze, underlying material	Eye, magnifying glass
Microstructure	10^{-4}–10^{-6} m	Crystallites + non-crystalline material	Microscopy (optical, electron)
Nanostructure	10^{-8} m	Carbon nanotube	High-resolution electron microscopy
Crystal structure	10^{-9} m or less	Atomic positions in crystals	Diffraction (X-ray, neutron, electron)

atomic species in the crystal, and which, when stacked in three-dimensions, reproduces the macroscopic crystal (Chapter 5). In many crystals the atoms are held in place by strong ionic, covalent or metallic bonds that extend throughout the matrix. Considering that chemical bonds tend to operate over only a few interatomic distances, it is rather surprising that so much of the solid state is crystalline. Nevertheless, this is so, and it is only with difficulty that many ordinary solids can be prepared in a non-crystalline form. 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 together by the weaker bonds described above. Such materials tend to have low melting points and to be rather frail. Single crystals are used for fundamental investigations of solid properties and are vital in many engineering applications.

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*.

3.2.3 Quasicrystals

In crystallography, because of space-filling requirements, a unit cell with overall five-fold, or greater than six-fold, rotational symmetry cannot exist, because none can be packed to form a crystal. In 1984, a metallic alloy of composition approximately $Al_{88}Mn_{12}$ was reported which possessed a ten-fold rotational symmetry axis and long-range translational order – something quite incompatible with classical crystallography. Initially effort was put into trying to explain the rotational symmetry as an artefact due to the presence of defects, but it was finally proven that the material really did have ten-fold rotational symmetry and the material violated the conventional laws of crystallography. Since then, many other alloys that show five-fold, eight-fold, ten-fold and 12-fold rotational symmetry have been discovered. The resulting materials rapidly became known as *quasiperiodic crystals*, or more compactly, as *quasicrystals*.

There are a number of ways that the contradiction between classical crystallography and the structures of quasicrystals can be resolved, all of which require a slight relaxation in the strict rules of conventional crystallography. A quasicrystal can be regarded as made up of certain structural units, all oriented in the same way, and separated by variable amounts of disordered material; that is, the materials show orientational order but *not* translational order.

The simplest way to explain this is in terms of *Penrose tilings*. Penrose tilings are two-dimensional patterns that are *aperiodic*: they cannot be described as having unit cells and do not show translational order. However, a Penrose tiling has a sort of translational order, in the sense that parts of the pattern are oriented identically, but they are not spaced in such a way as to generate a unit cell (Figure 3.3). Moreover, if the nodes in a Penrose tiling are replaced by atoms, computed diffraction patterns

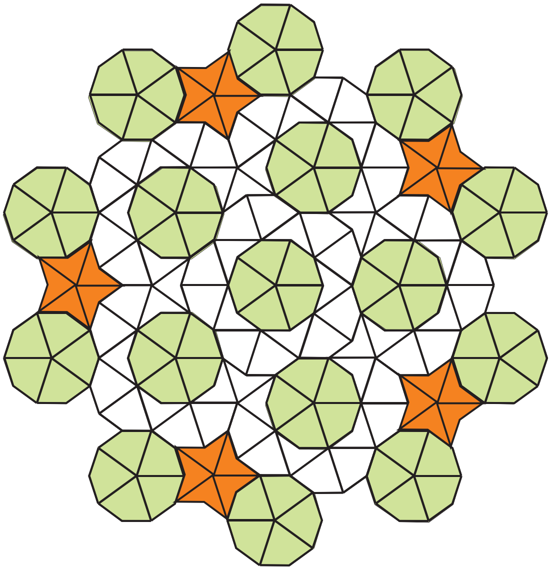


Figure 3.3 A Penrose tiling in which a number of polyhedra are oriented in the same way, but the pattern does not have a unit cell. Such tilings are two-dimensional equivalents of quasicrystals.

show sharp spots and five- and ten-fold rotational symmetries identical to those obtained experimentally from quasicrystals.

The Penrose model of quasicrystals consists of a three-dimensional Penrose tiling built of icosahedra. These can be joined to give an aperiodic structure which has the same sort of order as the two-dimensional tiling. That is, all of the icosahedra are in the same orientation, but do not show translational order. As with the planar tilings, these three-dimensional analogues give diffraction patterns that show both sharp spots and forbidden rotational symmetries. Icosahedral clusters of metal atoms are common (albeit transient) component entities of liquid metals, as these are a very compact way of packing spherical metal atoms. In favourable circumstances they are retained upon cooling to give a solid with the correct distribution of icosahedra to form a quasicrystal. It is not surprising, therefore, that quasicrystals were first found in alloys with complex composition, as these are often inhibited from crystallizing because of the degree of ordering required, but are able to solidify with the partial ordering associated with quasicrystals.

3.2.4 Non-crystalline solids

Non-crystalline solids do not have long-range order of the atoms in the structure. There is usually some short-range order, extending over a few atomic radii, but no correlation of atomic positions at longer distances. There are four important types of non-crystalline solid: glasses, polymers, amorphous solids and aerogels.

A *glass* is formed when a high-temperature liquid state is *undercooled* or *supercooled* to below the solidification point of a potential crystalline solid, and then solidifies without crystallizing. Glasses are often described as supercooled liquids. There is no one structure of glass any more than there is one structure of a crystal, and almost any solid can be produced in a glassy state if the melt is cooled sufficiently quickly. To some extent, glass can be thought of as a product of kinetics, and the structure of a glass can depend upon the rate at which the liquid is cooled. Theories of glass structure and formation must consider this. This status of glass is revealed by the behaviour upon warming. 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 which can be defined as a viscous liquid.

The best-known glasses, manufactured from silicon dioxide mixed with other oxides, are called *silicate glasses* (Section 6.3). However, both metals and organic compounds can also 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. In the case of most metals, a glassy state is much harder to achieve, and skill is required in formulating alloy compositions that reject crystallization on cooling (Section 6.1.4).

Glasses containing several components are often found to be inhomogeneous at a scale of about 10^{-6} m. Compositional 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 upon cooling, when some components separate by a process called *spinodal decomposition*.

Polymers are a class of substances that consist of very large molecules, *macromolecules*, built up from many multiples of small molecules or *monomers*. They can be synthetic (polythene, nylon) or natural (protein, rubber), and occur widely in nature as vital components of living organisms. 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. The long molecules themselves are linked by weak bonds and are usually disordered. 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.

Solids that are evaporated and then condensed 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 $\mathrm{m\,s^{-1}}$ compared with 5000 $\mathrm{m\,s^{-1}}$ in silica glass, with all values dependent upon the porosity. These materials find uses ranging from thermal insulation to high-energy nuclear particle detectors.

3.2.5 Partly crystalline solids

Although most solids turn out to be 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 or lesser extent by the structure of the polymer molecules. Most polymers have a chain-like form and longer chains are more difficult to crystallise. The presence or absence of side groups attached to the chain also has a considerable effect on the ease with which a polymer chain can crystallise. The partly crystalline structure of many linear polymers is typified by one of the simplest of polymers, polythene. Polythene molecules are 10^4 monomer units or more long, and resemble thin strings. If the liquid is cooled reasonably quickly, the chains remain in an extended form. The material has a low density, 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.

3.2.6 Nanoparticles and nanostructures

Nanoparticles and nanostructures are generally small enough that chemical and physical properties are observably different from the normal properties of bulk solids. For example, the energy levels of isolated atoms are sharp, while atoms in a solid 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. When this dimension is reached, a nanoparticle can be said to have formed.

The dimension at which non-bulk properties become apparent depends upon the phenomenon investigated. In the case of thermal effects, the boundary occurs at approximately the value of thermal energy, k_BT, which is about 4×10^{-21} J. In the case of optical effects, non-classical behaviour, that is, diffraction, is noted when the scale of the object illuminated is of the same size as a light wave, say about 5×10^{-7} m. For particles such as electrons, the scale is determined by the Heisenberg Uncertainty Principle, at about 3×10^{-8} m.

The bulk–nanostructure transition has been most studied in electronics, optoelectronics, and the related topics of magnetism and ferroelectricity. A thin film will have bulk properties modified towards atom-like properties in a direction normal to the layers. A thin layer of a semiconductor sandwiched between layers of a different semiconductor is called a *quantum well* (Figure 3.4a). 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 3.4b). Nanoscale superlattices of magnetic and ferroelectric materials can display remarkable properties not seen in bulk samples (see sections 11.3.11, 12.6).

A thin strip of semiconducting material is called a *quantum wire* (Figure 3.4c). In a quantum wire the electrons are confined in two dimensions and the structure behaves as a one-dimensional conductor. The ultimate degree of electron confinement occurs when the quantum wire is broken down into small units called a *quantum dot* (Figure 3.4d). These are regarded as zero-dimensional electronic structures.

Nanoparticles behave physically and chemically rather like free quantum dots. Inorganic

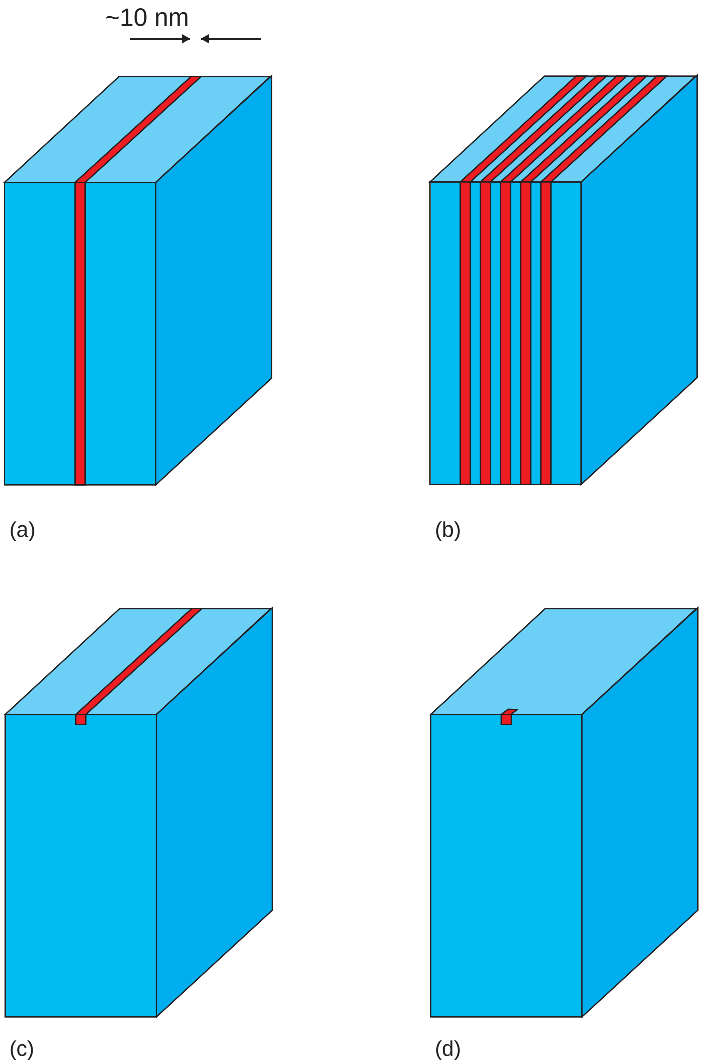


Figure 3.4 Electronic nanostructures: (a) a quantum well; (b) a series of quantum wells or superlattice; (c) a quantum wire; (d) a quantum dot.

nanoparticles such as those derived from semiconductors show fluorescent colour changes as the particle size grows smaller. This is a function of the way in which the bulk band structure gradually takes on atomic form. The colours of metal nanoparticles also vary with particle size and shape, a feature illustrated by medieval ruby glass which is coloured by gold nanoparticles embedded in the glassy matrix. As this indicates, metal nanoparticles are able to interact strongly with electromagnetic radiation, including light, studied in the emerging field of plamonics. *Plasmonic crystals* are generally metals containing a regularly spaced array of holes or particles that can manipulate light in prescribed ways.

Among the most widely investigated nanoparticles are those formed by carbon. *Fullerenes* are roughly spherical assemblies of carbon atoms linked by strong covalent bonds. The first example to be characterised, C_{60}, was called *Buckminsterfullerene* since the structure (Figure 3.5a) resembled the geodesic dome structure developed by Buckminster Fuller. Later these structures were also called *buckyballs*. The structure of C_{60} is a truncated icosahedron, and it is built of faces made up of pentagons and hexagons. A carbon atom is found at each vertex of the structure. Fullerenes have the electronic properties of quantum dots. Graphene, another material with important physical, optical and electronic properties, is a single layer of carbon atoms linked as in benzene and identical to a layer of graphite one atom thick (Figure 3.5b). Carbon nanotubes can be thought of as a layer of carbon atoms of the sort found in graphene coiled into a tube (Figure 3.5c). Carbon nanotubes behave as quantum wires and are being explored for connections in microelectronic circuits. The electronic and optical properties of both fullerenes and nanotubes can be modified by encapsulating other atoms, especially metal atoms, into the structure.

There are myriad examples of nanostructures in nature. Photonic crystals, ordered modulations of biological material, are to be found in many animals and give rise to iridescence in beetle cuticle, butterfly wings, feathers and countless other structures. These structures are being mimicked in the laboratory to produce novel optical materials for optoelectronics.

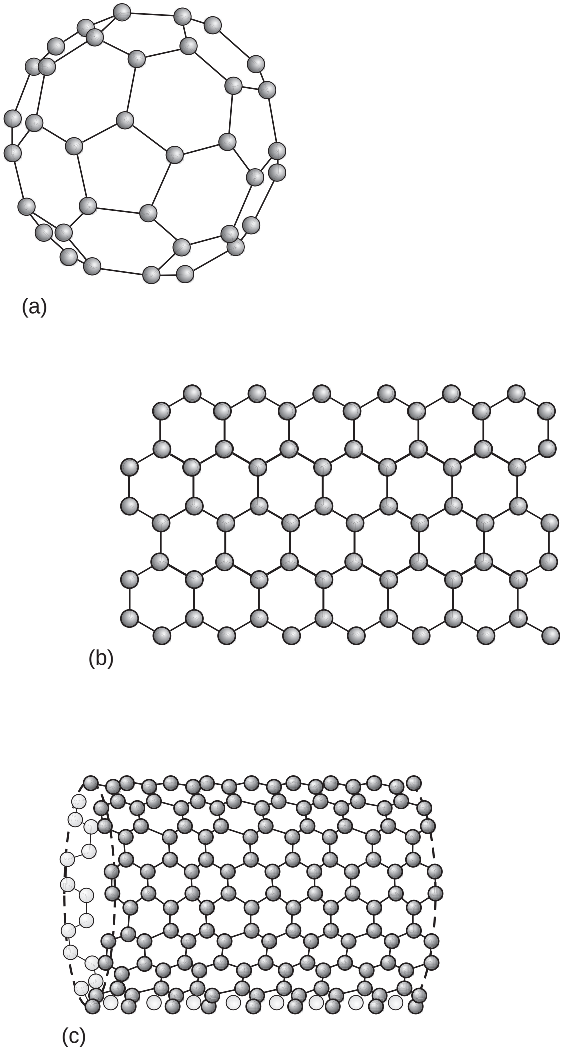


Figure 3.5 Carbon nanoparticles: (a) the truncated icosahedral (soccer ball) structure of a C_{60} buckyball; (b) the hexagonal structure of graphene; (c) carbon nanotubes, consisting of rolled-up graphene sheets. (Other tube configurations exist in addition to that drawn.)

3.3 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.

3.3.1 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 mica, quartz and feldspars (Figure 3.6a). Pure metals and alloys are also normally polycrystalline (Figure 3.6b). Many crystals grow from the melt with a branching shape or morphology that resembles a tree in form. These are called *dendrites* and the growth is called *dendritic growth* (Figure 3.6c). 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.

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 constraints 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.

3.3.2 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 upon working as the process introduces large numbers of defects (see below) 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 was the vulcanisation of rubber, which is 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.

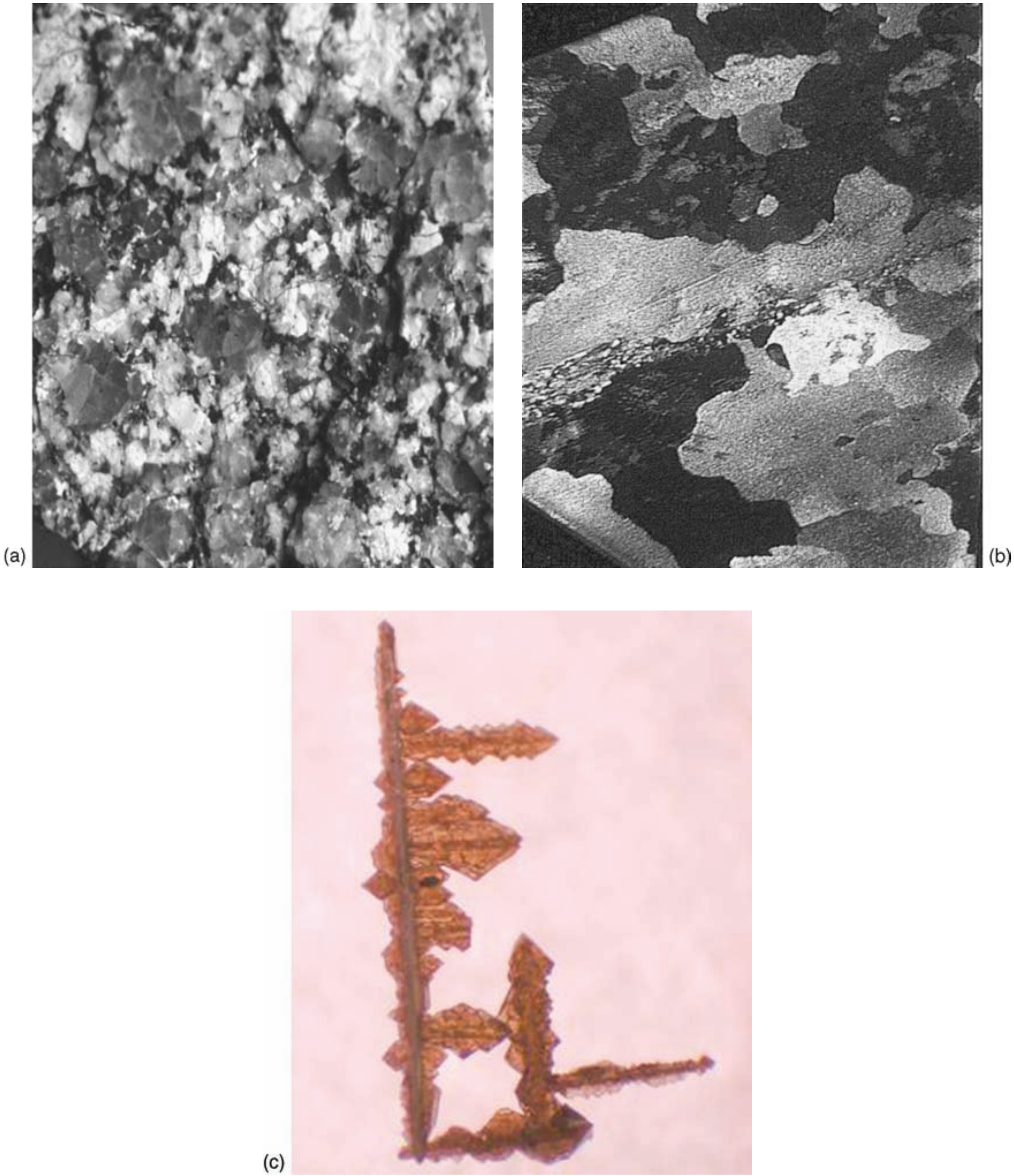


Figure 3.6 Optical micrographs of polycrystalline solids: (a) granite, composed of interlocking crystals of mica (black), approximately 2 mm in width, quartz and feldspars (colourless); (b) aluminium, consisting of interlocking grains up to 1 cm in width; (c) dendritic rutile (TiO_2) crystals approximately 5 mm long; each 'branch' is a separate crystal, and the whole group forms a polycrystalline solid.

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 (Section 8.1). 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.

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 Chapter 4). In this state, the fluid does not exert surface tension, and it is possible to remove it without collapse of the solid framework.

3.4 Point defects

Defects in crystalline solids are important because they modify properties. For example, just a trace of chromium impurity changes colourless aluminium oxide 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.

3.4.1 Point defects in crystals of elements

Crystals of solid elements such as silicon contain 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*.

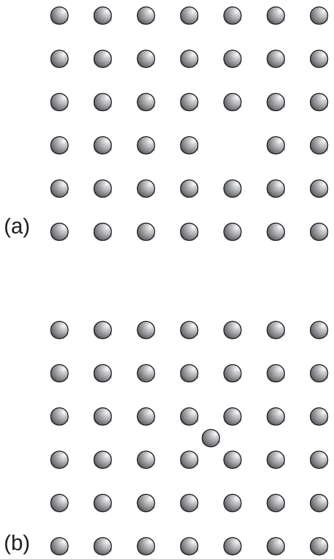


Figure 3.7 Point defects in pure crystals such as silicon: (a) a vacancy; (b) a (self-) interstitial.

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 3.7). Such vacancies and interstitials, which occur in even the purest of materials, are called *intrinsic defects*.

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 defects lowers the Gibbs energy, but ultimately large numbers of point defects results in an increase in Gibbs energy. The minimum in the curve represents the equilibrium situation that will exist at a given temperature (Figure 3.8). 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 is expressed by the formula:

$$n_\mathrm{d} \approx N \exp\left(\frac{-\Delta H}{k_\mathrm{B} T}\right) \tag{3.1}$$

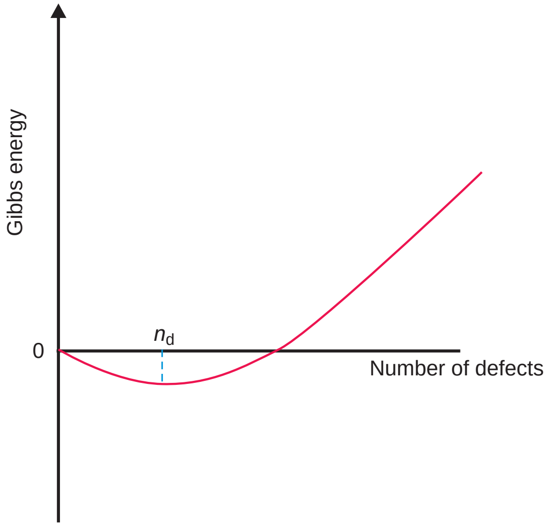


Figure 3.8 The Gibbs energy of a crystal as a function of the number of point defects present. At equilibrium, n_d defects are present in the crystal.

where n_d is the number of defects per unit volume, 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 *fraction* of atom sites that contain a defect, n_d/N, at any temperature can be calculated if the enthalpy of defect formation is known.

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 length of 0.5431 nm, and contains 8 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:

$$N = \frac{\rho \times N_A}{\text{molar mass}}$$

where N_A is Avogadro's constant.

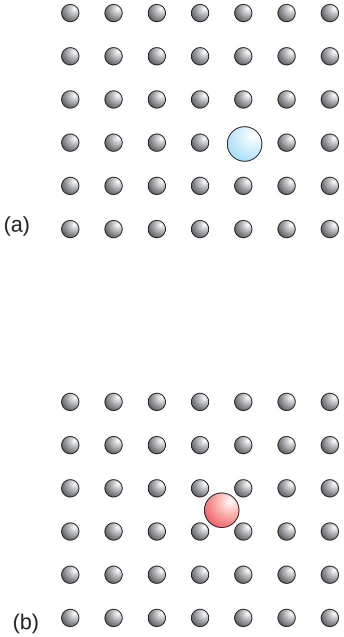


Figure 3.9 Impurity or dopant point defects in a crystal: (a) substitutional; (b) interstitial.

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 in 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 3.9). There is no simple thermodynamic formula for the number of impurities present in a crystal.

3.4.2 *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 crystal structure,

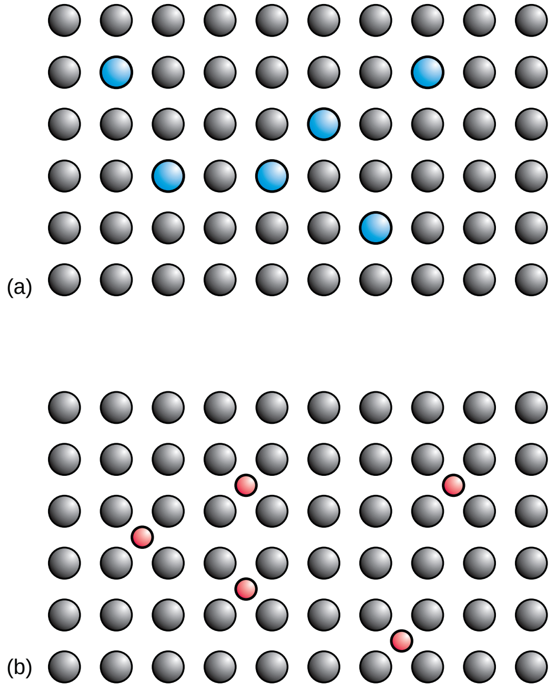


Figure 3.10 Solid solutions: (a) random substitutional solid solution; (b) random interstitial solid solution.

and in alloys the two metals are distributed at random (Figure 3.10a). 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. *Interstitial* solid solutions form when atoms enter spaces between the atoms in a crystal (Figure 3.10b). The impurities must be small and are typically elements from the first row of the periodic table, such as C and N. Steel is composed of iron containing about 0.5% of interstitial carbon.

3.4.3 Schottky defects

In ionic compounds the charges must remain balanced when point defects are introduced into the crystal. Take the compound sodium chloride, which contains equal numbers of sodium (Na^+) and chlorine (Cl^-) ions. To separate out the effects of the anions from those of the cations it is convenient to refer to the *anion sublattice* for the Cl^- array and the *cation sublattice* for the Na^+ array.

Vacancies on the cation sublattice will change the composition of the compound. As the constituents are charged, this will also alter the charge balance in the crystal. If x vacancies occur, the formula of the crystal will now be $Na_{1-x}Cl$ and the overall material will have an excess negative charge of $x-$, because the number of chloride ions is greater than the number of sodium ions by this amount. The formula could be written $[Na_{1-x}Cl]^{x-}$. If x vacancies are found on the anion sublattice, the material will take on an overall positive charge, because the number of sodium ions now outnumbers the chlorine ions, and the formula becomes $[NaCl_{1-x}]^{x+}$. Now crystals of sodium chloride do not normally show an overall negative or positive charge or have a formula different to NaCl. This means that equal numbers of vacancies must occur on both sublattices.

The defects arising from balanced populations of cation and anion vacancies in *any* crystal (not just NaCl) are known as *Schottky defects* (Figure 3.11a). Any ionic crystal of formula MX must contain equal numbers of cation vacancies and anion vacancies. In such a crystal, one Schottky defect consists of one cation vacancy plus one anion vacancy. (These vacancies need not be near to each other in the crystal.) The number of Schottky defects in a crystal of formula MX is equal to one half of the total number of vacancies. In crystals of a more complex formula, charge balance is also preserved. The ratio of two anion vacancies to one cation vacancy will hold in all compounds of formula MX_2, such as titanium dioxide, TiO_2. Schottky defects in this material will introduce twice as many anion vacancies as cation vacancies into the structure. In crystals with a formula M_2X_3, a Schottky defect will consist of two vacancies on the cation sublattice and three vacancies on the anion sublattice. In Al_2O_3, for example, two Al^{3+} vacancies will be balanced by three O^{2-} vacancies.

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 3.8. This means that Schottky

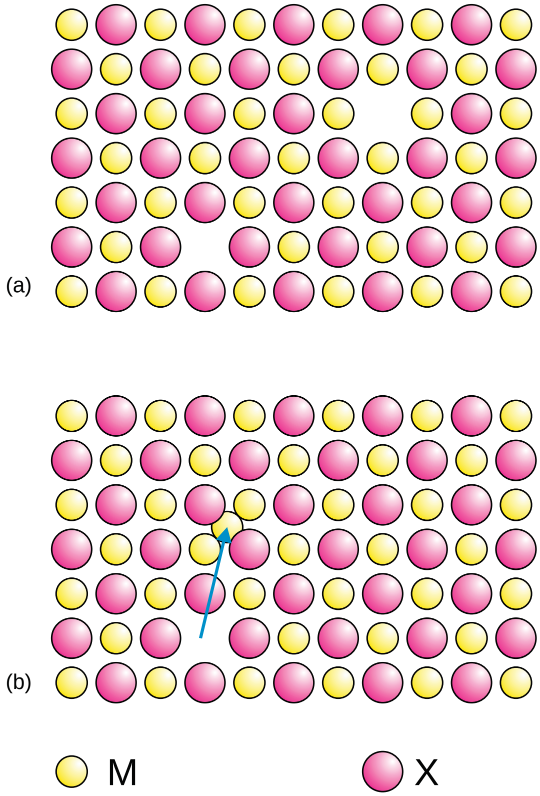


Figure 3.11 Point defects in an ionic crystal of formula MX: (a) Schottky defects; (b) Frenkel defects.

Table 3.3 The formation enthalpy of Schottky defects in some alkali halide compounds of formula MX^*

Compound	ΔH_S/kJ mol^{-1}	ΔH_S/eV**
LiF	225.2	2.33
LiCl	204.1	2.12
LiBr	173.4	1.80
LiI	102.4	1.06
NaF	233.1	2.42
NaCl	225.7	2.34
NaBr	203.0	2.10
NaI	140.9	1.46
KF	262.0	2.72
KCl	244.5	2.53
KBr	224.6	2.33
KI	153.0	1.59

*All compounds have the *halite* (NaCl) structure.

**The energy unit eV, electron-volt, is widely used for defect energies.

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 a crystal with formula MX at equilibrium is given by:

$$n_S \approx N e^{-\Delta H_S / 2k_B T} \tag{3.2}$$

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). Sometimes equation (3.2) is written in the form:

$$n_S \approx N e^{-\Delta H_S / 2RT} \tag{3.3}$$

In this case ΔH_S is in J mol^{-1}, and represents the enthalpy required to form 1 mole of Schottky defects, and R is the gas constant. (Note that equations 3.2 and 3.3 only apply to materials with a composition MX.)

The fraction of vacant sites at any temperature, n_S/N, can be calculated if the enthalpy of defect formation, ΔH_S is known. To obtain the absolute number of defects in the solid, it is necessary to know the number of ions, N, of the appropriate type in a unit volume of the crystal. This is obtained from the unit cell dimensions or the density of the material, as described above.

3.4.4 Frenkel defects

It is possible to imagine a defect in ionic crystals similar to the interstitial defects described above. Such defects are known as *Frenkel defects*. In this case, an ion from one sublattice moves to a normally empty place in the crystal, leaving a vacancy behind. One Frenkel defect consists of an interstitial ion plus a vacancy (Figure 3.11b). 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

Table 3.4 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*	139.7	1.45	CaF_2**	261.4	2.71
AgBr*	109.0	1.13	SrF_2**	167.4	1.74
β-AgI*	57.8	0.60	BaF_2**	184.3	1.91

*Frenkel defects on the cation sublattice of a halite (NaCl) structure compound.

**Frenkel defects on the anion sublattice of a fluorite (CaF_2) structure compound.

interstitial site. It is not necessary to displace two ions to form the Frenkel defect.

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 in some of the normally occupied silver sites. (The Br^- ions are not involved in the defects.) Frenkel defects in AgBr make possible both black and white and colour film photography.

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. There is one small difference compared with these equations: 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:

$$n_F \approx (NN^*)^{1/2} e^{-\Delta H_F/2k_BT} \qquad (3.4)$$

when ΔH_F is the enthalpy of formation of single Frenkel defect, k_B is the Boltzmann constant, and T the absolute temperature. This is also often expressed in molar quantities:

$$n_F \approx (NN^*)^{1/2} e^{-\Delta H_F/2RT} \qquad (3.5)$$

where 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 (see Table 3.4). 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.

3.4.5 Non-stoichiometric compounds

Although molecules have a fixed formula and composition, many non-molecular solids are found to exist over a range of composition. 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 compositional 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 uses 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

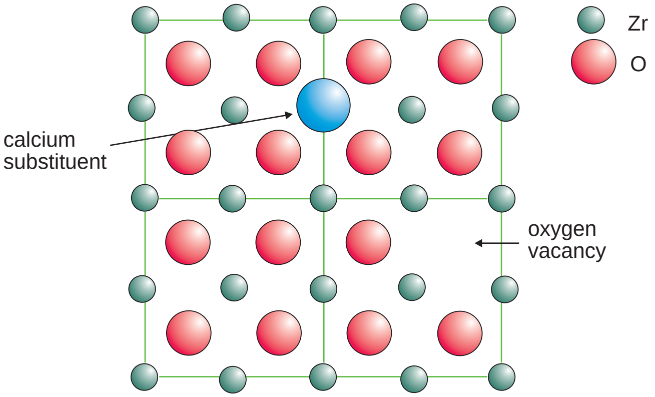


Figure 3.12 The structure of calcia-stabilised zirconia. The crystal contains a number of Ca^{2+} ions substituted for Zr^{4+} ions. Each Ca^{2+} ion is accompanied by an O^{2-} vacancy to maintain charge neutrality. The unit cell of the parent structure is outlined.

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 3.12). 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 $Ca_x^{2+} Zr_{1-x}^{4+} O_2$. The crystal compensates for the extra negative charge by leaving some of the anion sites unoccupied. The number of vacancies in 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 $Ca_x^{2+} Zr_{1-x}^{4+} O_{2-x}$.

Calcia-stabilised zirconia provides a good example of the consequences of incorporating an ion with a lower valence into a crystal structure. When a material is 'doped' with substitutional impurity cations of lower charge, anion vacancies are a common method of achieving charge balance. This has a significant effect on the properties of the solid, as the diffusion coefficient of ions on the sublattice containing the vacancies is greatly increased. Calcia-stabilised zirconia is widely used as a solid electrolyte in electrochemical cells and sensors (Section 9.2.3).

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 are 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 9.3). They are typified by disulphides such as TiS_2 (Figure 3.13) and NbS_2. Small atoms such as Li can enter the structure between these layers to form non-stoichiometric phases such as Li_xTiS_2. Because the bonding between the layers is weak, this process is easily

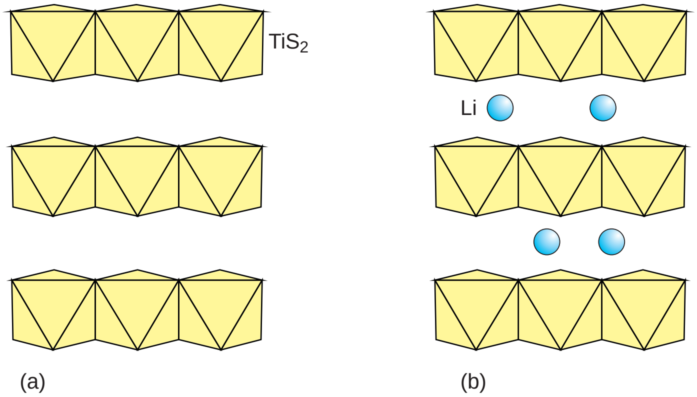


Figure 3.13 (a) The structure of TiS_2, composed of layers of TiS_6 octahedra, linked by weak bonds. (b) Insertion of Li (or other) atoms between the layers.

reversible and the compound can act as a convenient reservoir of Li atoms in lithium batteries.

Another group of insertion compounds of importance is 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, the so-called *van der Waals gap*. 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 (Section 10.5.1)

3.4.6 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 3.5). 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

Table 3.5 The Kröger-Vink notation for defects in crystals*

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	$Y_i^{n'}$
Free electron	e'	Free hole	$h^\bullet$
Associated defects	$(V_M V_Y)$		

*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.

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.

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 a 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 nonmetal 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 $+$.

To illustrate this concept, let us determine the effective charge on a sodium vacancy, V_{Na}, in the NaCl structure. The real charge on the vacancy is 0. The real 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 disulphide, 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.

3.5 Linear, planar and volume defects

3.5.1 Edge 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 (Section 10.3). The 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 half plane of atoms inserted into the crystal (Figure 3.14). These dislocations are instrumental in allowing metals to undergo plastic deformation. 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 3.15). The Burgers circuit starts and ends 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.

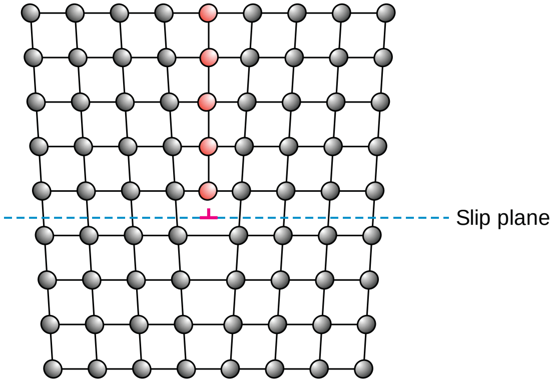


Figure 3.14 An edge dislocation, consisting of an extra half plane of atoms inserted in a crystal. The dislocation line, marked ⊥, is perpendicular to the plane of the figure.

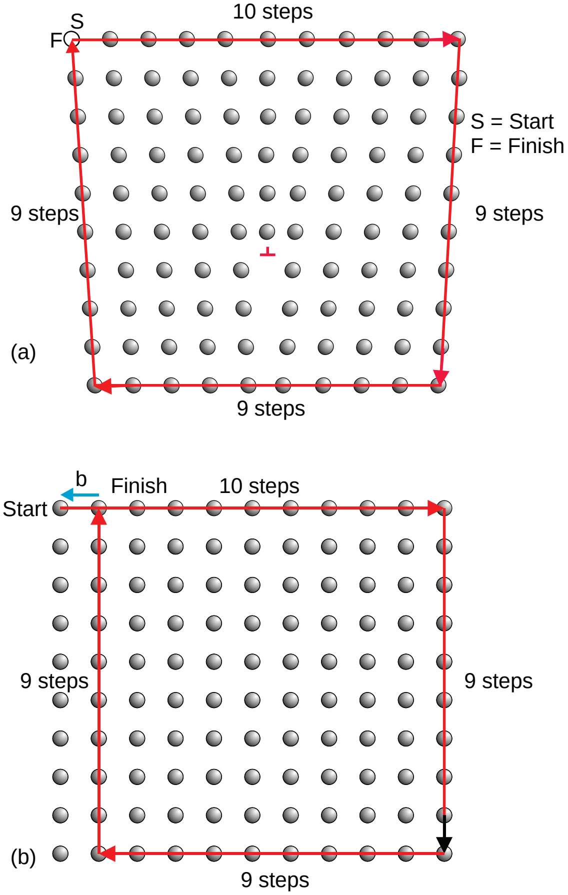


Figure 3.15 The Burgers vector of an edge dislocation: (a) a circuit around an edge dislocation, and (b) the corresponding circuit in a perfect crystal. The vector linking the finishing atom to the starting atom in (b) is the Burgers vector of the dislocation.

The Burgers vector of an edge dislocation is perpendicular to the dislocation line.

3.5.2 Screw dislocations

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 atomic planes (Figure 3.16). 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 3.17). Screw dislocations play an important part in crystal growth.

3.5.3 Partial and mixed dislocations

Dislocations can be imaged and their Burgers vectors determined using transmission electron microscopy. This technique has shown that many dislocations have a Burgers vector that is less than the repeat distance of the structure. These are called *partial dislocations*.

A dislocation line separates a region of crystal that has moved 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. Point defects become increasingly mobile at high temperatures, and if a metal is held at a temperature near to its melting point, these defects can migrate from site to site. If sufficient defects aggregate, a dislocation loop can form (Figure 3.18). 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 character 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*.

3.5.4 Planar defects

In many cases, the most important planar defects in a solid are the *external surfaces*. The creation of

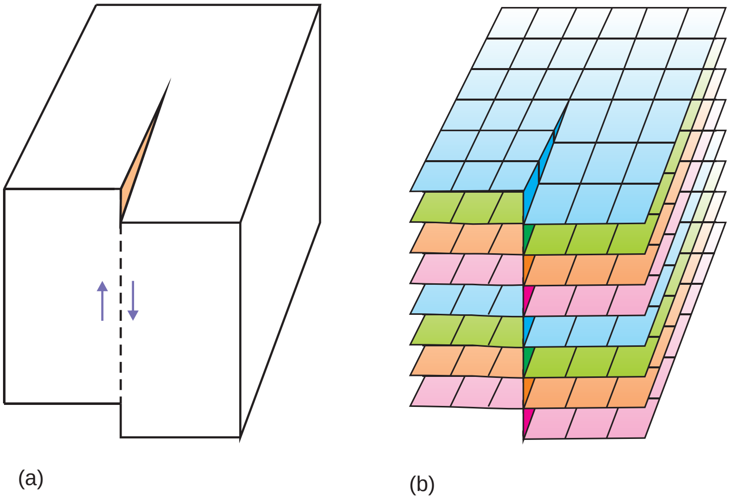


Figure 3.16 (a) A screw dislocation can be formed (hypothetically) by cutting a crystal and displacing the halves. (b) The planes in a screw dislocation coil around the dislocation line like a spiral staircase.

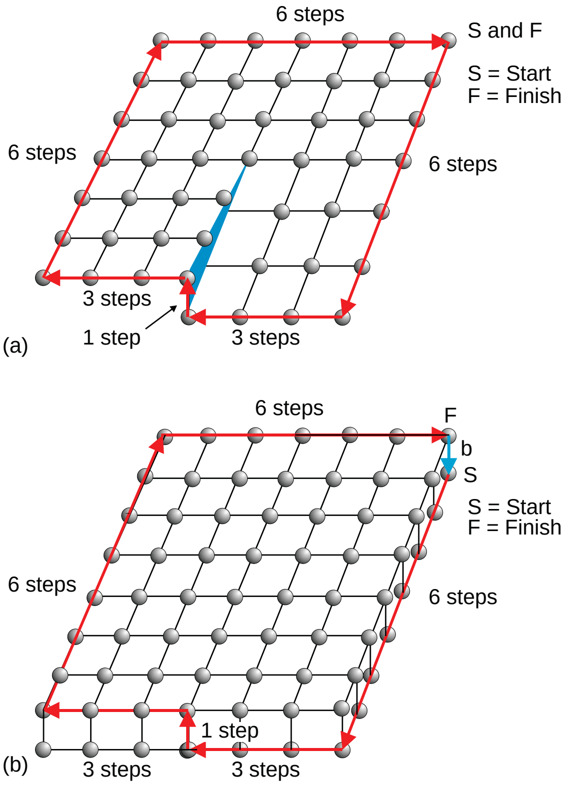


Figure 3.17 The Burgers vector of a screw dislocation: (a) a circuit around a screw dislocation, and (b) the corresponding circuit in a perfect crystal. The vector linking the finishing atom to the starting atom in (b) is the Burgers vector of the dislocation.

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.

Grain boundaries are interfaces between crystallites in a polycrystalline array (Figure 3.19a). The energy of these boundaries, much of which is 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. Grain boundaries 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. Grain boundaries also increase the electrical resistance of a polycrystalline solid compared with 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 3.19b). 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 (APBs) are planes within a crystal across which one part of the crystal has been displaced with respect to the other side (Figure 3.19c). 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 3.19d). 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 $Ti_{10}O_{19}$.

3.5.5 Volume defects: precipitates

Volume defects are regions of an impurity phase in the matrix of a material. *Precipitates* in a solid compose most volume defects (Figure 3.20). Precipitates form in a variety of circumstances. Solid solutions

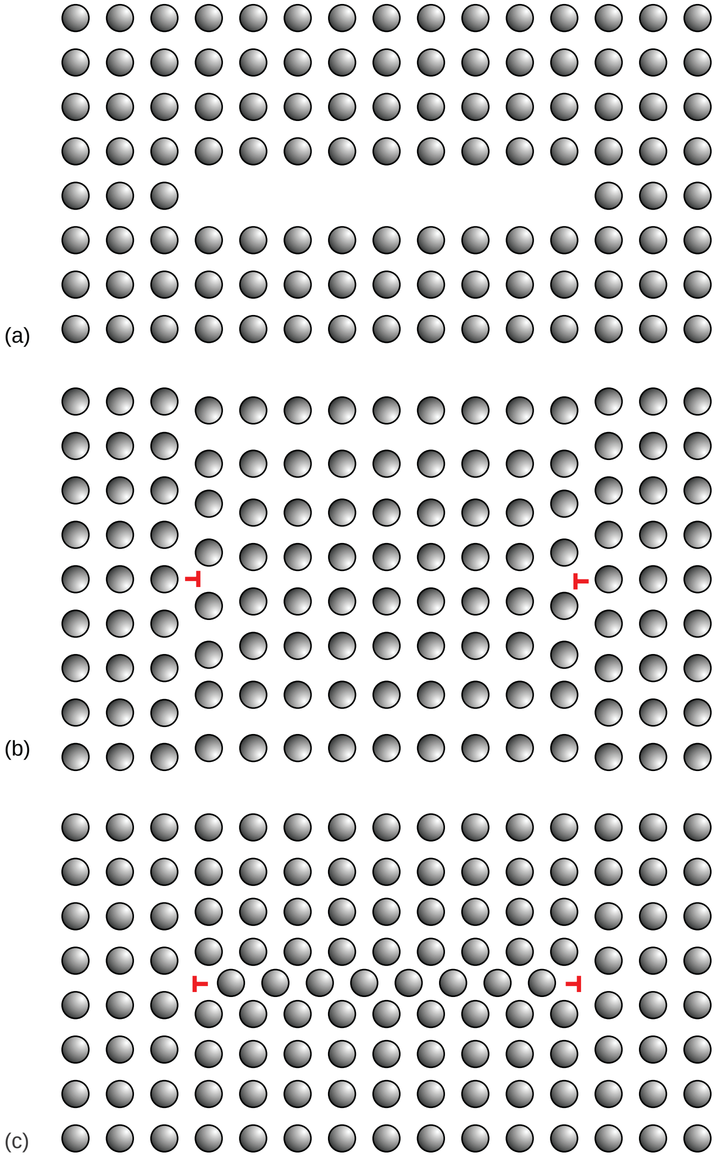


Figure 3.18 Formation of dislocation loops: (a) the aggregation of vacancies onto a plane in a crystal; (b) collapse of the crystal to form a dislocation loop; (c) aggregation of interstitials to form a dislocation loop.

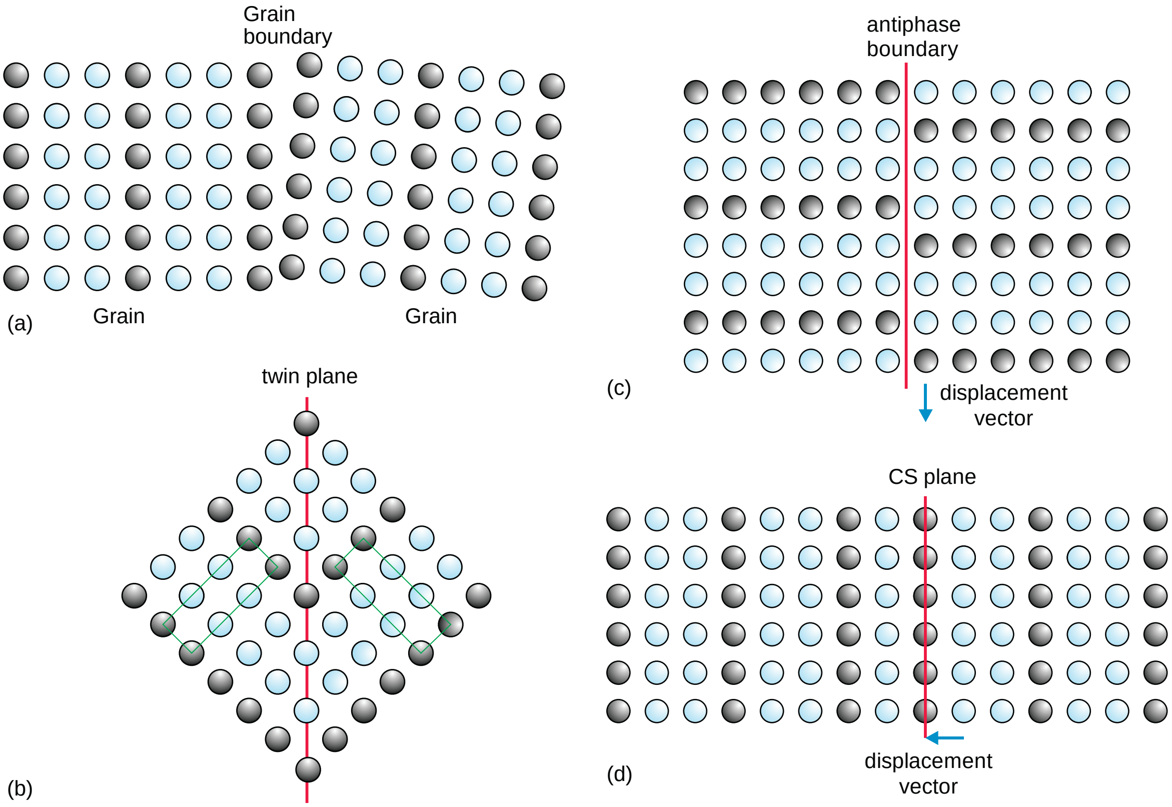


Figure 3.19 Surfaces and boundaries in a crystal: (a) grain boundaries; (b) twin plane; (c) antiphase boundary; (d) crystallographic shear (CS) plane.

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 recrystallize 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.

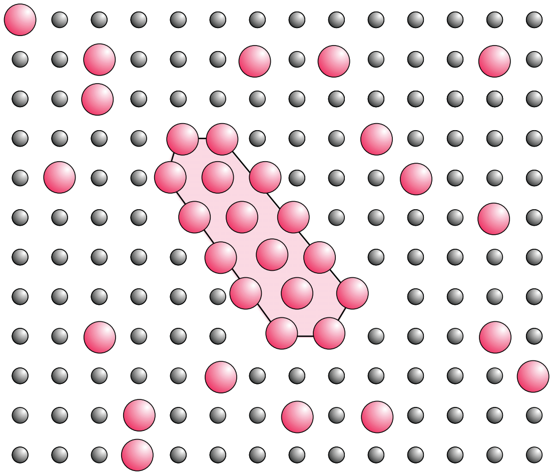


Figure 3.20 A precipitate formed by clustering of impurity atoms in a crystal.

Further reading

Callister, W.D. (2000) *Materials Science and Engineering, an Introduction*, 5th edn. John Wiley & Sons, Ltd., New York.

Eckert, J., Stucky, G.D. and Cheetham, A.K. (1999) Partially disordered inorganic materials. *Materials Research Society Bulletin*, **24**: 31.

Kingery, W.D. (1987) A role for ceramic materials science in art, history and archaeology. *Journal of Materials Education*, **9**: 679–718.

Kingery, W.D., Bowen, H.K. and Uhlmann, D.R. (1960) *Introduction to Ceramics*. Wiley-Interscience, New York.

Kraynik, A.M. (2003) Foam structure: from soap froth to solid foams. *Materials Research Society Bulletin*, **28**: 275.

Megaw, H. (1973) *Crystal Structures*. W.B. Saunders, Philadelphia.

Smith, W.F. (1993) *Foundations of Materials Science and Engineering*, 2nd edn. McGraw-Hill, New York.

Tilley, R.J.D. (2008) *Defects in Solids*. John Wiley & Sons Ltd., Hoboken, NJ.

An overview of the properties of graphene is given in a series of reviews in *Nature Supplement* 483 [No. 7389] S29–S74 (2012).

Problems and exercises

Quick quiz

1 Hydrogen bonding is found in:
(a) Solid and liquid hydrogen.
(b) Hydrocarbons.
(c) Compounds containing oxygen and hydrogen.

2 Van der Waals bonds are due to:
(a) Dispersion forces.
(b) Ion–dipole forces.
(c) Dipole–dipole forces.

3 A refractory oxide is:
(a) An oxide that is difficult to prepare.
(b) An oxide that is rare.
(c) An oxide resistant to high temperatures.

4 The microstructure of a solid is at a scale of:
(a) $1–10^{-2}$ m.
(b) $10^{-4}–10^{-6}$ m.
(c) $10^{-7}–10^{-9}$ m.

5 A polycrystalline solid is composed of:
(a) Polymer chains.
(b) Glass and crystallites.
(c) Many small crystals.

6 A glass is:
(a) A non-crystalline inorganic solid.
(b) A crystalline inorganic solid.
(c) A solid containing silica.

7 Sintering is a process that involves:
(a) Cross-linking polymer chains.
(b) Heating powdered ceramics or metals.
(c) Cold-working a solid to make it harder.

8 An important step in the fabrication of glass ceramics is:
(a) Devitrification.
(b) Sintering.
(c) Quenching.

9 In a quantum well, electrons are confined in:
(a) Three dimensions.
(b) Two dimensions.
(c) One dimension.

10 Fullerenes are:
(a) Large carbon molecules with a roughly spherical form.
(b) Large carbon molecules with a roughly tubular form.
(c) Large carbon molecules with a chain-like form.

11 The number of different intrinsic point defects possible in a single crystal of a pure element is:
(a) One.
(b) Two.
(c) Three.

12 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.

13 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.

14 The Burgers vector and dislocation line are normal to each other in:

(a) A screw dislocation.

(b) A partial dislocation.

(c) An edge dislocation.

15 The Burgers vector and dislocation line are parallel to each other in:

(a) A screw dislocation.

(b) A partial dislocation.

(c) An edge dislocation.

16 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.

17 Dislocations are:

(a) Planar defects.

(b) Line defects.

(c) Point defects.

18 A twin boundary in a solid is an example of:

(a) A line defect.

(b) A planar defect.

(c) A volume defect.

19 A precipitate is a:

(a) Point defect.

(b) Planar defect.

(c) Volume defect.

Calculations and questions

3.1 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) estimate the equilibrium interatomic separation of this molecule.

3.2 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 3.1. Compare this with the boiling point of argon (question 3.22).

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

3.4 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.

3.5 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 3.1 and 3.4, estimate the Lennard-Jones constants for krypton, Kr.

3.6 Derive the relationship:

$$V_{LJ} = 4V_m \left[\left(\frac{r(0)}{r} \right)^6 - \left(\frac{r(0)}{r} \right)^{12} \right]$$

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

3.7 The enthalpy of formation of vacancies in pure nickel is $\Delta H = 97.3$ kJ mol^{-1}. What is the fraction of sites vacant at 1100 °C?

3.8 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?

3.9 The enthalpy of formation of vacancies in pure gold is $\Delta H = 123.5\,\text{kJ mol}^{-1}$. The density of gold is 19281 kg m^{-3}. What number of atom positions is vacant at 1000 °C?

3.10 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}. What number of atom positions is vacant at 600 °C?

3.11 Calculate how the fraction of Schottky defects in a crystal of KCl varies with temperature if ΔH_S is 244 kJ mol^{-1}.

3.12 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.

3.13 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 Frenkel defect formation in a crystal of AgBr at 300 K.

3.14 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.

3.15 Calculate the enthalpy of formation of Frenkel defects in NaBr, using the data on the number of defects present given in the 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}

3.16 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 °C and 2000 °C. How many vacancies are present at these temperatures? CaO has a density of 3300 kg m^{-3}.

3.17 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}.

3.18 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.

3.19 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}.

3.20 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.

3.21 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 $1.12 \times 10^{22}\,m^{-3}$ at 600 °C. Calculate the energy of formation of these defects.

3.22 The melting points and boiling points of the noble gases are given in the table. Explain these trends, and 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

3.23 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.

3.24 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?

3.25 What defects will form in the crystals made by adding small amounts of compound a to compound b?

a. LiBr, b. $CaBr_2$.
a. $CaBr_2$, b. LiBr.
a. MgO, b. Fe_2O_3.
a. MgO, b. NiO.

3.26 What defects will form in the crystals made by adding small amounts of compound a to compound b.

a. $CdCl_2$, b. NaCl.
a. NaCl, b. $CdCl_2$.
a. Sc_2O_3, b. ZrO_2.
a. ZrO_2, b. HfO_2.

3.27 Show that the number of metal atom sites N in a crystal of composition MX is given by:

$$N = \frac{\rho \times N_A}{\text{molar mass}}$$

where ρ is the density of MX and N_A is the Avogadro constant.

3.28 Sodium chloride has a density of 2165 kg m^{-3}. The unit cell, which is cubic, with a cell edge of 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.

4

Phase diagrams

- What is a binary phase diagram?
- What is a eutectic point?
- What is the difference between carbon steel and cast iron?

4.1 Phases and 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 the 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 under specified conditions, most often composition, temperature and pressure. Phase diagrams mostly relate to *equilibrium conditions*. If a diagram represents non-equilibrium conditions, it is called an *existence diagram*. Phase diagrams essentially display thermodynamic information and can be constructed using thermodynamic data (Section 4.5). The conditions limiting the existence and coexistence of phases are given by the (*Gibbs*) *Phase Rule*.

Phases are made up of various combinations of *components*, which are the chemical substances sufficient to construct the phase diagram. A component can be an element, such as carbon, or a compound, such as sodium chloride. The components chosen to display phase relations are the simplest that allow all phases to be described.

4.1.1 One-component (unary) systems

In a *one-component*, or *unary*, system, only one chemical component is required, for example, Fe, H_2O or CH_4. There are many one-component systems, including all pure elements and compounds. 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 (1 atmosphere $= 1.01325 \times 10^5$ Pa), plotted along the ordinate (y-axis) (Figure 4.1).

The areas of the diagram within which a single phase exists are labelled with the name of the phase present. The phase or phases occurring at a given temperature and pressure are read from the diagram. The areas over which single phases occur are

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

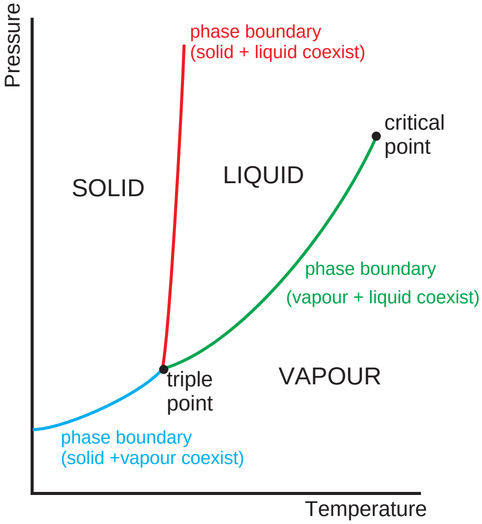


Figure 4.1 The generalised form of a one-component phase diagram.

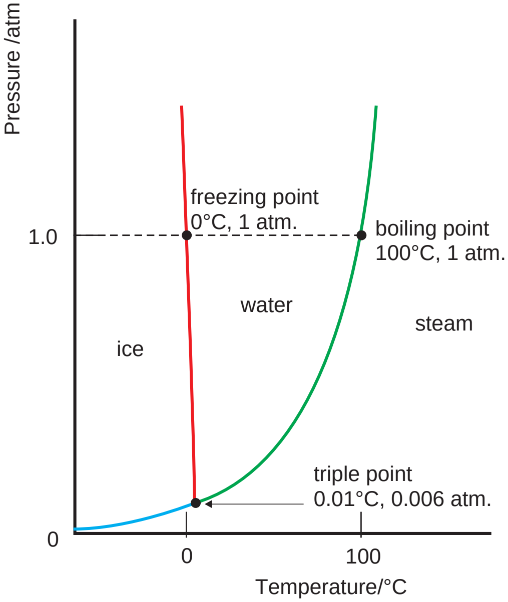


Figure 4.2 The approximate phase diagram for water, not to scale.

bounded by lines called phase boundaries. 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*.

Perhaps the most important one-component system for life on Earth is that of water (Figure 4.2). The three phases found are ice (solid), water (liquid), and steam (vapour). The ranges of temperature and pressure over which these phases are found are read from the diagram. For example, at 1 atm pressure and 50°C, water is the phase present. In a single-phase region, *both* the pressure and the temperature can be changed independently of one another without changing the phase present. For example, liquid water exists over a range of temperature and pressure, and either can be varied (within the limits given on the phase diagram) without changing the situation.

On the phase boundaries, two phases coexist indefinitely, ice and water, water and steam, or ice and steam. If a variable is changed, the two-phase equilibrium is generally lost. In order to preserve two-phase equilibrium, *one* variable, *either* pressure *or* temperature, can be changed at will, but the other *must also change*, by *exactly the amount specified in the phase diagram*, to maintain two phases in co-existence and so to return to the phase boundary.

The critical point of water, at 374°C and 218 atm, is the point at which water and steam become identical. The triple point is found at 0.01°C and 0.00605 atm. Only at this point do the phases water, ice and steam occur together. Any change in either the temperature or the pressure destroys the three-phase equilibrium.

The slopes of the phase boundaries give some information about the change of boiling and freezing points as the pressure varies. For example, the phase boundary between water and steam slopes upwards to the right. This indicates that an increase in pressure will favour liquid compared to 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 for supposing that liquid water might be found at depths under the surface of some of the cold outer moons in the solar system.

A phase diagram can be used to explain the pattern of temperature changes observed as a substance cools (Figure 4.3). For example, a sample of water at A will cool steadily until point B, on the water–ice phase boundary, is reached. The slope of the temperature versus time plot, called a *cooling curve*, will change smoothly. At point B, the melting/freezing point, T_m, if there is any further cooling, ice will begin 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 again to point C and a smooth cooling curve will again be obtained.

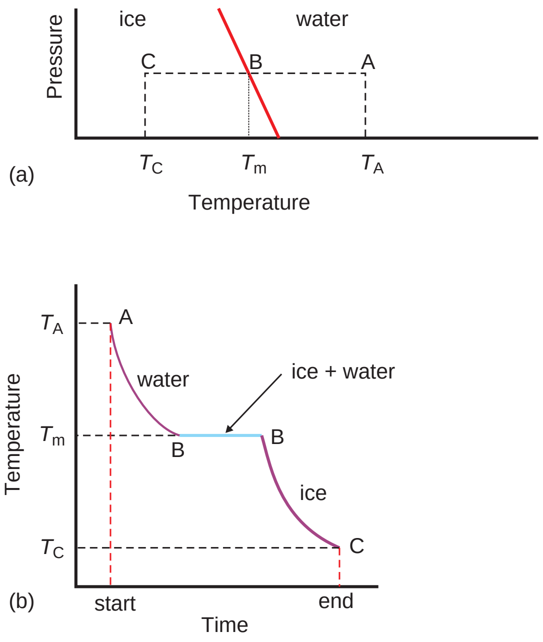


Figure 4.3 (a) A small part of the water phase diagram. (b) The cooling curve generated as a uniform sample of water cools from temperature T_A, liquid, to temperature T_C, solid.

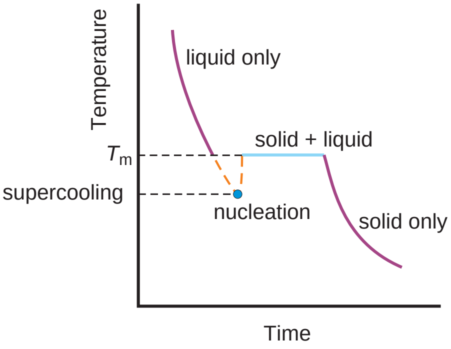


Figure 4.4 A cooling curve showing supercooling.

A change in slope of a cooling curve is an indication that the system is passing across a phase boundary, irrespective of the complexity of the system. Cooling curves are therefore useful in mapping out the presence of phase boundaries and in the construction of phase diagrams.

This form of cooling curve will be found in any one-component system if a sample is cooled slowly through a phase boundary, so that the system is always at equilibrium. Normal rates of cooling are faster, and experimental cooling curves for liquids often continue below the break expected at a phase boundary (Figure 4.4). 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.

4.1.2 The phase rule for one-component (unary) systems

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

$$P + F = C + 2 \tag{4.1}$$

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:

$$P + F = 3 \qquad (4.2)$$

The conditions limiting the existence and coexistence of phases can now be assessed. With reference to the phase diagram for water (Figure 4.2), when only one phase is present, for example ice, equation (4.2) becomes: $F = 2$. That is, the region over which ice is stable has to be specified in terms of two variables, normally temperature and pressure. This means, as stated above, that 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 equation (4.2): $F = 0$. There are no degrees of freedom available, and so 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 equation (4.2): $F = 1$. This phase boundary is characterised by one degree of freedom, which 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 (Figure 4.3) 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 one atmosphere, 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.

4.2 Binary phase diagrams

4.2.1 Two-component (binary) systems

Binary systems contain two components, for example, $Fe + C$, $NaNbO_3 + LiNbO_3$, or $Pb + Sn$. 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 (Figure 4.5a). A single phase will be represented by a *volume* in the diagram. Phase boundaries form *two-dimensional surfaces* in the representation, and three phases will coexist along a *line*.

As most experiments are carried out at atmospheric pressure, a planar diagram using temperature and composition as variables is usually sufficient (Figure 4.5b). These sections are called *isobaric* phase diagrams. 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*. That is, the total weight is expressed as 100 grams (or kilograms), and the amount of each component is given as x grams and $(100 - x)$ grams. In chemical work, *atom percentages* or *mole percentages* are used. In these cases, the amounts of each component are given by x atoms (moles) and $(100 - x)$ atoms (moles). Note carefully that the composition gives the amount of the components present, not the composition of the phases present, unless the composition falls in a single phase region. In constant-pressure diagrams, the temperature is specified as °C. A single phase occurs over an area in the figure, phase boundaries are drawn as lines, and three phases occur at a point. In all of the binary phase diagrams discussed later, it is assumed that pressure is fixed at one atmosphere.

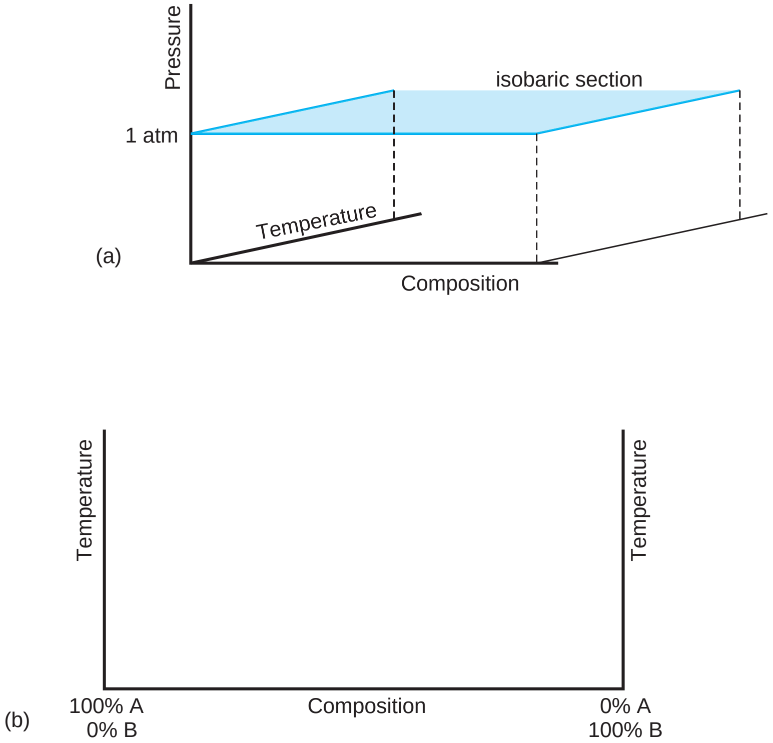


Figure 4.5 (a) A three-axis pressure–temperature–composition frame required to display the phase relations in a binary system. (b) Isobaric sections, in which the pressure is fixed, use only temperature and composition.

4.2.2 The phase rule for two-component (binary) systems

The phase rule for a binary system is given by equation (4.3):

$$P + F = 4 \qquad (4.3)$$

When only one phase is present, $P = 1$, so from equation (4.3), $F = 3$. 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 equation (4.3), $F = 2$. The variance is seen to be equal to two, and 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 coexist is three (e.g. solid, liquid and vapour). In this case, $P = 3$ and equation (4.3) gives $F = 1$. With only one degree of freedom available to the system, three phases will coexist along a line in the phase diagram (Table 4.1).

4.2.3 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

Table 4.1 A comparison of phase diagrams of one-component and two-component systems

	Unary	Binary
Axes	2	3
Single-phase region	Area	Volume
Phase boundary	Line	Surface
Three-phase coexistence	Point	Line

system provides a good example (Figure 4.6). At the top of the diagram, corresponding to the highest temperatures, one homogeneous phase, liquid, occurs. In this, the Cu and Ni atoms are mixed up together at random. In the copper-rich part of the diagram, the liquid can be considered as a solution of nickel in molten copper, and in the nickel-rich region, the liquid can be considered a solution of copper in liquid nickel.

At the bottom of the diagram, corresponding to 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.

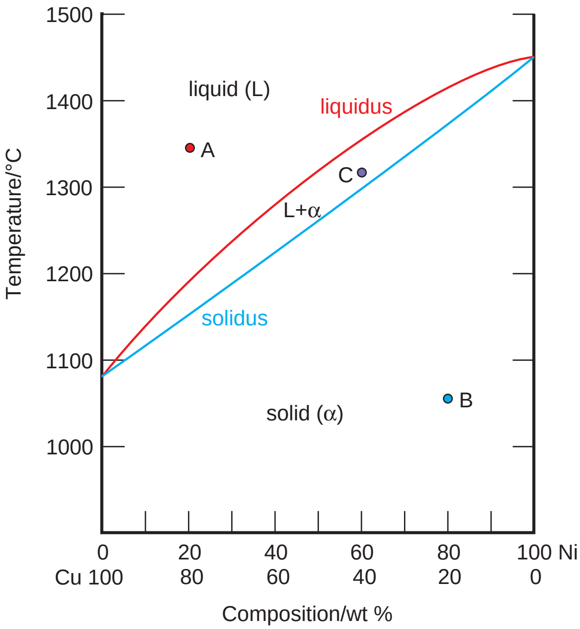


Figure 4.6 The copper (Cu)–nickel (Ni) phase diagram at atmospheric pressure.

Between the liquid and solid phases, phase boundaries delineate a lens-shaped region. Within this area, solid α and liquid L 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*.

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. Moreover, breaks in the smooth curve occur as the sample passes both the liquidus, at T_l and the solidus, at T_s (Figure 4.7). Cooling curves for samples spanning the whole compositional range can be used to map out the positions of the solidus and liquidus.

The most obvious information found in the diagram is the phase or phases present at any temperature. Thus, suppose that a mixture of 50 grams of copper and 50 grams of nickel is heated. At 1400°C, one phase will be present, a homogeneous liquid. At 1100°C, one phase will also be present, a homogeneous solid, the α phase. At 1250°C two phases are present, liquid (L) and solid (α).

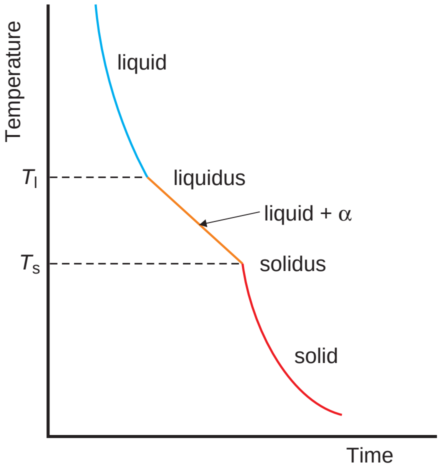


Figure 4.7 A cooling curve for a sample passing through a two-phase liquid + solid region.

The composition of any point in the diagram is simply read from the composition axis. Thus, point A in Figure 4.6 has a composition of 80 weight % (wt%) copper (and thus 20 wt% nickel). Point B has a composition of 20 wt% copper (and thus 80 wt% nickel). Point C has an *average* composition of 40 wt% copper (and thus 60 wt% nickel). The average is quoted for point C because there are two phases present, solid and liquid. To determine the compositions of each of these phases, it is simply necessary to 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 4.8). 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* (Figure 4.8). The fraction of solid phase x_s, is given by:

$$x_s = \frac{c_0 - c_l}{c_s - c_l}$$

The fraction of liquid phase, x_l, is given by:

$$x_l = \frac{c_s - c_0}{c_s - c_l}$$

In these equations, c_0 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. These compositions are read from the composition axis as described above. Note that if the composition scale is uniform, these amounts can simply be measured as a distance.

4.2.4 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 4.9). 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

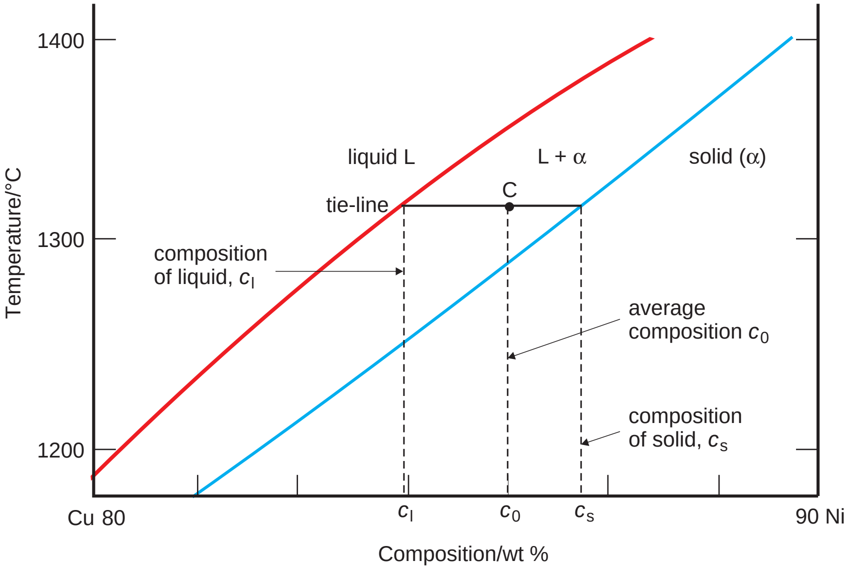


Figure 4.8 Part of the copper–nickel phase diagram (not to scale), demonstrating phase compositions.

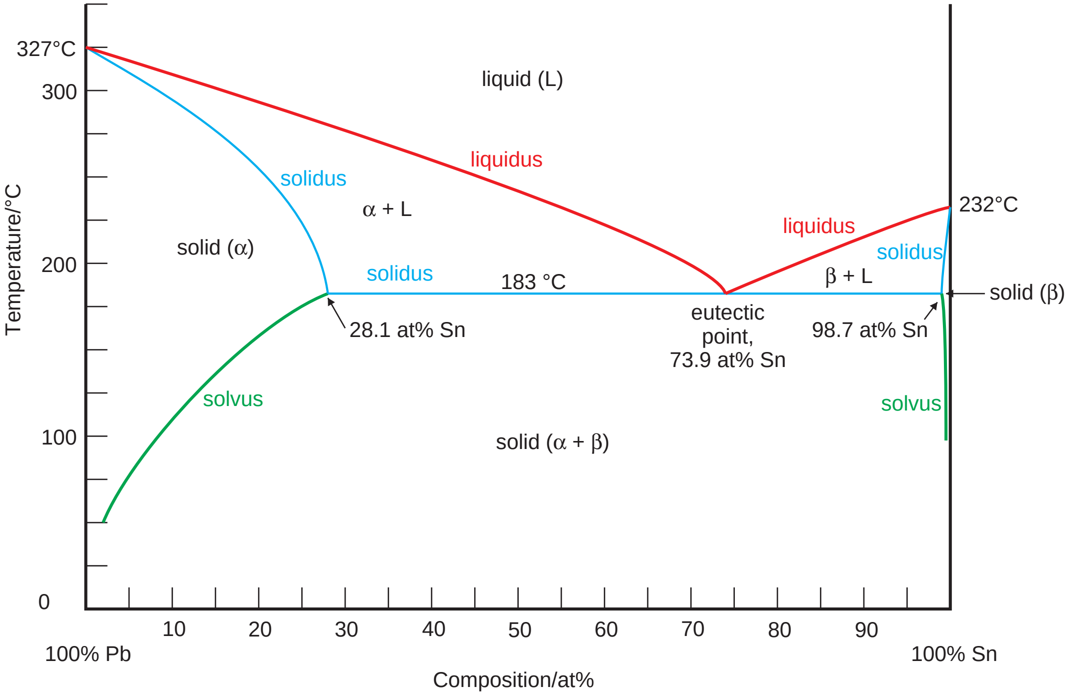


Figure 4.9 The lead (Pb)–tin (Sn) phase diagram at atmospheric pressure.

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 as defects. The β-phase has the same crystal structure as that of tin, and the lead atoms are distributed at random within the crystal as defects. 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. This is because increasing temperature leads to greater atomic vibration, which allows more flexibility in the accommodation of the foreign atoms.

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 in Figure 4.10 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, 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 4.11) is 40 at.% Sn, 60 at.% Pb. The composition of the α-phase is 18 at.% Sn, 82 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

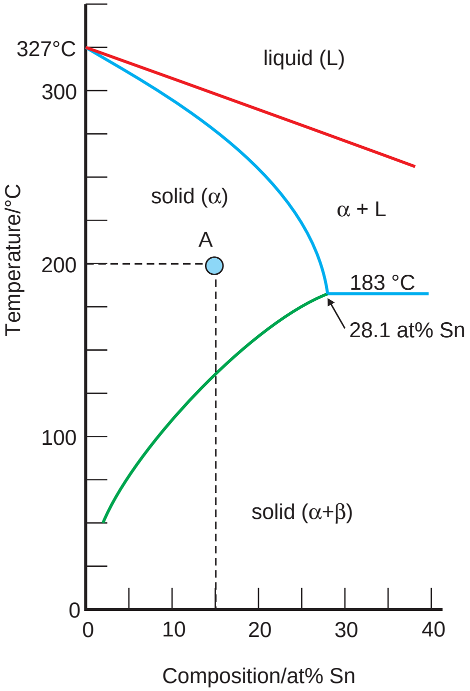


Figure 4.10 The lead-rich region of the lead–tin phase diagram.

of the lever rule, using these compositions. Thus:

$$\text{Amount of } \alpha\text{-phase} = \frac{99 - 40}{99 - 18} = 72.8\%$$

$$\text{Amount of } \beta\text{-phase} = \frac{40 - 18}{99 - 18} = 27.2\%$$

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 coexistence 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 upon cooling through

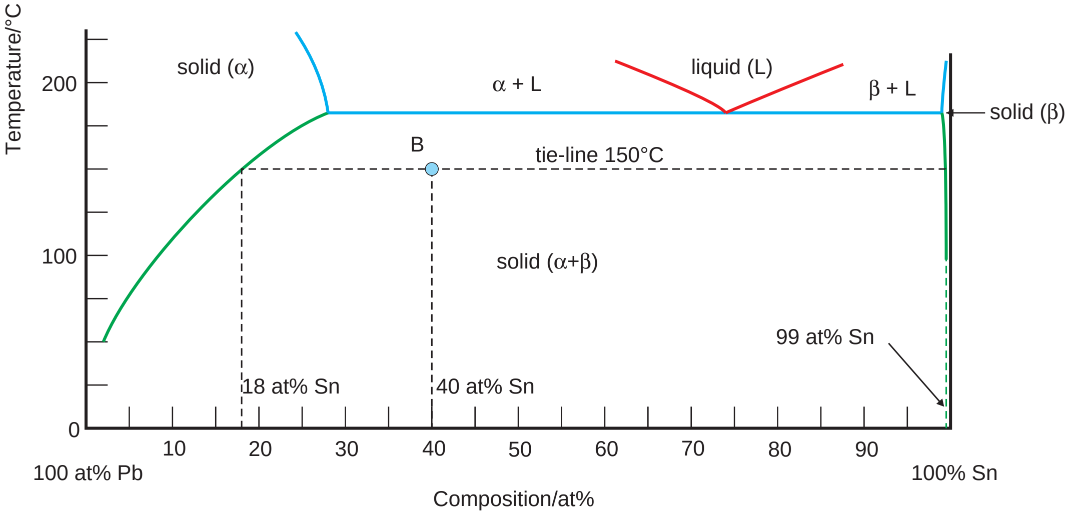


Figure 4.11 The central region of the lead–tin phase diagram.

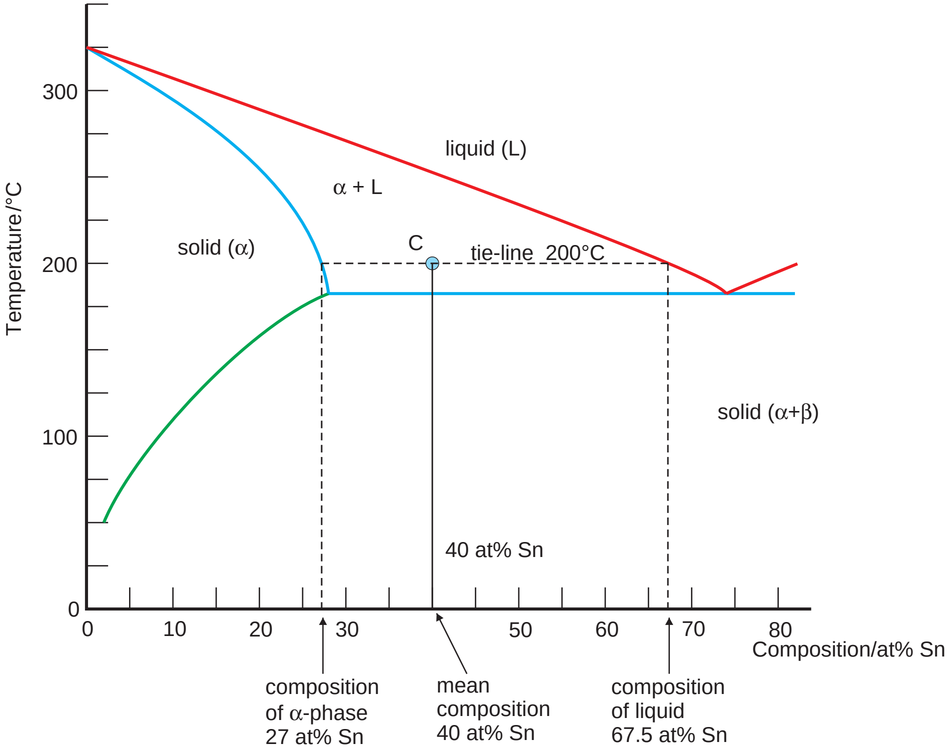


Figure 4.12 The lead-rich region of the lead–tin phase diagram.

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:

$$L \rightarrow \alpha + \beta$$

A cooling curve shows a horizontal break on passing through a eutectic. 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 upon heating.

Solidification over the rest of the phase diagram involves the passage through a two-phase solid plus liquid region. For example, a composition on the Pb-rich side of the eutectic, on passing through the liquidus, will consist of solid α plus liquid. A composition on the Sn-rich side of the eutectic, on passing through the liquidus, will consist of solid β together with liquid. The compositions of the solids are 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 4.12) corresponds to an overall 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 (Section 8.6).

4.2.5 Intermediate phases and melting

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 4.13):

$$CaSiO_3 + CaAl_2O_4 \rightarrow Ca_2Al_2SiO_7$$

None of the phases form solid solutions or show a compositional 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 4.13 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 above can be used to obtain quantitative information in each part of the diagram.

The phase diagram shows that gehlenite melts without any 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 is called a *peritectic* point 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:

$$\text{liquid} + \text{solid } 1 \leftrightharpoons \text{solid } 2$$

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 4.14). 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:

$$MgV_2O_5(s) \rightarrow Mg_2V_2O_7(s) + L(32\,\text{mol\%}\,MgO)$$

As the reaction is reversible, slow-cooling the resultant liquid will regenerate MgV_2O_5.

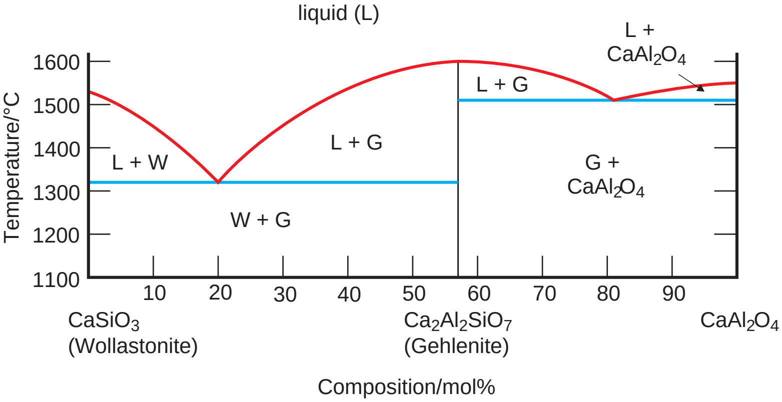


Figure 4.13 The $CaSiO_3$ (wollastonite)–$CaAl_2O_4$ phase diagram showing the intermediate phase $Ca_2Al_2SiO_7$ (gehlenite).

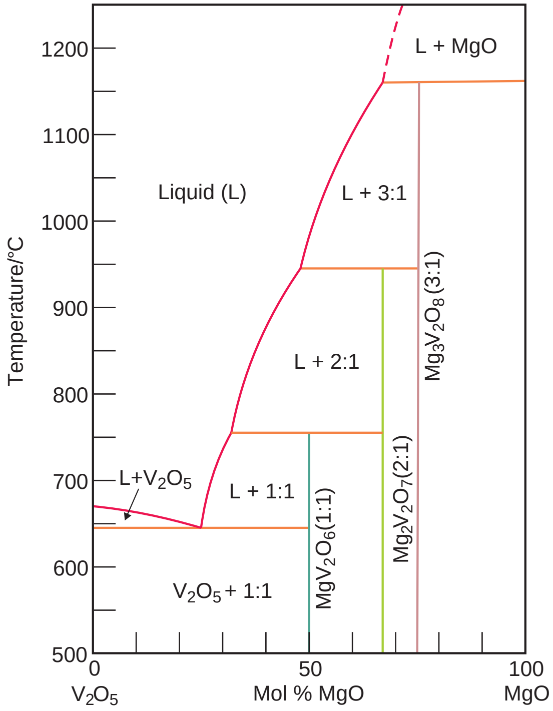


Figure 4.14 The V_2O_5–MgO phase diagram.

4.3 The iron–carbon system near to iron

4.3.1 The iron–carbon phase diagram

The systematic understanding of the iron–carbon phase diagram at the end of the 19th century and the early years of the 20th century was at the heart of the technological advances that characterised these years. This is because steel is an alloy of carbon and iron, and knowledge of the iron–carbon phase diagram allowed metallurgists to fabricate steels of known mechanical properties on demand. Apart from this historical importance, the phase diagram shows a number of interesting features in its own right.

The low-carbon region of the phase diagram is the region relevant to steel production. The version most used is that in which the composition axis is specified in weight % carbon (Figure 4.15). In fact, this is not the *equilibrium* phase diagram of the system. 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.

On the left-hand side of the diagram, the forms of pure iron are indicated, with a melting point of 1538°C. Below the melting point, 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 5.3.4) 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°C and 912°C, was β-iron, but this terminology is no longer in use. Between the temperatures of 912–1394°C the allotrope γ-iron is stable. This phase adopts the face-centred cubic structure (Section 5.3.3). At the highest temperatures, between 1394°C 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 forming interstitial defects. 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 interstitial 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. This material is called *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.

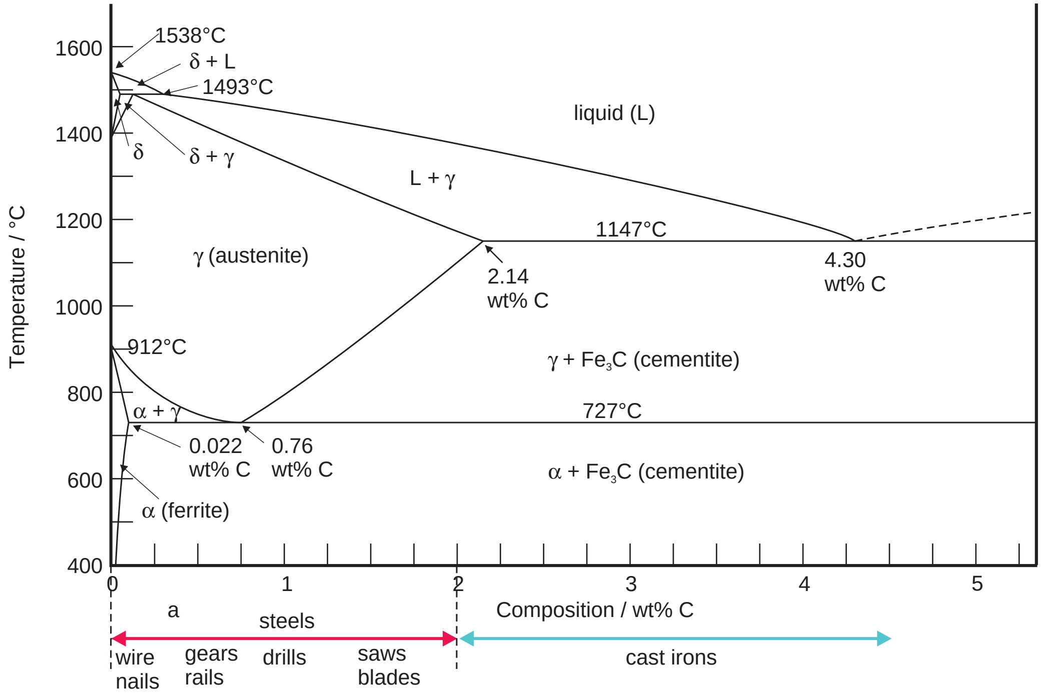


Figure 4.15 The iron-rich region of the iron–carbon metastable existence diagram. The phase cementite (Fe_3C) (not shown) occurs at 6.70 wt% carbon. The α (ferrite) and δ phase fields have been expanded for clarity.

4.3.2 Steels and cast irons

The phase 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% C (Figure 4.15), 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 0.15–0.25 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. Recently, *ultrahigh-carbon steels*, containing between 1 and 2 wt% C, have been studied and found to accept extreme deformation before fracture, called *superelasticity* (Section 10.1.5).

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 cast into the desired shape, and is referred to as *cast iron*. Commercial cast irons rarely contain much more than about 4.5 wt% C.

4.3.3 Invariant points

There are three invariant points in the iron–carbon diagram. A eutectic point is found at 4.30 wt% C and 1148°C. Recall that at a eutectic point a liquid phase transforms to two solids on cooling:

$$L \rightarrow \gamma + Fe_3C$$

A *eutectoid point* occurs at the lowest temperature of the austenite phase field, 727°C, and at 0.76 wt% C. At a eutectoid point, a *solid* transforms to two solids on cooling:

$$\gamma \rightarrow \alpha + Fe_3C$$

This eutectoid transformation is of great importance in steel-making (Section 8.6). The phase diagram also contains a *peritectic point*, at the highest temperature of the austenite phase region, 1493°C. Recall that at a peritectic point, a liquid plus a solid transform into a different solid on cooling.

$$\delta + L \rightarrow \gamma$$

4.4 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 4.16a). Each of the three faces of the prism is the binary phase diagram A–B, A–C or B–C.

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 4.16b). The compositions in isothermal sections are most easily plotted using triangular graph paper. The composition of a point on one of the edges is read directly from the diagram. For example, point D (Figure 4.16b) 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 below.

The composition of an internal point such as E is also found from the compositional 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 4.17 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 $W_{18}O_{49}$.

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 4.18 are given by:

$$\text{Amount of } W_{18}O_{49} = \frac{a}{a+b}$$

$$\text{Amount of } ZrO_2 = \frac{b}{a+b}$$

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 4.18). The

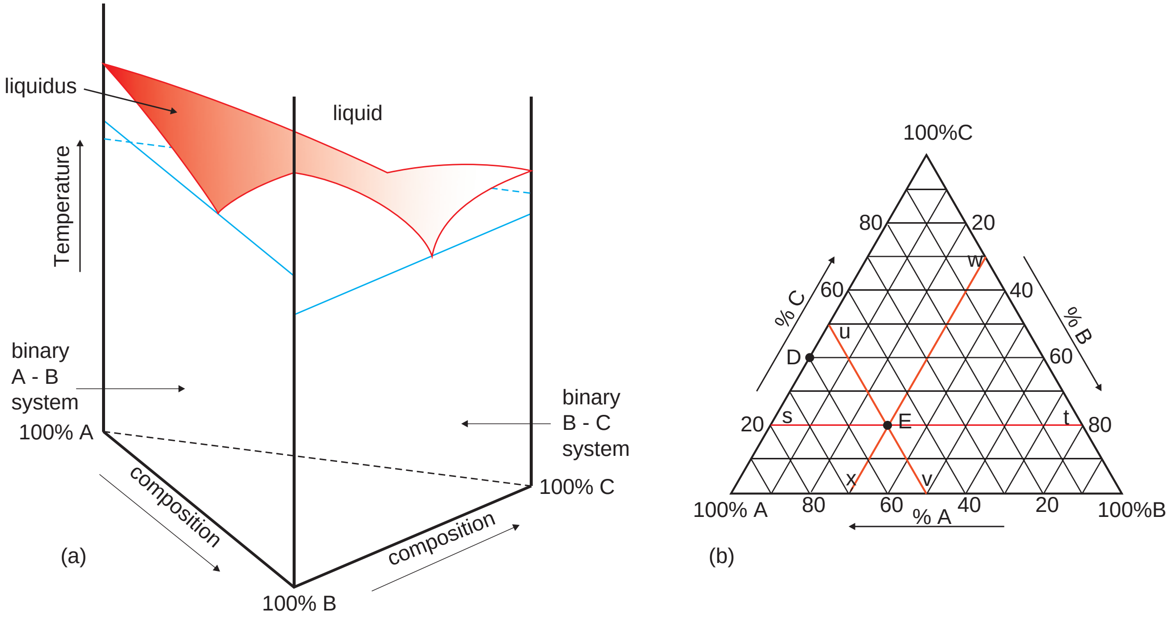


Figure 4.16 (a) The general form of an isobaric ternary phase diagram. (b) Representations of compositions on an isothermal three-component phase diagram section through (a).

amounts of the phases are then given by the lengths of these lines. For example:

$$\text{Amount of } W_{18}O_{49} = \frac{\text{Distance e to G}}{\text{Distance e to } W_{18}O_{49}}$$

$$\text{Amount of } ZrO_2 = \frac{\text{Distance d to G}}{\text{Distance d to } ZrO_2}$$

$$\text{Amount of } WO_2 = \frac{\text{Distance c to G}}{\text{Distance c to } WO_2}$$

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

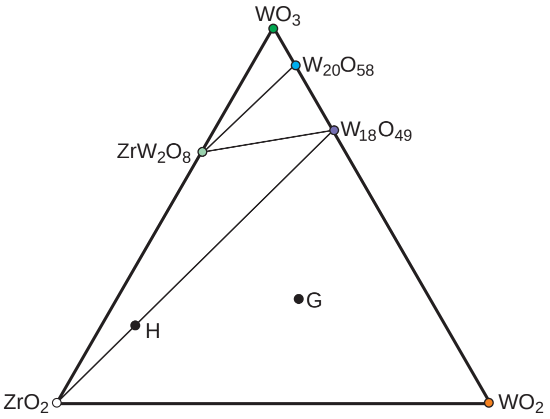


Figure 4.17 The simplified WO_3–WO_2–ZrO_2 phase diagram.

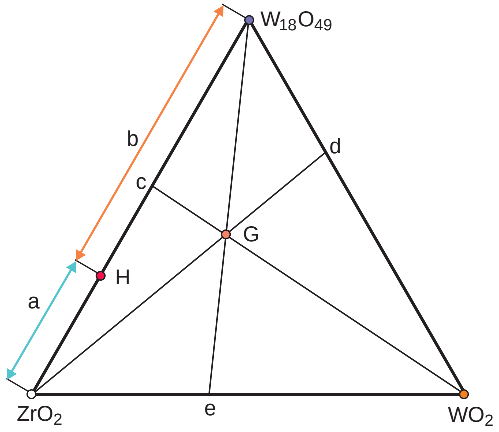


Figure 4.18 The method of determination of compositions on an isothermal phase diagram.

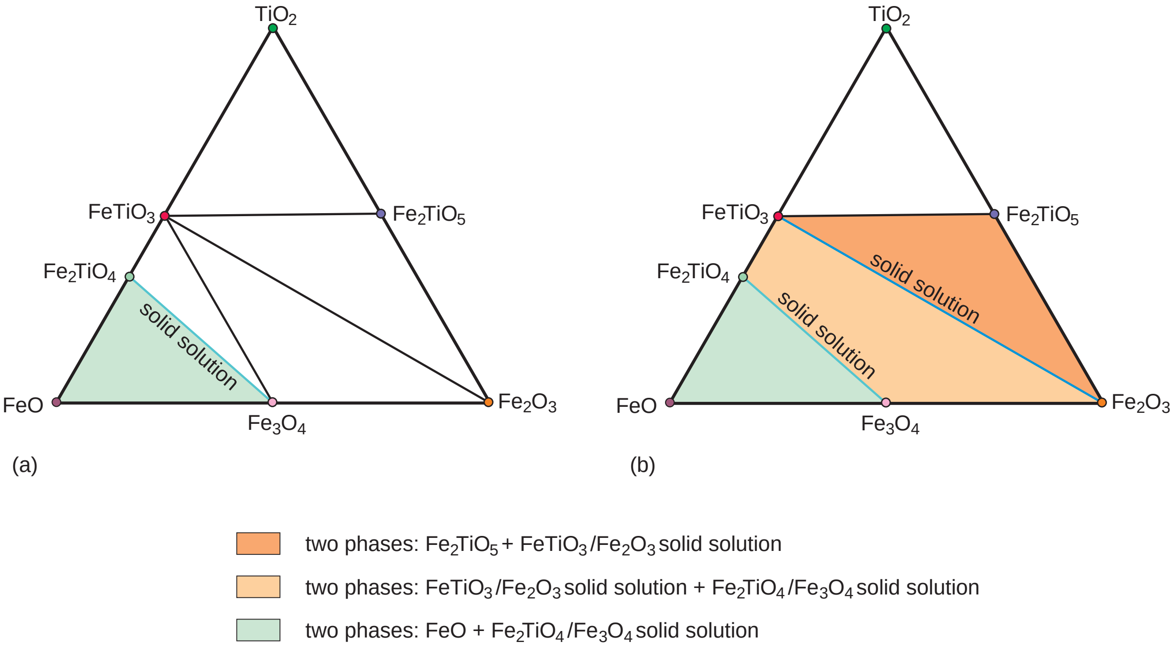

TiO_2
$FeTiO_3$
Fe_2TiO_5
Fe_2TiO_4
solid solution
FeO
Fe_3O_4
Fe_2O_3
(a)
TiO_2
$FeTiO_3$
Fe_2TiO_5
Fe_2TiO_4
solid solution
solid solution
FeO
Fe_3O_4
Fe_2O_3
(b)
two phases: Fe_2TiO_5 + $FeTiO_3/Fe_2O_3$ solid solution
two phases: $FeTiO_3/Fe_2O_3$ solid solution + Fe_2TiO_4/Fe_3O_4 solid solution
two phases: FeO + Fe_2TiO_4/Fe_3O_4 solid solution

Figure 4.19 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 contain three phases.

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 compositional ranges, which may be due to solid solution formation or non-stoichiometry. In these cases the appearance of the 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 adopt the spinel structure. The phase triangle limited by FeO–Fe_3O_4–Fe_2TiO_4 will then only contain two phases, FeO and spinel solid solution, while all other phase triangles will contain three phases (Figure 4.19a). However at temperatures close to 700°C the phase $FeTiO_3$ forms a complete solid solution with Fe_2O_3; both adopt the corundum (Al_2O_3) structure. In this case all of the phase triangles below the $FeTiO_3$–Fe_2TiO_5 tie line contain just two phases (Figure 4.19b).

Ceramic bodies are typically multiphase materials and phase diagrams can indicate the likely phase make-up of these solids. The MgO–Al_2O_3–SiO_2 system contains the important industrial cordierite and steatite ceramics (Figure 4.20). Although these

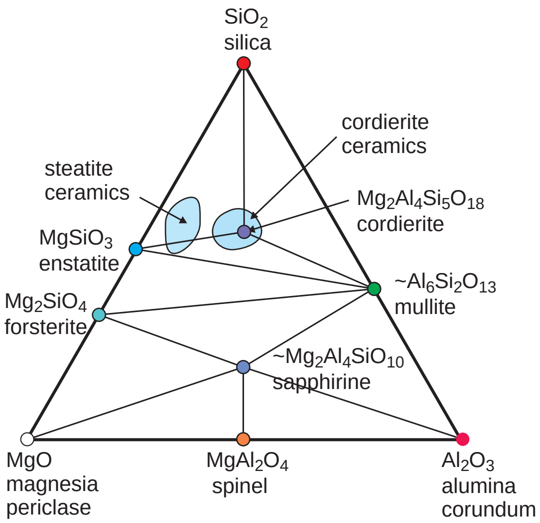


Figure 4.20 Approximate phase diagram for the MgO–Al_2O_3–SiO_2 system, containing the important ceramic materials cordierite and steatite.

materials exist over considerable compositional 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 have compositional ranges that have been denoted by representative points in the figure.)

4.5 Calculation of phase diagrams: CALPHAD

The principle of calculating phase diagrams that lies behind the CALculation of PHAse Diagrams (CALPHAD) method uses basic thermodynamics. That is, the most stable phase or collection of phases, when a system is at equilibrium, is that which possesses the lowest Gibbs energy.

The Gibbs energy, G, of a stoichiometric line phase is a function of the temperature, pressure, magnetic state and electronic state. When the phase has a compositional range, as with solid solutions and non-stoichiometric materials, the composition must also be included. To carry out the calculations the Gibbs energy is usually written as a series involving all or some of these variables. The simplest case, the temperature variation of a single stoichiometric phase, is usually written in terms of the series:

$$G^{\phi} = \mathrm{a} + bT + cT\ln T + dT^2 + eT^3 + f/T \ldots \tag{4.4}$$

where a, b, c . . . are empirical parameters. Broadly speaking, the Gibbs energy of a phase in a multicomponent system is given by the sum of three terms:

$$G^{\phi} = G^0 + G^{\mathrm{ideal}} + G^{\mathrm{xs}} \tag{4.5}$$

where G^0 is the free energy arising from mechanical mixing of the pure components, G^{ideal} is the Gibbs energy arising when the components form an ideal solution, and G^{xs} is the excess Gibbs energy arising from other factors, such as ionic interactions, defects or order–disorder transformations.

The simplest case is for a stoichiometric binary phase, when the Gibbs energy is given by an expression of the type:

$$G^{\phi} = \mathrm{x}_{\mathrm{A}}G_{\mathrm{A}}^{\,0} + \mathrm{x}_{\mathrm{B}}G_{\mathrm{B}}^{\,0} + \Delta G_{\mathrm{f}}$$

where x_{A} is the mole fraction of component A, $G_{\mathrm{A}}^{\,0}$ is the standard Gibbs energy of pure component A, and similarly for component B, and ΔG_{f} is the energy of formation of the phase. The first two of these terms, added together, are equivalent to G^0 in equation (4.5). Because the phase is stoichiometric there is no necessity to include a solution term G^{ideal}, and the excess Gibbs energy G^{xs} is, in this case, ΔG_{f}. Each of these Gibbs energy symbols must, of course, be represented by the appropriate series (equation 4.4) for the purposes of computation.

If the phase exhibits a compositional range that is considered to be an ideal solution of one component in the other, the term G^{ideal}, given by $RT(x_{\mathrm{A}} \ln x_{\mathrm{A}} + x_{\mathrm{B}} \ln x_{\mathrm{B}})$, is included as well as an appropriate expression for G^{xs}:

$$G^{\phi} = x_{\mathrm{A}}G_{\mathrm{A}}^{\,0} + x_{\mathrm{B}}G_{\mathrm{B}}^{\,0} + RT(x_{\mathrm{A}}\ln x_{\mathrm{A}} + x_{\mathrm{B}}\ln x_{\mathrm{B}}) + \Delta G^{\mathrm{xs}}$$

where R is the gas constant and T the absolute temperature. The excess Gibbs energy term takes into account the interactions between the components in the solution over and above those pertaining to the ideal solution model. This is usually written as a series – the *Redlich-Kister* formula:

$$\Delta G^{\mathrm{xs}} = x_{\mathrm{A}}x_{\mathrm{B}}\left[{}^0L + {}^1L(x_{\mathrm{A}} - x_{\mathrm{B}}) + {}^2L(x_{\mathrm{A}} - x_{\mathrm{B}})^2 + {}^3L(x_{\mathrm{A}} - x_{\mathrm{B}})^3 + \ldots\right]$$

$$\Delta G^{\mathrm{xs}} = x_{\mathrm{A}}x_{\mathrm{B}}\textstyle\sum_{j=0}^{n} {}^jL(x_{\mathrm{A}} - x_{\mathrm{B}})^j$$

where ${}^jL$ is the *binary interaction parameter*, generally expressed as a series similar to that given above for $G\phi$:

$${}^jL = \mathrm{a} + bT + cT\ln T + dT^2 + eT^3 + f/T \ldots$$

Usually only the first one or two terms are needed.

Similar but more complex expressions are written if the phases contain defects, or if other interactions (ionic, magnetic, order–disorder and so on) are important. These aspects are often treated by subdividing the solid into sublattices and handling each sublattice separately.

Having set up the framework of equations for the Gibbs energy of the system, the various coefficients a, b . . . need to be evaluated – a process known as *assessment* or *optimisation*. Assessment relies upon analysis of experimental data using mathematical techniques and extrapolation to obtain optimum values. The larger the data-set that is available, the more accurately the parameters can be evaluated. In addition, it is necessary to have adequate computational power and sophisticated algorithms to be able to calculate the many data points needed to construct even a simple binary phase diagram. The computational problems are considerably more severe for ternary and higher systems.

Serious work on the computation of phase diagrams began just after the middle of the last century. Progress was limited, though, because of the lack of both data and computing power. This is no longer so, and a variety of computer software packages are available, such as ChemSage, MTDATA, Thermo-Calc and PANDAT, which are able to process sophisticated phase diagram simulations. Indeed, present studies go far beyond the presentation of phase diagrams and include a wide range of thermodynamic computations. The addition of kinetic data also allows diffusion profiles, solidification and precipitation reactions to be simulated successfully (Chapters 7 and 8)

Further Reading

Ehlers, E.H. (1972) *The Interpretation of Geological Phase Diagrams*. W.H. Freeman, San Francisco.

Lukas, H.L., Fries, S.G. and Sundman, B. (2007) *Computational Thermodynamics, the Calphad Method*. Cambridge University Press, Cambridge.

Massalski, T.B. (ed.) (1990) *Binary Alloy Phase Diagrams*, 2nd edn, Vols **1–3**. ASM International, Materials Park, OH.

Phase Diagrams for Ceramicists, Vol 1. (1964) to Vol 10, (1994), a continuing series, with changing editors. American Ceramic Society, Westerville, OH.

Villars, P., Prince, A. and Okamoto, H. (1995) *Ternary Phase Diagrams*, Vols **1–10**. ASM International, Materials Park, OH.

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

1 A phase is:
(a) A compound in a system.
(b) An element in a system.
(c) A homogeneous part of a system.

2 The phase diagram of a one-component system is described in terms of:
(a) One variable.
(b) Two variables.
(c) Three variables.

3 How many phases coexist at a triple point in the iron phase diagram?
(a) One.
(b) Two.
(c) Three.

4 On a phase boundary in the sulphur system:
(a) One phase exists.
(b) Two phases exist.
(c) Three phases exist.

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.

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.

7 The number of variables needed to specify phase relations in a binary system are:
(a) Two.
(b) Three.
(c) Four.

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.

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.

10 A tie line is drawn:
(a) Parallel to the pressure axis.
(b) Parallel to the temperature axis.
(c) Parallel to the composition axis.

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.

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.

13 A eutectic point on a binary phase diagram is:
(a) An invariant point.
(b) A triple point.
(c) A critical point.

14 On cooling a homogeneous liquid sample through a eutectic point:
(a) A liquid plus a solid form.
(b) Two solid phases form.
(c) A homogeneous solid forms.

15 A line phase on a binary phase diagram is a phase that:
(a) Exists along a phase boundary.
(b) Has a compositional range along a tie line.
(c) Has no apparent compositional range.

16 Solid phases that melt congruently will 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.

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.

18 Pure iron has:
(a) Two allotropes.
(b) Three allotropes.
(c) Four allotropes.

19 Steel is an alloy of iron and carbon in which the carbon occupies:
(a) Substitutional sites.
(b) Interstitial sites.
(c) Vacancies.

20 Ferrite is:
(a) An allotrope of iron
(b) An intermediate phase.
(c) An iron–carbon alloy.

21 Austenite has:
(a) No appreciable compositional range.
(b) A narrow compositional range compared with ferrite.
(c) A wide compositional range compared with ferrite.

22 Austenite has the same crystal structure as:
(a) α-iron.
(b) β-iron.
(c) γ-iron.

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.

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.

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.

26 A ternary system is one in which there are:
(a) Three components.
(b) Three variables.
(c) Three phases present.

27 To represent all possible phase relations a ternary system needs:
(a) Five axes.
(b) Four axes.
(c) Three axes.

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

4.1 A copper–nickel alloy is made up with 28 grams of copper in 100 g alloy. What is the at.% Ni in the alloy?

4.2 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 grams of sample?

4.3 A copper–zinc alloy is made up with 35 g Cu in 100 g alloy (35 wt%). What is the at.% Zn in the alloy?

4.4 A solder contains 50 wt% Sn and 50 wt% Pb. What are the atomic percentages of Sn and Pb in the solder?

4.5 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?

4.6 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 (Figures 4.6 and 4.8):

(a) What is the at.% 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?

4.7 For the sample in the previous question, the alloy is held at a temperature of 1340°C.

(a) What phase(s) are present?
(b) How much solid is present, if any?
(c) How much liquid is present, if any?

4.8 With reference to the Cu–Ni phase diagram (Figures 4.6 and 4.8), 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% that of the solid, calculate the vol.% of the solid and liquid present.

4.9 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 (Figure 4.21).

(a) What is the wt% 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?

4.10 With reference to Figure 4.21, 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?

4.11 With reference to Figure 4.21, a sample of composition 30 mol% Cr_2O_3 is held at 2080°C.

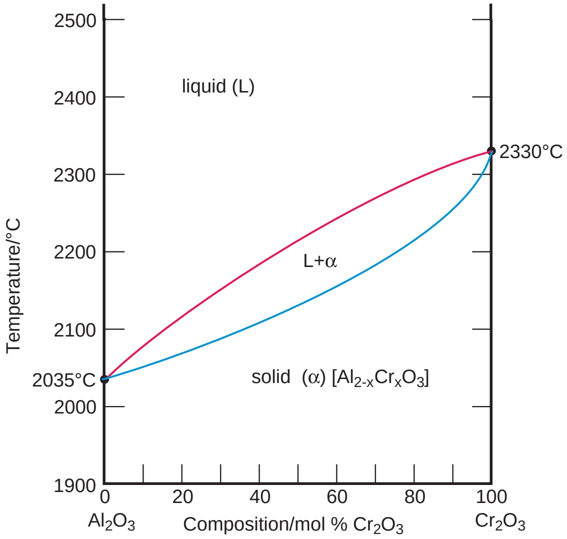


Figure 4.21 The Al_2O_3–Cr_2O_3 phase diagram.

(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?

4.12 With reference to Figure 4.21, 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?

4.13 The phase diagram of the ruthenium (Ru)–rhenium (Re) system is given in Figure 4.22.

(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 grams 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?

4.14 With reference to Figure 4.22, 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?

4.15 The phase diagram of the BeO–Y_2O_3 system is given in Figure 4.23. Explain why this figure differs from that of the lead–tin system given in Figure 4.9. Why does the composition axis use $^1/_2[Y_2O_3]$ instead of Y_2O_3?

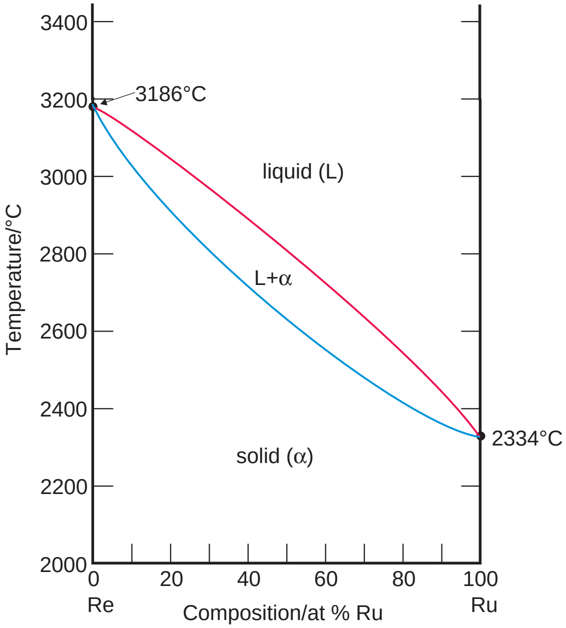


Figure 4.22 The Re–Ru phase diagram.

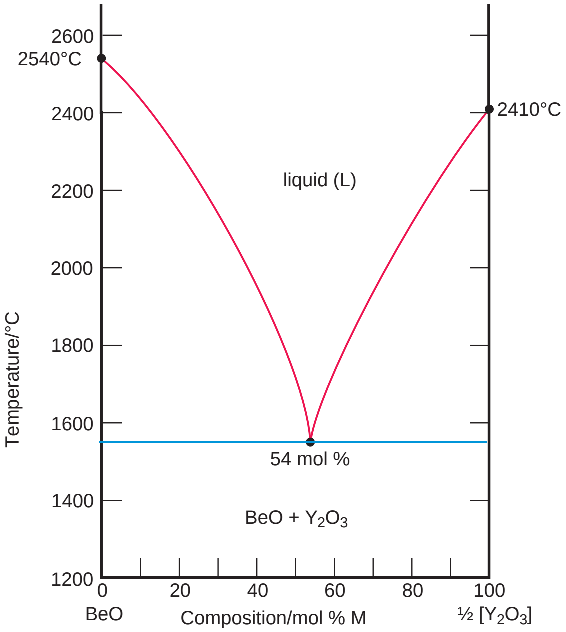


Figure 4.23 The BeO–Y_2O_3 phase diagram.

4.16 With respect to Figure 4.23, 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?

4.17 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?

4.18 With respect to the Pb–Sn phase diagram (Figures 4.9–4.12), 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) 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?

4.19 With respect to the Pb–Sn phase diagram (Figures 4.9–4.12), 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?

4.20 With respect to the Pb–Sn phase diagram (Figures 4.9–4.12), 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?

4.21 With respect to the Fe–C phase diagram (Figure 4.15), an alloy with a composition of 1.5 wt% C is homogenised by heating for a long period at 1000°C.

(a) What phase(s) are present?

(b) How much of each phase is present?

(c) What are the compositions of the phase(s)?

4.22 With respect to the Fe–C phase diagram (Figure 4.15), an alloy with a composition of 1.5 wt% C, homogenised by heating for a long period at 1000°C, is subsequently cooled slowly to 800°C.

(a) What phase(s) are present?

(b) How much of each phase is present?

(c) What are the compositions of the phase(s)?

4.23 With respect to the Fe–C phase diagram (Figure 4.15), 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?

4.24 With respect to the phase diagram of the WO_3–WO_2–ZrO_2 system (Figure 4.17), 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?

4.25 With respect to the phase diagram of the WO_3–WO_2–ZrO_2 system (Figure 4.17), 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?

4.26 According to the SiO_2–MgO–Al_2O_3 phase diagram (Figure 4.20), what phases may be present in steatite ceramics and cordierite ceramics?

5

Crystallography and crystal structures

- How does a lattice differ from a crystal structure?
- What is a unit cell?
- How are crystal structures determined?

Crystallography describes the ways in which the component atoms are arranged in crystals. Many chemical and physical properties depend upon crystal structure, and knowledge of crystallography is essential if the properties of materials are to be understood.

In earlier centuries, crystallography developed via two independent routes. The first of these was observational. It was long supposed that the beautiful shapes of mineral crystals were an expression of internal order, and this order was described by the classification of external shapes, the morphology or *habit* of crystals. The regularity of crystals, together with 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. A second route, the mathematical description of the arrangement of objects in space, was developed in the latter years of the 19th century. Both of these play a part in helping us to understand crystals and their properties. The two approaches were unified with the exploitation of X-ray and other diffraction methods, which are now used to determine crystal structures on a routine basis.

5.1 Crystallography

5.1.1 Crystal lattices

Crystal structures and crystal lattices are different, although these terms are frequently (and incorrectly) 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 edges and the angles between them are

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

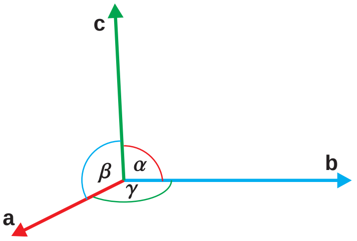


Figure 5.1 The reference axes used to characterise the crystal systems.

collectively called the *lattice parameters* or (*unit*) *cell parameters*.

Clearly there are any number of ways of choosing **a**, **b** and **c** and the unit cell. In order to avoid ambiguity when describing lattices and crystal structures, 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 5.1). 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**. Just seven different unit cell types (arrangements of axes) are needed in order to specify all three-dimensional lattices (Table 5.1).

Table 5.1 The crystal systems

System	Unit cell parameters
Cubic (isomorphic)	$a = b = c$; $\alpha = 90°$, $\beta = 90°$, $\gamma = 90°$
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*	$a = b = c$; $\alpha = \beta = \gamma \neq 90°$ $a' = b' \neq c'$; $\alpha' = 90°$, $\beta' = 90°$, $\gamma' = 120°$

*Rhombohedral unit cells are often specified in terms of a bigger hexagonal unit cell.

The seven unit cell types give rise to 14 three-dimensional lattices, called *Bravais* lattices (Figure 5.2). 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 8 neighbouring unit cells, and as each unit cell has 8 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.

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.

5.1.2 Crystal systems and crystal structures

Observation suggested that all crystals could be classified into one of seven *crystal systems*, that is, seven unit cell types (Table 5.1). They are identical to the unit cells required to describe the Bravais lattices.

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 just sufficient, when repeated at each of the lattice points, to construct the whole of the structure. A crystal structure is built up from a lattice plus a motif.

The position of an atom within the unit cell is given as fractions *x*, *y*, *z*, of the *a*, *b* and *c* lattice parameters 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 ($^1/_2$, $^1/_2$, $^1/_2$), 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 ($^1/_2$, $^1/_2$, 0), (0, $^1/_2$, $^1/_2$), or ($^1/_2$, 0, $^1/_2$), 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 ($^1/_2$, 0, 0), (0, $^1/_2$, 0) or (0, 0, $^1/_2$). 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 5.3). 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.

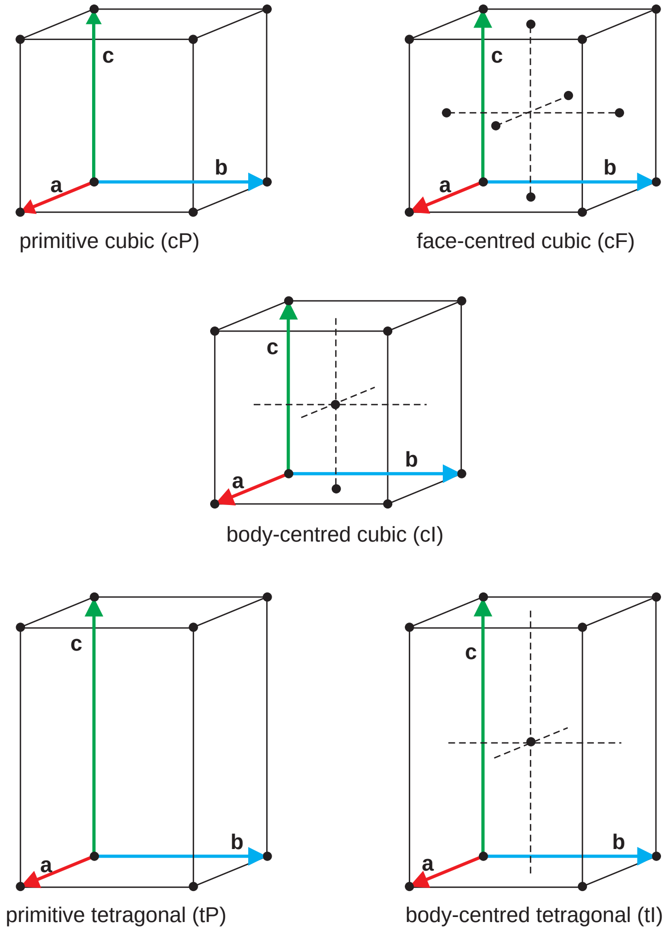


Figure 5.2 The 14 Bravais lattices. Note that the lattice points are *not* atoms. The monoclinic lattices have been drawn with the **b**-axis vertical, to emphasise that it is normal to the plane containing the **a**- and **c**-axes.

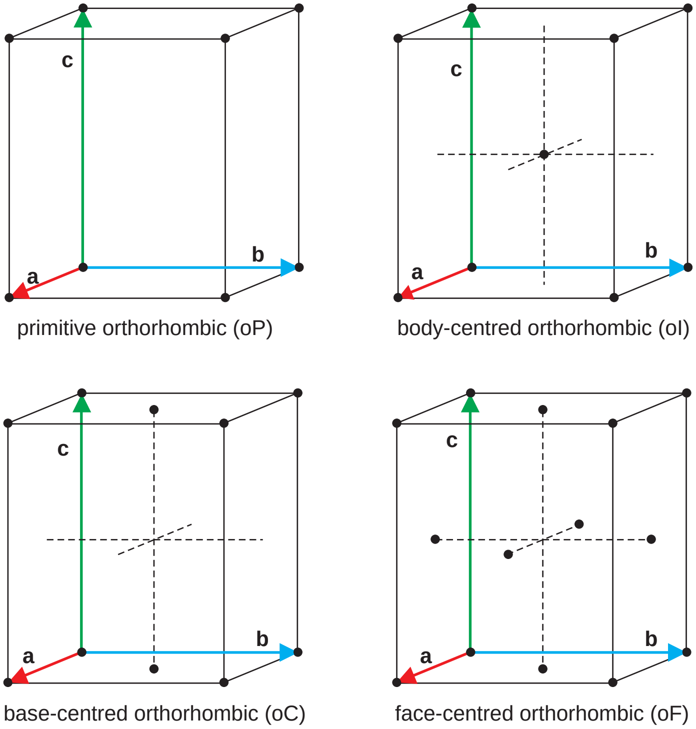

c
b
a
primitive orthorhombic (oP)
c
b
a
body-centred orthorhombic (oI)
c
b
a
base-centred orthorhombic (oC)
c
b
a
face-centred orthorhombic (oF)

Figure 5.2 (*Continued*)

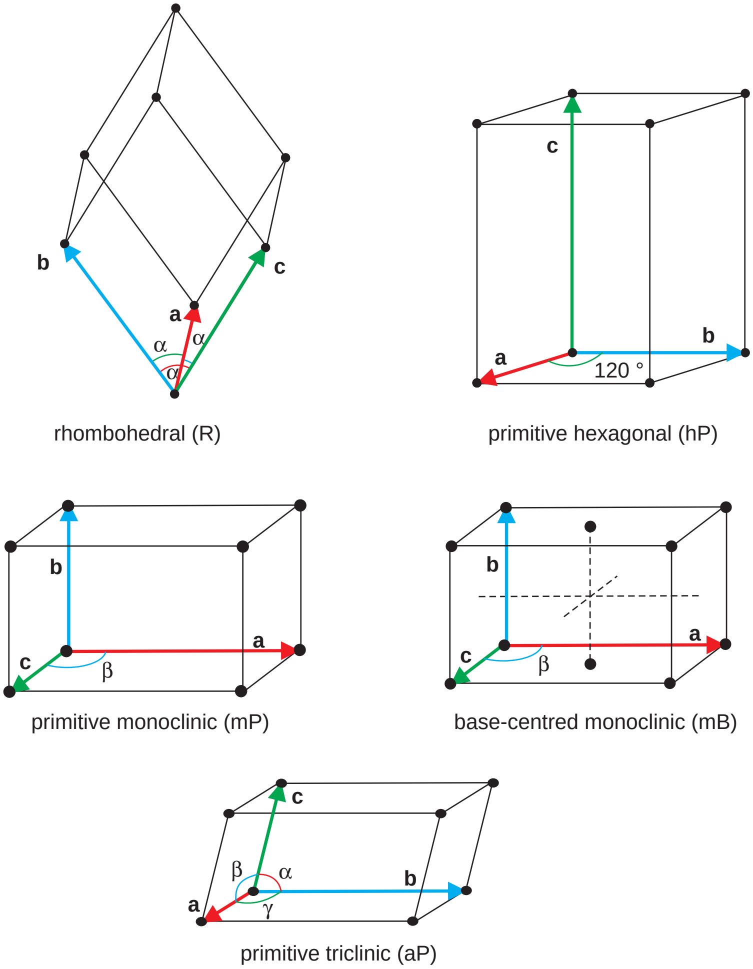
Different compounds that crystallise with the same crystal structure, for example the two alums, $NaAl(SO_4)_2.12H_2O$ and $NaFe(SO_4)_2.12H_2O$, are said to be *isomorphous*[1] or *isostructural*. Sometimes the crystal structure of a compound will change with temperature and with applied pressure. This is called *polymorphism*. Polymorphs of elements are known as *allotropes*. Graphite and diamond are two allotropes of carbon, formed at different temperatures and pressures.

5.1.3 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 is 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 the physical properties that can be observed. For example, the piezoelectric effect (Section 11.2)

[1] This description originally applied to the same external form of the crystals rather than the internal arrangement of the atoms.



rhombohedral (R)

primitive hexagonal (hP)

primitive monoclinic (mP)

base-centred monoclinic (mB)

primitive triclinic (aP)

Figure 5.2 (*Continued*)

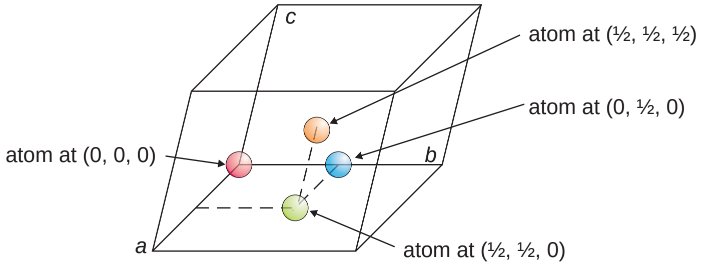


Figure 5.3 The positions of atoms in a unit cell, specified as fractions of the cell edges, not with respect to Cartesian axes.

is only found in crystals that lack a centre of symmetry. A unit cell with 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 a SiO_4 tetrahedron, a group that lacks a centre of symmetry (Figure 5.4). The oxygen atom at $+z$ on the **c**-axis is not paired with a similar oxygen atom at $-z$.

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 placing of atoms within the unit cell. The determination of a crystal structure generally starts with the determination of the correct space group for the sample.

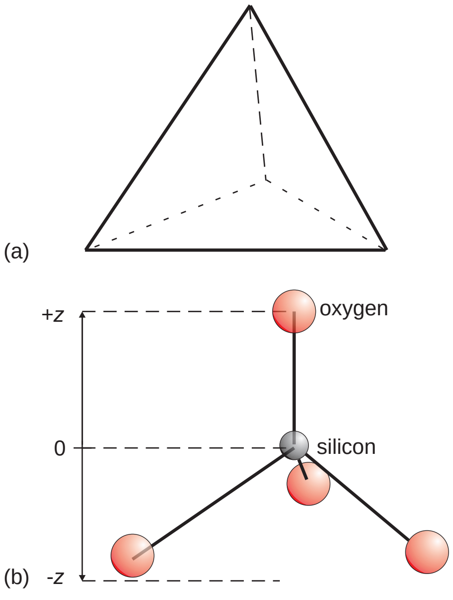


Figure 5.4 (a) An ideal tetrahedron. (b) An ideal tetrahedral (SiO_4) unit. An oxygen atom at $+z$ does not have a counterpart at $-z$, and the unit is not centrosymmetric.

5.1.4 Crystal planes and 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 5.5a). 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 5.5b). Similarly, parallel planes cutting the **a**-edge at a/3 would have Miller indices of (300) (Figure 5.5c). Any *general plane* parallel to (100) is written (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 the c cell-edge has indices (00l) (Figure 5.5d,e).

Planes that cut two edges and lie parallel to a third are described by indices (hk0), (0kl) or (h0l) (Figure 5.6). 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). As well as the (110) plane, a similar plane also cuts the

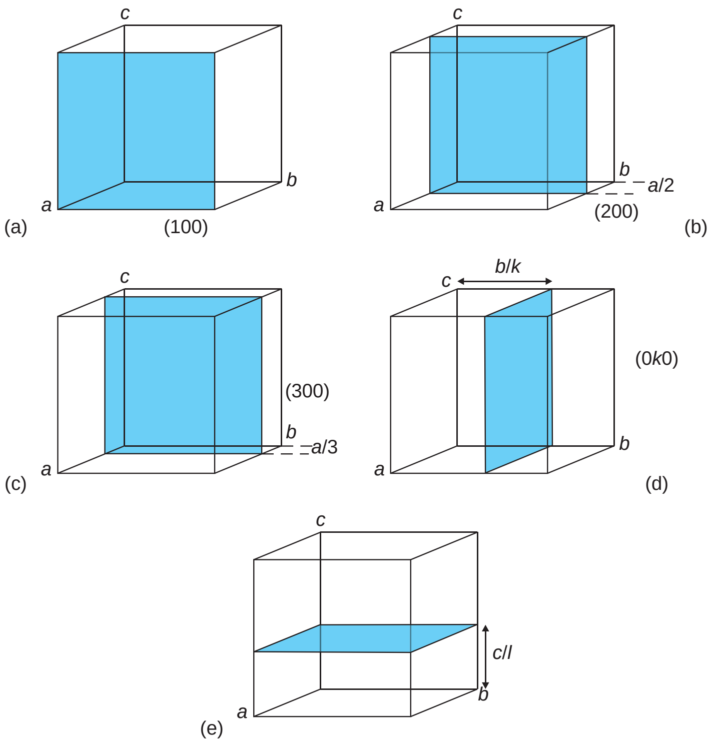


Figure 5.5 Miller indices of crystal planes: (a) (100); (b) (200); (c) (300); (d) (0*k*0); (e) (00*l*).

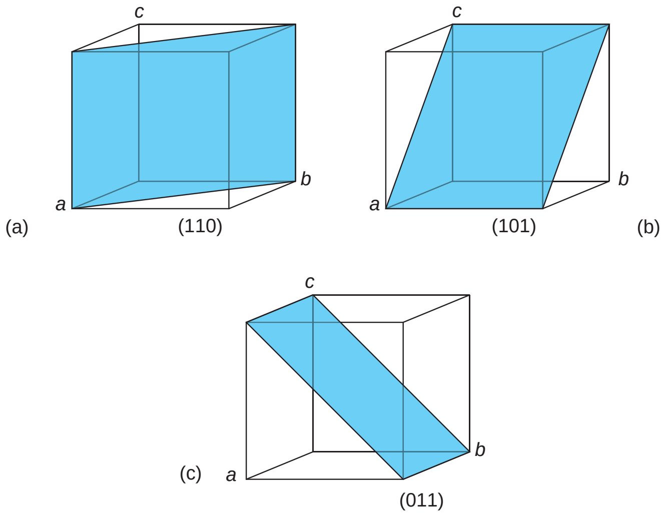


Figure 5.6 Miller indices of crystal planes: (a) (110); (b) (101); (c) (011).

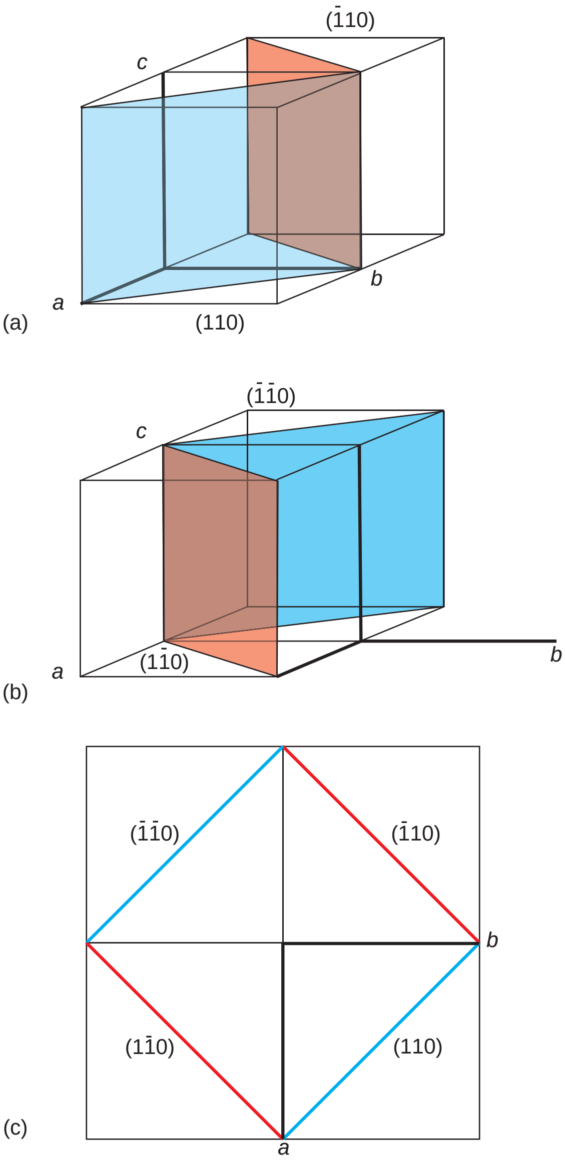


Figure 5.7 Miller indices of crystal planes: (a) (110) and $(\bar{1}10)$; (b) $(1\bar{1}0)$ and $(\bar{1}\bar{1}0)$; (c) projection down the **c**-axis, showing all four equivalent {110} planes.

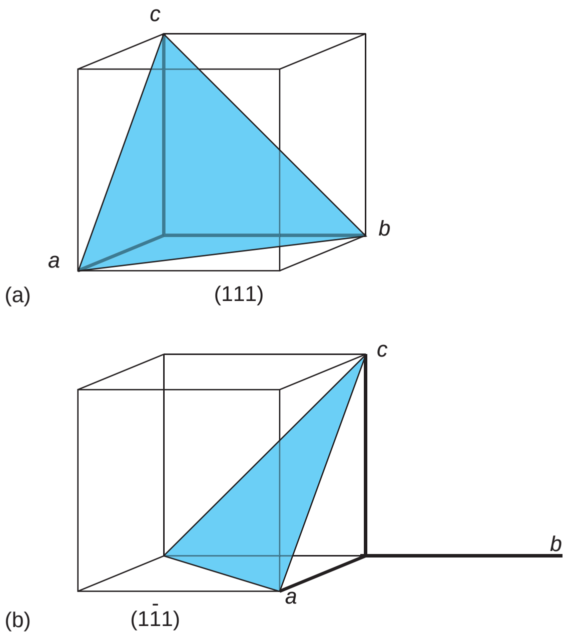


Figure 5.8 Miller indices of crystal planes: (a) (111); (b) $(1\bar{1}1)$.

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 5.7a). 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 5.7b). Because the Miller indices (*hkl*) refer to a set of planes $(\bar{1}\bar{1}0)$ is equivalent to (110), as the position of the axes is arbitrary. Similarly, the plane with Miller indices $(1\bar{1}0)$ is equivalent to $(\bar{1}10)$ (Figure 5.7c). This notation is readily extended to cases where a plane cuts all three unit cell edges (Figure 5.8).

In order to find the Miller indices of a plane, first draw the other members of the set, if not already present (Figure 5.9a,c). This is done by drawing parallel planes through each lattice point (Figure 5.9b,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 is the −**a** direction this gives index $\bar{h}$. Repeat this along **b** to get *k* and along **c** to get *l*.

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

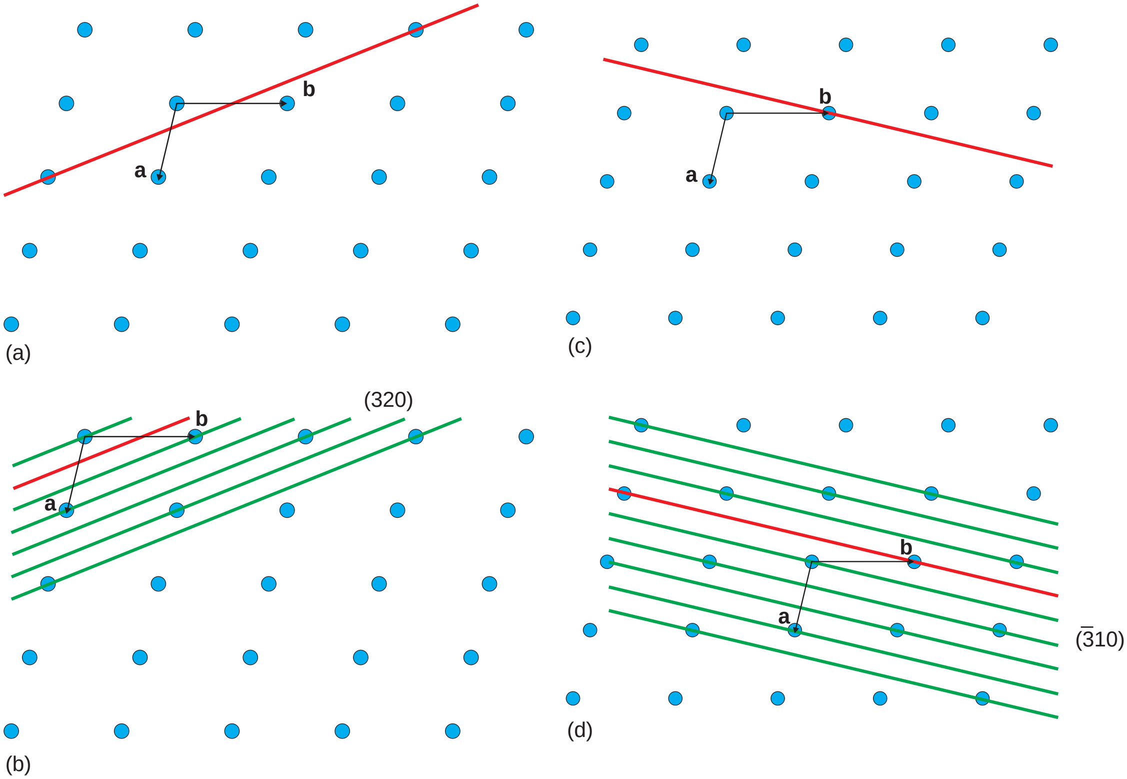


Figure 5.9 Determination of Miller indices. (a, c) For the plane shown, draw parallel planes through all the lattice points. (b, d) Count the spaces between the planes in a distance of $1a$ and $1b$. Spaces counted in a negative direction are given negative indices.

(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).

5.1.5 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 5.10 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}$0) and ($\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:

$$h + k + i = 0, \text{ that is, } i = -(h + k)$$

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),

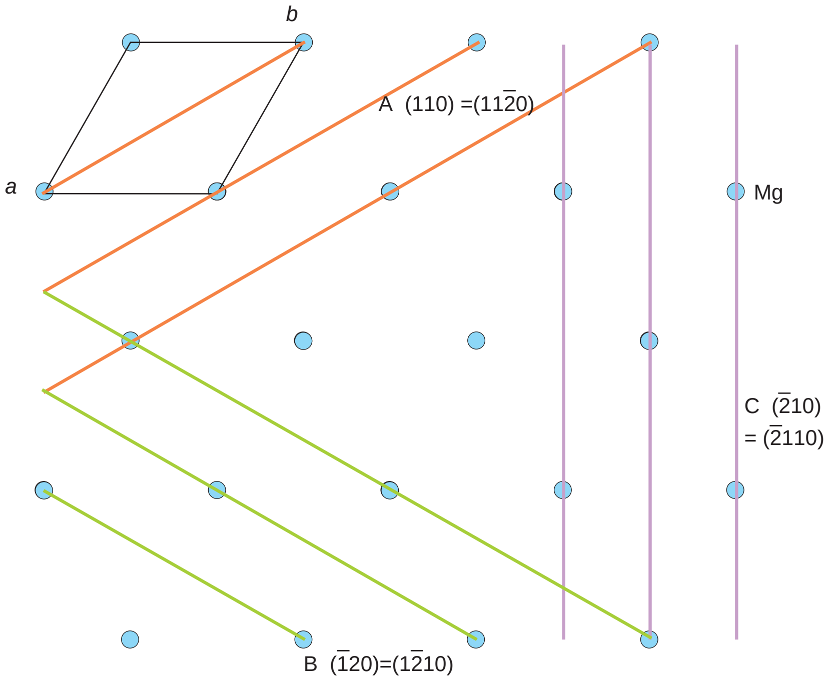


Figure 5.10 Miller indices in hexagonal crystals.

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 (*hk.l*). This nomenclature emphasises that the hexagonal system is under discussion without actually including a value for *i*.

5.1.6 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 5.11a). For example, the direction [100] is parallel to **a**, the direction [010] is parallel to **b**, and [001] is parallel to **c**. Because directions are vectors, [*uvw*] is not identical to [$\bar{u}\,\bar{v}\,\bar{w}$], 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 5.11b). A North wind is always a North wind, regardless of where you stand.

As with 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}$0], [001] and [00$\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

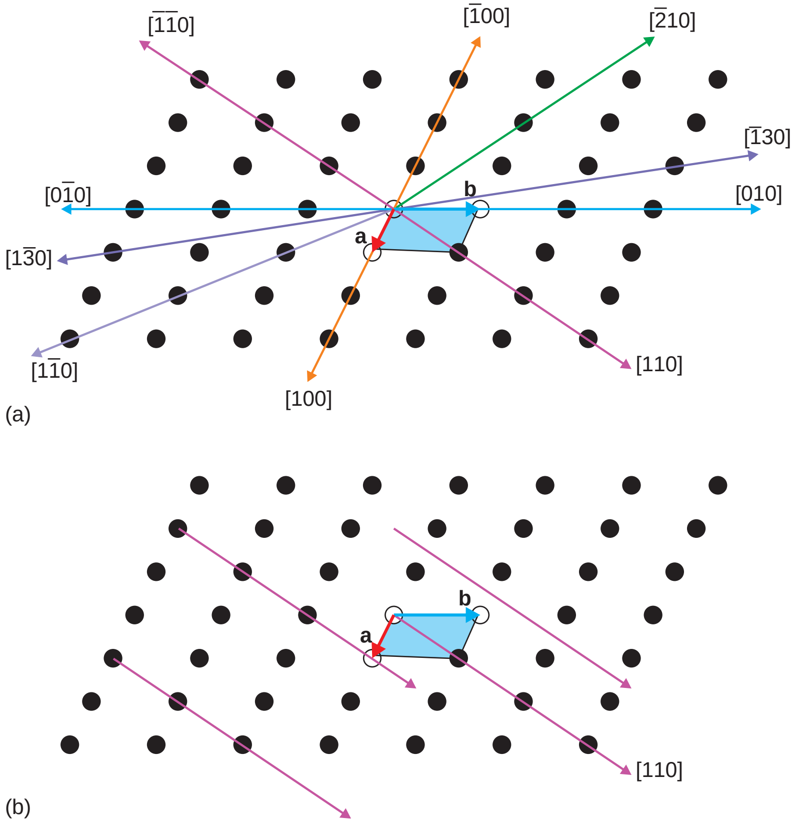


Figure 5.11 (a) Directions in a lattice. (b) Parallel directions all have the same indices.

(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 above, 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 $2a[100]$.

As with Miller indices, directions in hexagonal crystals are sometimes specified by a four-index system, [u' v' t w'], called *Weber indices*. The conversion of a three-index set to a four-index set is given by:

$$[uvw] \longrightarrow [u'\,v'\,t\,w']$$

$$u' = n/3(2u - v)$$

$$v' = n/3(2v - u)$$

$$t = -(u' + v')$$

$$w' = nw$$

In these equations, n is a factor *sometimes* needed to make the new indices into smallest integers. Thus directions [$00l$] always transform to [$000l$]. The three equivalent directions in the basal (0001) plane of a hexagonal crystal structure such as magnesium (Figure 5.12) are obtained by using the

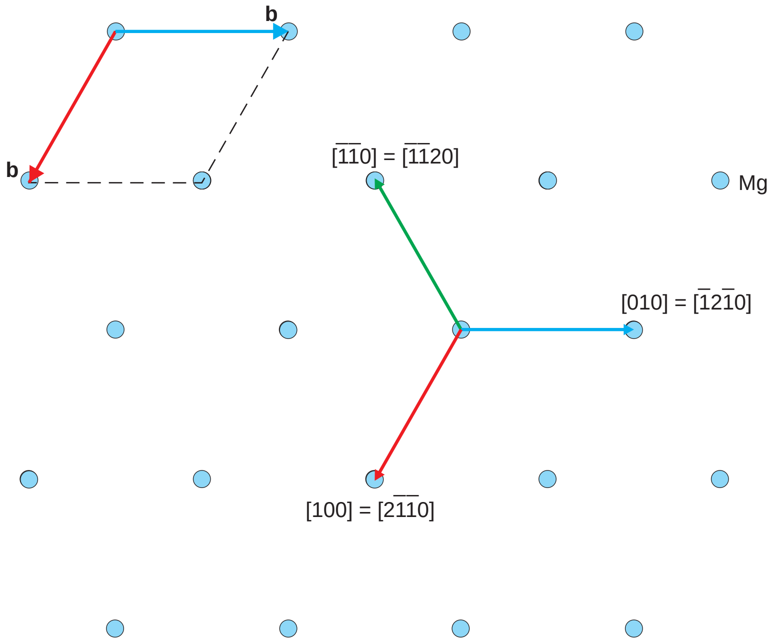

b
b
$[\bar{1}\bar{1}0] = [\bar{1}\bar{1}20]$
Mg
$[010] = [\bar{1}2\bar{1}0]$
$[100] = [2\bar{1}\bar{1}0]$

Figure 5.12 Directions in the basal (001) plane of a hexagonal crystal structure, given in terms of three indices, [*uvw*], and four indices [u′ *v*′ *t w*′].

above transformations. The correspondence is:

$$
\begin{aligned}
[100] &= [2\bar{1}\bar{1}0]\\
[010] &= [\bar{1}2\bar{1}0]\\
[\bar{1}\bar{1}0] &= [\bar{1}\bar{1}20]
\end{aligned}
$$

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*).

5.1.7 Crystal geometry and the reciprocal lattice

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 5.2.

Many of the physical properties of crystals, as well as the geometry of the three-dimensional patterns of radiation diffracted by crystals, are most easily described by using the *reciprocal lattice*. Crystal structures and Bravais lattices, sometimes called *direct lattices*, are said to occupy *real space*, while reciprocal lattices occupy *reciprocal space*.

The reciprocal lattice is defined in terms of three basis vectors labelled $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$. The *lengths* of the basis vectors of the reciprocal lattice are:

$$a^* = 1/d_{100},\ b^* = 1/d_{010},\ c^* = 1/d_{001}$$

where d_{100} is the spacing between (100) planes in the (real) crystal or lattice, and so on. Each reciprocal lattice point with coordinates *hkl* is associated with a set of crystal planes with Miller indices (*hkl*) and to the spacing between them, d_{hkl}.

The steps in the construction of a reciprocal lattice (illustrated for a section of a monoclinic unit cell) are:

1. Draw the unit cell (Figure 5.13a).
2. Draw lines perpendicular to the (100), (010) and (001) planes. These lines are perpendicular to the *end faces* of the unit cell and form the axes of the reciprocal lattice (Figure 5.13b,c).
3. Mark the lattice points at distances of $(1/d_{100}) = \mathbf{a}^*$, $(1/d_{010}) = \mathbf{b}^*$ and $(1/d_{001}) = \mathbf{c}^*$, and fill in the

Table 5.2 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*	$1/d^2_{hkl} = [1/V^2]\ \{[S_{11}h^2] + [S_{22}k^2] + [S_{33}l^2]$ $+ [2S_{12}hk] + [2S_{23}kl] + [2S_{13}hl]\}$	$abc \sqrt{}(1 - \cos^2 \alpha - \cos^2 \beta - \cos^2 \gamma$ $+ 2 \cos \alpha \cos \beta \cos \gamma)$
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	$1/d^2_{hkl} = \{[(h^2 + k^2 + l^2 \sin^2 \alpha) + 2(hk + kl + hl)$ $(\cos^2 \alpha - \cos \alpha]/[a^2(1 - 3\cos^2 \alpha + 2\cos^3 \alpha)]\}$	$a^3\ \sqrt{}(1 - 3\cos^2\alpha + 2\cos^3 \alpha)$

*$S_{11} = b^2c^2 \sin^2 \alpha$; $S_{22} = a^2c^2 \sin^2\beta$; $S_{33} = a^2b^2 \sin^2 \gamma$; $S_{12} = abc^2(\cos \alpha \cos \beta - \cos \gamma)$; $S_{23} = a^2bc(\cos \beta \cos \gamma - \cos \alpha)$; $S_{13} = ab^2c(\cos \gamma \cos \alpha - \cos \beta)$; $V =$ unit cell volume.

lattice by extending these over the required region of reciprocal space (Figure 5.13d).

4. Index the reciprocal lattice points with the same indices as the Miller indices of the appropriate *hkl* planes.

5. The distance of any point *hkl* from the origin of the lattice will be found to be $1/d_{hkl}$ and the direction from the origin to the *hkl* lattice point is normal to the (hkl) planes in real space.

For cubic, tetragonal and orthorhombic crystals:

$$a^* = 1/a,\ b^* = 1/b,\ c^* = 1/c$$

The reciprocal lattice axes are parallel to the direct lattice axes, which themselves are parallel to the

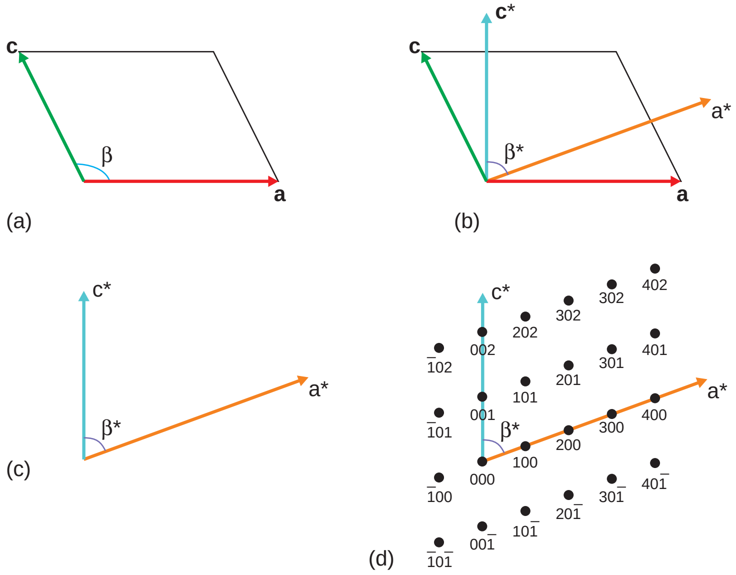


Figure 5.13 Construction of the reciprocal lattice of a monoclinic crystal.

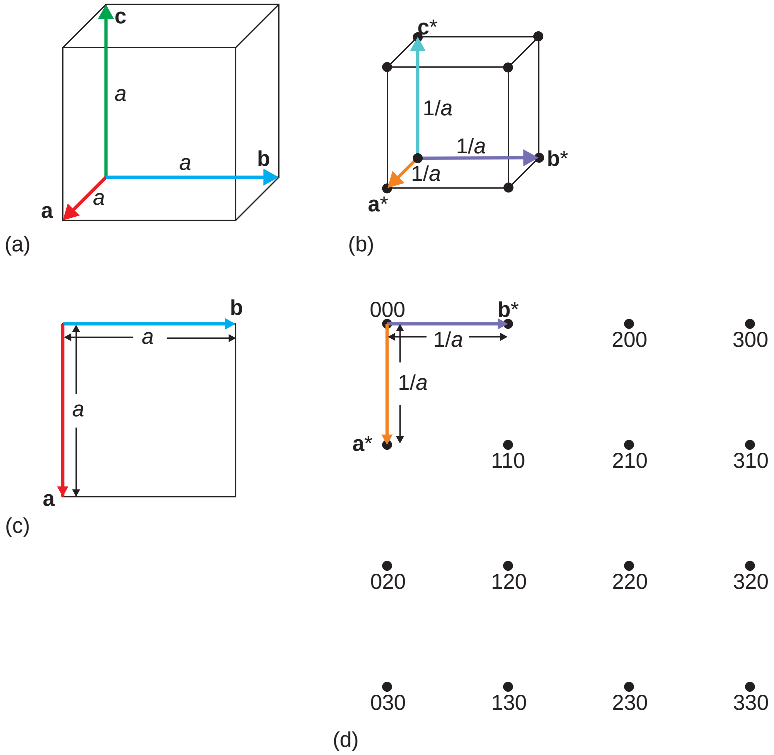


Figure 5.14 The direct lattice and reciprocal lattice of a cubic crystal. (a), (c) The direct lattice, specified by vectors **a**, **b** and **c**, with unit cell edges a ($=b=c$). (b), (d) The reciprocal lattice, specified by vectors **a***, **b*** and **c***, with unit cell edges $1/a$ ($=1/b=1/c$). The vector **a*** is parallel to **a**, **b*** parallel to **b**, and **c*** parallel to **c**.

unit cell edges, and the spacing of the lattice points hkl along the three reciprocal axes is equal to the reciprocal of the unit cell dimensions, $1/a=1/b=1/c$ (Figure 5.14). For some purposes it is convenient to multiply the length of the reciprocal axes by a constant. Thus, physics texts usually multiply the axes by 2π, and crystallographers by λ, the wavelength of the radiation used to obtain a diffraction pattern.

5.2 The determination of crystal structures

Crystal structures are determined by using diffraction (Section 14.7). X-ray diffraction is the most widespread technique used for structure determination, but diffraction of electrons and neutrons is also of great importance, as these reveal features that are not readily observed with X-rays.

The physics of diffraction by crystals has been worked out in detail. It is found that the incident radiation is diffracted in a characteristic way, called a *diffraction pattern*. If the positions of the diffracted beams (also called reflections) are recorded, they map out the reciprocal lattice of the crystal. The intensities of the beams are a function of the arrangements of the atoms in space and some other atomic properties, especially the atomic number of the atoms. Thus, if the positions and the intensities of the diffracted beams are recorded, it is possible to deduce the arrangement of the atoms in the crystal and their chemical nature.

5.2.1 Single crystal X-ray diffraction

In this technique, which is the most important structure determination tool, a small single crystal of the material, of the order of a fraction of a millimetre in size, is mounted in a beam of X-rays. The diffraction pattern used to be recorded photographically, but now the task is carried out electronically. The technique has been used to solve enormously complex structures, like that of huge proteins, or DNA.

Problems still remain, though. Any destruction of the perfection in the crystal structure degrades the sharpness of the diffracted beams. (This in itself can be used for crystallite size determination.) Poorly crystalline material gives poor information and truly amorphous samples give virtually no crystallographic information at all.

The intensity of radiation diffracted from a set of (*hkl*) planes depends upon the relative phases of the waves from adjacent planes, which in turn depends upon the symmetry of the structure and the types of atoms that are in the unit cell. When these waves are completely in step the diffracted beam is intense. When they are completely out of step, the diffracted intensity is zero. The relative phases of the beams are not preserved when the intensities are recorded but must be determined by indirect methods in order to solve the structure. Determination of the phases still poses problems in complex crystal structure determinations.

Some reflections, not present because of the symmetry of the lattice, are termed *systematic absences*. These are:

- F Bravais lattice, reflections are *present* for h, k and l all even or all odd, and *absent* for mixed even and odd combinations.
- I Bravais lattice, reflections are *absent* when $h + k + l$ is odd.
- C Bravais lattice, reflections are *absent* when $h + k$ is odd.

For example, the first five reflections from cubic crystals that are based upon an F Bravais lattice, such as NaCl, KCl and $MgAl_2O_4$, are {111}, {220}, {311}, {400} and {422}. The intensities of these allowed reflections will depend upon the atoms present in the crystal. Thus the {111} set is absent in the X-ray diffraction pattern of KCl because the scattering from the K and Cl atoms cancels, although it is present in the diffraction patterns from NaCl and $MgAl_2O_4$. If the reciprocal lattice of a crystal is drawn with lattice points corresponding to absent reflections omitted, and with each remaining *hkl* lattice point given a diameter proportional to the intensity of the diffracted beam from the (hkl) planes, it is called the *weighted reciprocal lattice*.

5.2.2 Powder X-ray diffraction and crystal identification

The diffraction pattern from a powder placed in the path of an X-ray beam gives rise to a series of cones rather than spots, because each plane in the crystallite can have any orientation (Figure 5.15a). The positions and intensities of the diffracted beams are recorded along a narrow strip (Figure 5.15b), and the diffracted beams are often called *lines* (Figure 5.15c). The *position* of a diffracted beam (not the intensity) is found to depend only upon the interplanar spacing, d_{hkl}, of the plane of atoms giving rise to the diffraction and the wavelength of the X-rays used, λ. Bragg's Law, equation (5.1), gives the connection between these quantities:

$$\lambda = 2d_{hkl}\sin\theta \qquad (5.1)$$

where θ is the diffraction angle (Figure 5.16). (Although the geometry of Figure 5.16 is identical to that of reflection, the physical process occurring is diffraction.)

The positions of the lines on the diffraction pattern of a single phase can be used to derive the unit cell dimensions of the material. The unit cell of a solid with a fixed composition is a constant. If the solid has a compositional 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

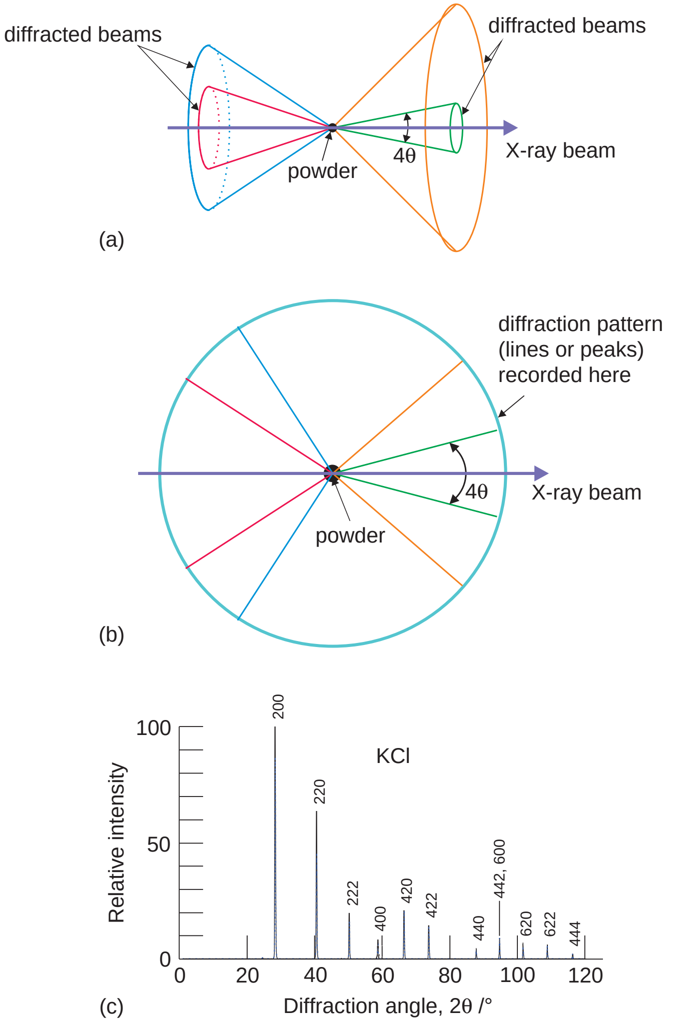


Figure 5.15 Powder X-ray diffraction: (a) a beam of X-rays incident upon a powder is diffracted into a series of cones; (b) the diffracted beams are recorded along a circle, to give a diffraction pattern; (c) the diffraction pattern from powdered potassium chloride, KCl, a cubic crystal. The numbers above the lines are the Miller indices of the diffracting planes.

of the two end members of the compositional range (Figure 5.17).

$$x = \frac{a_{ss} - a_1}{a_2 - a_1}$$

where a_1 and a_2 are the lattice parameters of the parent phases, a_{ss} is the lattice parameter of the solid solution, 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

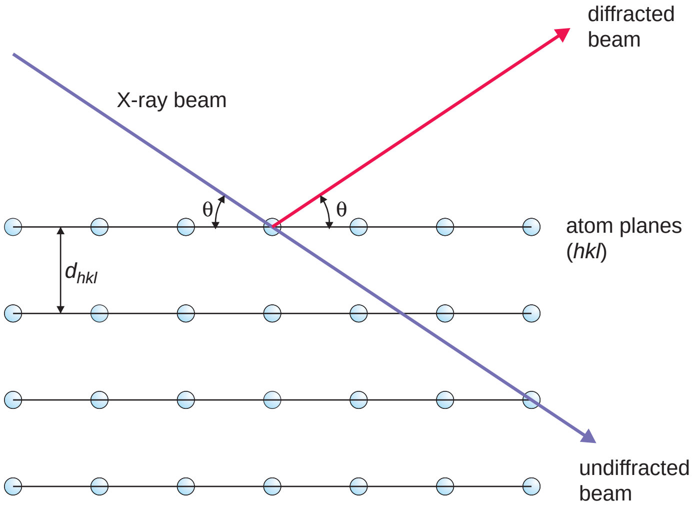


Figure 5.16 The geometry of diffraction from a set of crystal planes (hkl), with interplanar spacing d_{hkl}.

all cases, a plot of composition versus cell parameters can be used to determine the composition of intermediate compositions in a solid solution.

When the intensity and the positions of the diffracted beams are taken into account, every crystalline substance has a unique X-ray powder pattern. These can be likened to fingerprints, and mixtures of different crystals can be analysed if a reference set of patterns is consulted. The technique is routine in both metallurgical and mineralogical laboratories and is widely used in the determination of phase diagrams.

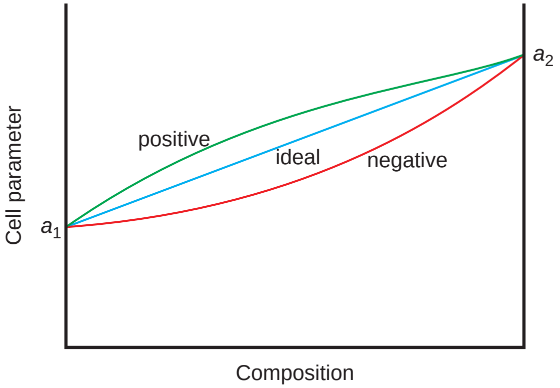


Figure 5.17 Vegard's law relating unit cell parameters to composition.

The experimental procedure can be illustrated with reference to the sodium fluoride–zinc fluoride (NaF–ZnF_2) system. Suppose that pure NaF is mixed with a few per cent of pure ZnF_2 and the mixture heated at 600°C until the reaction is complete. The X-ray powder diffraction pattern will show the presence of two phases: NaF, which will be the major component, and a small amount of a new compound (point A, Figure 5.18). A repetition of the experiment with gradually increasing amounts of ZnF_2 will yield a similar result, but the amount of the new phase will increase relative to the amount of NaF until a mixture of 1 part NaF plus 1 part ZnF_2 is heated. At this composition, only the new compound will be indicated on the X-ray powder diagram. It has the composition $NaZnF_3$.

A slight increase in the amount of ZnF_2 in the reaction mixture again yields an X-ray pattern that shows two phases to be present. Now, however, the compounds are $NaZnF_3$ and ZnF_2 (point B, Figure 5.18). This state of affairs continues as more ZnF_2 is added to the initial mixture, with the amount of $NaZnF_3$ decreasing and the amount of ZnF_2 increasing until pure ZnF_2 is reached. Careful preparations reveal the fact that NaF or ZnF_2 only appear alone on the X-ray films when they are pure, and $NaZnF_3$ only appears alone at the exact

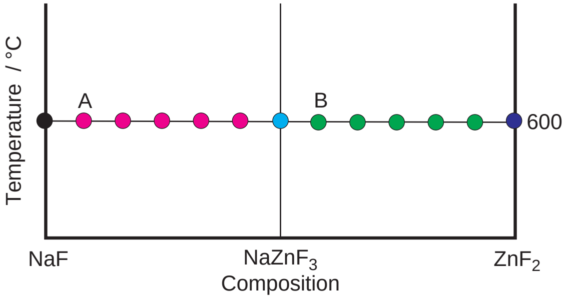


Figure 5.18 The determination of phase relations using X-ray diffraction. The X-ray powder patterns will show a single material to be present only at the exact compositions NaF, $NaZnF_3$ and ZnF_2. At points such as A, the solid will consist of NaF and $NaZnF_3$. At points such as B, the solid will consist of $NaZnF_3$ and ZnF_2.

composition of one mole NaF plus one mole ZnF_2. In addition, over all the compositional range studied, the unit cell dimensions of each of these three phases will be unaltered.

An extension of the experiments to higher temperatures will allow the whole of the solid part of the phase diagram to be mapped.

5.2.3 Neutron diffraction

Neutron diffraction is very similar to X-ray diffraction in principle, but quite different in practice, because neutrons need to be generated in a nuclear reactor. One advantage of using neutron diffraction is that it is often able to distinguish between atoms that are difficult to distinguish with X-rays. This is because the scattering of X-rays depends upon the atomic number of the elements, but this is not true for neutrons, and in some instances neighbouring atoms have quite different neutron-scattering capabilities, making them easily distinguished. Another advantage is that neutrons have a spin and so interact with unpaired electrons in the structure. Thus neutron diffraction gives rise to information about the magnetic properties of the material. The antiferromagnetic arrangement of the Ni^{2+} ions in nickel oxide, for example, was determined by neutron diffraction (Section 12.4).

5.2.4 Electron diffraction

Electrons are charged particles and interact very strongly with matter. This has two consequences for structure determination. Firstly, electrons will only pass through a gas or very thin solids. Secondly, each electron will be diffracted many times in traversing the sample, making the theory of electron diffraction more complex than the theory of X-ray diffraction, where each photon is presumed to be scattered only once in traversing the crystal. For this reason the relationship between the position and intensity of a diffracted beam is not easily related to the atomic positions in the unit cell. Moreover, delicate molecules are easily damaged by the intense electron beams needed for a successful diffraction experiment. Electron diffraction, therefore, is not used in the same routine way for structure determination as X-ray diffraction.

Electrons, however, do have one advantage. Because they are charged, they can be focused by magnetic lenses to form an image. The mechanism of diffraction as an electron beam passes through a thin flake of solid allows defects such as dislocations to be imaged with a resolution close to atomic dimensions. Similarly, diffraction of electrons from surfaces of thick solids allows surface details to be recorded, also with a resolution close to atomic scales. Thus, although electron diffraction is not widely used in structure determination, it is used as an important tool in the exploration of the microstructures and nanostructures of solids.

5.3 Crystal structures

5.3.1 Unit cells, atomic coordinates and nomenclature

Irrespective of the complexity of a crystal structure, it can be constructed by the packing together of unit cells. This means that the positions of all of the atoms in the crystal do not need to be given, but only those in a unit cell. The minimum amount of information needed to specify a crystal structure is thus the unit cell type, the cell parameters and the positions of the atoms in the unit cell. For example, the unit cell of the rutile form of titanium dioxide has a tetragonal unit cell, with cell parameters[2] of $a\ (=b) = 0.459\,\text{nm}$, $c = 0.296\,\text{nm}$.

The (x, y, z) coordinates of the atoms in each unit cell are expressed as fractions of a, b and c, the cell sides (Figure 5.3 and Section 5.1.2). 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 in the unit cell, written as Z. The unit cell of rutile, given above, has $Z = 2$, which means that there are two TiO_2 units in the cell, that is, two Ti atoms and four O atoms.

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. Metallurgical texts often refer to the structures of metals using a symbol for the structure. These symbols were employed by the journal *Zeitschrift für Kristallographie* in the catalogue of crystal structures *Strukturberichte* Vol. 1, published in 1920, and are called *Strukturberichte* symbols. For example, all solids with the same crystal structure as copper are grouped into the A1 structure type. These labels remain useful shorthand for simple structures, but become cumbersome when applied to complex materials, when the mineral name is often more convenient.

5.3.2 The density of a crystal

The atomic contents of the unit cell give the *composition* of the material. 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:

$$m_A = \frac{\text{molar mass}}{N_A}$$

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 is divided by the unit cell volume, V:

$$\rho = \frac{\sum_{i=1}^{q} n_i m_i / N_A}{V}$$

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.

[2] The unit cell dimensions are often specified in terms of the Ångström unit, Å. 10 Å = 1 nm.

5.3.2.1 The density of sodium chloride as an example

The unit cell of the halite structure contains 4 sodium (Na) and 4 chlorine (Cl) atoms (Section 5.3.8) and has a cubic unit cell with $a = 0.5500$ nm. The mass of the unit cell, m, is then given by:

$$\begin{aligned} m &= [(4 \times 22.99) + (4 \times 35.453)]/1000 \times N_{\mathrm{A}} \\ &= 3.882 \times 10^{-25}\ \mathrm{kg} \end{aligned}$$

where 22.99 g mol^{-1} is the molar mass of sodium, 35.453 g mol^{-1} is the molar mass of chlorine, N_{A} is the Avogadro constant, $6.02214 \times 10^{23}\ \mathrm{mol}^{-1}$, and the factor 1000 is to convert grams to kilograms.

The volume, V, of the cubic unit cell is given by a^3, thus:

$$V = (0.5500 \times 10^{-9})^3\ \mathrm{m}^3 = 1.66375 \times 10^{-28}\ \mathrm{m}^3$$

The density, ρ, is given by the mass m divided by the volume, V:

$$\begin{aligned} \rho &= \frac{3.882 \times 10^{-25}\ \mathrm{kg}}{1.66375 \times 10^{-28}\ \mathrm{m}^3} \\ &= 2333\ \mathrm{kg\ m}^{-3} \end{aligned}$$

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.

5.3.2.2 The density of materials with a variable composition

For a solid that has a variable composition, such as an alloy or a non-stoichiometric phase, the density will vary across the phase range. Similarly, the unit cell dimensions are found to change in a regular way across the phase range. These two techniques can be used in conjunction with each other to determine the most likely point defects that may be responsible for the composition change. (Note that as both these techniques are averaging techniques, they say nothing about the real 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 $\mathrm{MX}_{1.0}$ (Section 5.3.8). Wüstite has a composition that is always oxygen-rich compared with the ideal formula of $\mathrm{FeO}_{1.0}$. Data[3] 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, as interstitials, 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 interstitial oxygen atoms being present, over and above those on the normal anion positions. The ideal unit cell of the structure contains 4 Fe and 4 O, and so, in this model, the unit cell must contain 4 atoms of Fe and $4(1 + x)$ atoms of oxygen. The unit cell contents are $\mathrm{Fe}_4\mathrm{O}_{4+4x}$ and the (measured) composition is $\mathrm{FeO}_{1.058}$.

[3] The data are from the classical paper: Jette and Foote (1933) *J. Chem. Phys.*, **1**: 29.

The mass of 1 unit cell is m_A:

$$\begin{aligned} m_A &= [(4 \times 55.85) + (4 \times 16 \times (1 + x))]/N_A \\ &= [(4 \times 55.85) + (4 \times 16 \times 1.058)]/N_A \text{ grams} \\ &= 4.834 \times 10^{-25} \text{ kg} \end{aligned}$$

The volume, V, of the cubic unit cell is given by a^3, thus:

$$\mathrm{V} = (0.4301 \times 10^{-9})^3 \text{ m}^3 = 7.9562 \times 10^{-29} \text{ m}^3.$$

The density, ρ, is given by the mass m_A divided by the volume, V:

$$\rho = \frac{4.834 \times 10^{-25} \text{ kg}}{7.9562 \times 10^{-29} \text{ m}^3} = 6076 \text{ kg m}^{-3}$$

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 $\mathrm{Fe}_{4-4\mathrm{x}}\mathrm{O}_4$ and the (measured) composition is $\mathrm{Fe}_{1/1.058}\mathrm{O}_{1.0}$ or $\mathrm{Fe}_{0.945}\mathrm{O}$.

The mass of one unit cell is m_B:

$$\begin{aligned} m_B &= [(4 \times (1 - x) \times 55.85) + (4 \times 16)]/N_A \\ &= [(4 \times 0.945 \times 55.85) + (4 \times 16)]/N_A \text{ grams} \\ &= 4.568 \times 10^{-25} \text{ kg} \end{aligned}$$

The density, ρ, is given by m_B divided by the volume, V, to yield:

$$\rho = \frac{4.568 \times 10^{-25} \text{ kg}}{7.9562 \times 10^{-29} \text{ m}^3} = 5741 \text{ kg m}^{-3}$$

The difference in the two values from the different models 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 $\mathrm{Fe}_{0.945}\mathrm{O}$.

5.3.3 The cubic close-packed (A1) structure

- General formula: M; example, copper, Cu.
- Lattice: cubic face-centred, $a = 0.360$ nm.
- $Z = 4$ Cu.
- Atom positions: 0, 0, 0; ½, ½, 0; 0, ½, ½; ½, 0, ½.

There are four lattice points in the face-centred unit cell, and the motif is one atom at 0, 0, 0. The structure is typified by copper (Figure 5.19), but the cubic close-packed structure is adopted by many metals (Figure 6.1) and the noble gases, Ne(s), Ar(s), Kr(s), Xe(s). This structure is also called the face-centred cubic (fcc) structure or referred to by the Strukturbericht symbol, A1. 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 (sections 6.1.1 and 6.1.2).

5.3.4 The body-centred cubic (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.

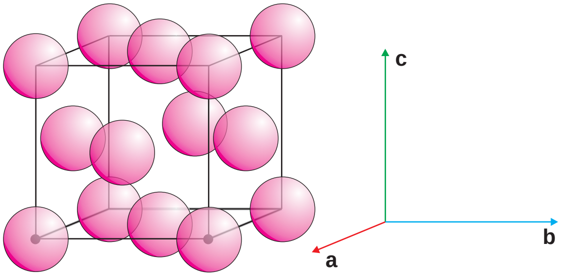


Figure 5.19 The A1 structure of copper.

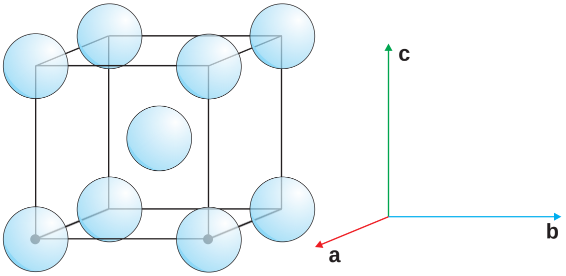


Figure 5.20 The A2 structure of tungsten.

There are two lattice points in the body-centred unit cell, and the motif is one atom at 0, 0, 0. The structure is adopted by tungsten, W (Figure 5.20), and many other metallic elements (Figure 6.1). This structure is often called the body-centred cubic (bcc) structure or referred to by the Strukturbericht symbol, A2. In this structure, each atom has 8 nearest neighbours and 6 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. This is less than either the A1 structure or the A3 structure (following), both of which have a volume fraction of occupied space of 0.7405. The bcc structure is often the high-temperature structure of a metal that has a close-packed structure at lower temperature (Sections 6.1.1 and 6.1.2).

5.3.5 The hexagonal (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; $^1/_3$, $^2/_3$, $^1/_2$.

The lattice is primitive, and so there is only one lattice point in each unit cell. The motif is two atoms, one atom at (0, 0, 0) and one atom at $^1/_3$, $^2/_3$, $^1/_2$. The structure is represented by magnesium, Mg (Figure 5.21). 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 $\sqrt{8}/\sqrt{3} \approx 1.633$. Many metals adopt the hexagonal A3 structure, some over a limited temperature range (Figure 6.1, sections 6.1.1 and 6.1.2).

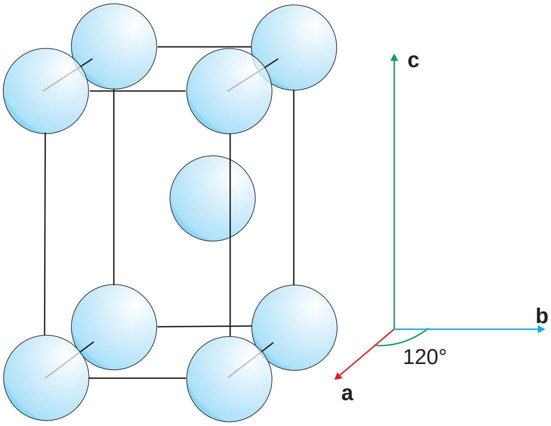


Figure 5.21 The A3 structure of magnesium.

5.3.6 The diamond (A4) structure

- General formula: M; example, diamond, C.
- Lattice: cubic face-centred, $a = 0.356$ nm.
- $Z = 8$ C.
- Atom positions: 0, 0, 0; $^1/_4$, $^1/_4$, $^1/_4$; repeated in the face-centred pattern.

There are four lattice points in the face-centred unit cell, and the motif is two atoms, one at 0, 0, 0 and one at $^1/_4$, $^1/_4$, $^1/_4$. 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 5.22). The bonds, of length 0.154 nm, are extremely strong sp^3-hybrids. The crystal can be regarded as a giant molecule.

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.

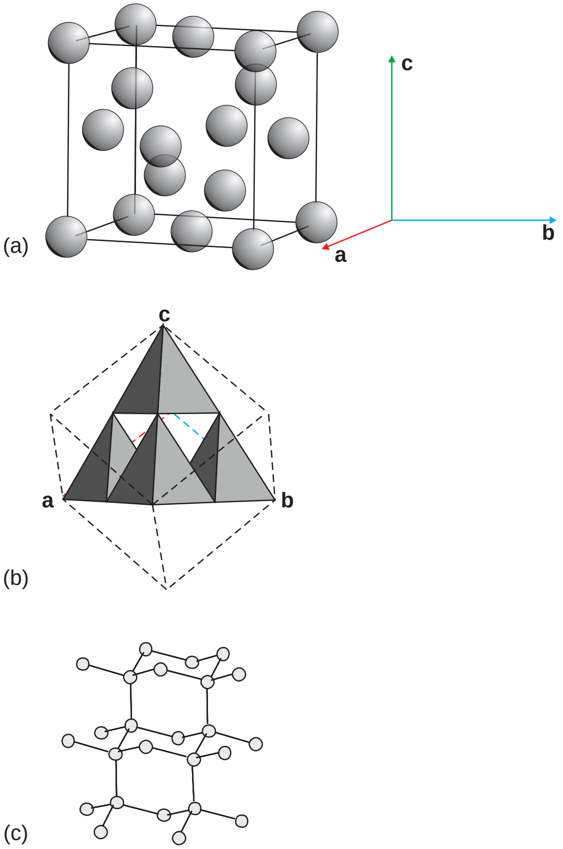


Figure 5.22 The A4 structure of diamond: (a) atoms in the unit cell; (b) cell projected approximately down [111] to show the structure as carbon-centred tetrahedral; (c), as (b), revealing the tetrahedral bond geometry.

5.3.7 The graphite (A9) 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, $^1/_2$; $^1/_3$, $^2/_3$, 0; $^2/_3$, $^1/_3$, $^1/_2$.

The lattice is primitive, and so there is only one lattice point in each unit cell. The motif is four atoms, at the positions specified above.

Graphite is a form of elemental carbon. The bonding in this material 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 5.23). Above and below the layers, π-bonds form a cloud of delocalised electrons which, 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.

Graphene (Figure 3.5b) is a single graphite layer and carbon nanotubes (Figure 3.5c) can be thought of as coiled up sheets of a single graphite layer.

5.3.8 The halite (rock salt, sodium chloride, B1) 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 $^1/_2$, $^1/_2$, $^1/_2$ 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 $^1/_2$, $^1/_2$, $^1/_2$. In this structure, called the halite, rock salt or sodium chloride structure, each ion is surrounded by six ions of the opposite type at the corners of a regular octahedron (Figure 5.24).

This structure is adopted by many oxides, sulphides, halides and nitrides with a formula MX.

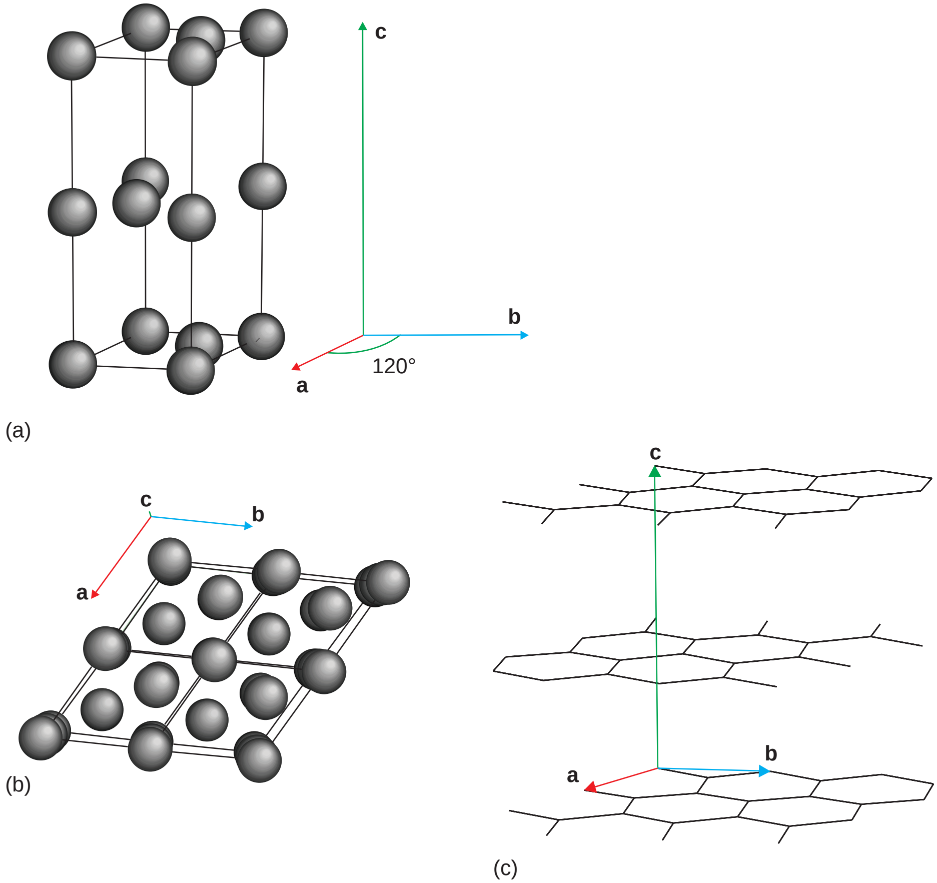


Figure 5.23 The A9 structure of graphite: (a) perspective view; (b) projection down [001]; (c) the structure drawn as hexagonal sheets. A carbon atom lies at each hexagonal vertex.

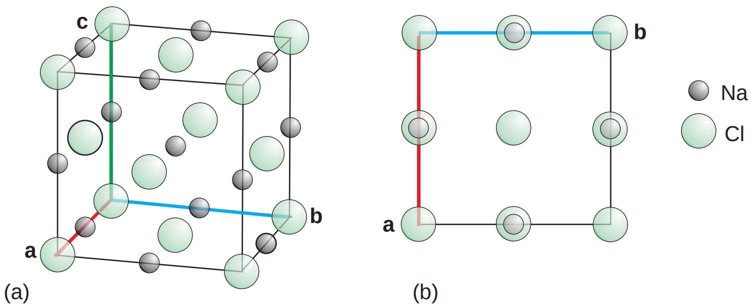


Figure 5.24 The B1 halite structure of NaCl: (a) perspective view; (b) projection down [001].

5.3.9 The spinel ($H1_1$) 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 $MgAl_2O_4$ complexes. This structure is named after the mineral spinel, $MgAl_2O_4$. 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 5.25). Thus 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 Mg and Al atoms are inserted into this array in an ordered fashion. To a good approximation, all of the magnesium atoms are surrounded by four oxygen atoms in the form of a tetrahedron and are said to occupy tetrahedral positions, or sites, in the structure. Similarly, to a good approximation, the aluminium atoms are surrounded by six oxygen atoms and are said to occupy octahedral positions or sites. When the structure is viewed down [111] the oxygen atoms can be seen to form cubic close packed layers, emphasising the relationship with the halite (NaCl) structure.

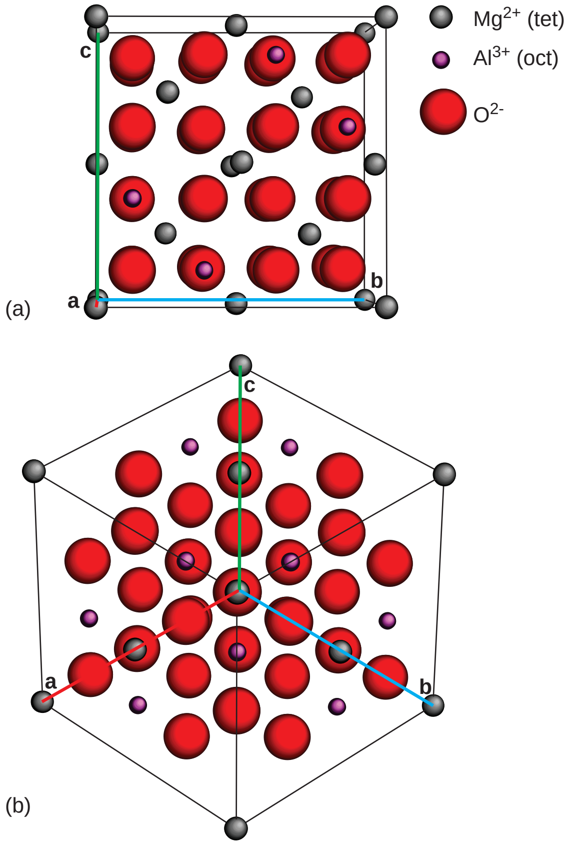


Figure 5.25 The $H1_1$ normal spinel structure of $MgAl_2O_4$: (a) projected down [100]; (b) projected down [111].

The mineral spinel, $MgAl_2O_4$, has given its name to an important group of compounds with the same structure, collectively known as spinels, which includes halides, sulphides and nitrides as well as oxides. Oxide spinels are often regarded as ionic compounds. The formula of the oxide spinels, AB_2O_4, is satisfied by a number of combinations of cations, the commonest of which is A^{2+} and B^{3+}, typified by Mg^{2+} and Al^{3+} in spinel itself.

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 half of the available octahedral positions and an eighth 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 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]O_4, with the tetrahedral cations enclosed () and the octahedral cations enclosed [ ]. 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]O_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:

$$\lambda = \frac{B^{3+}{}_{\text{tet}}}{B^{3+}{}_{\text{total}}}$$

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.

5.4 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.

5.4.1 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 an infinite number of ways of doing this, just two main arrangements are sufficient to describe many crystal 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 5.26). 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 5.27). 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 . . . . The structure has a *hexagonal* symmetry and unit cell. The **a**- and **b**-axes lie 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.

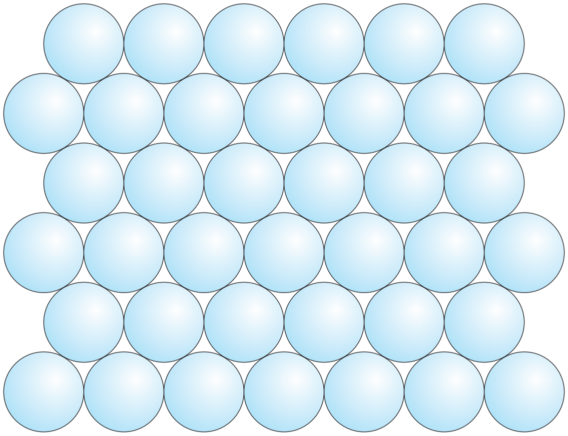


Figure 5.26 A single close packed layer of spheres.

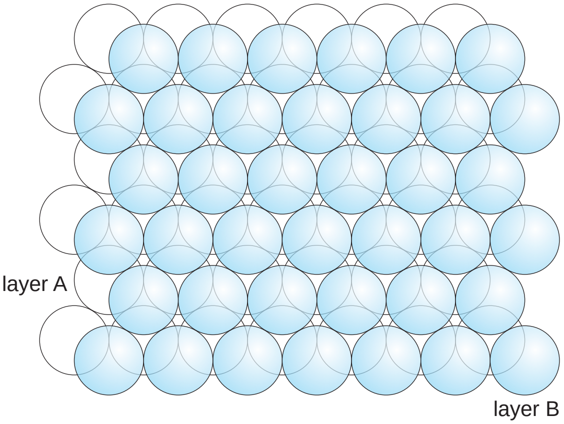


Figure 5.27 Hexagonal closest packing of spheres. The relative position of the layers follows the sequence . . . ABABAB . . . .

There are two spheres (two atoms) in a unit cell, at positions 0, 0, 0 and $^1/_3$, $^2/_3$, $^1/_2$. 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:

$$d = \frac{\sqrt{8r}}{\sqrt{3}} \approx 1.633r$$

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 A3 structure, although the c/a ratio departs from the ideal value of 1.633 in 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 5.28). This three-layer stacking is repeated indefinitely, thus: . . . ABCABC . . . . Although this structure can be described in terms of a hexagonal unit cell, the structure turns out to be *cubic*, and this description is always chosen. The close-packed layers of spheres lie perpendicular to the [111] direction and form (111) planes (Figure 5.29). The spacing of the close-packed planes for an ideal packing, d, is one third the body diagonal of the cubic unit cell, i.e. $a/\sqrt{3}$. If

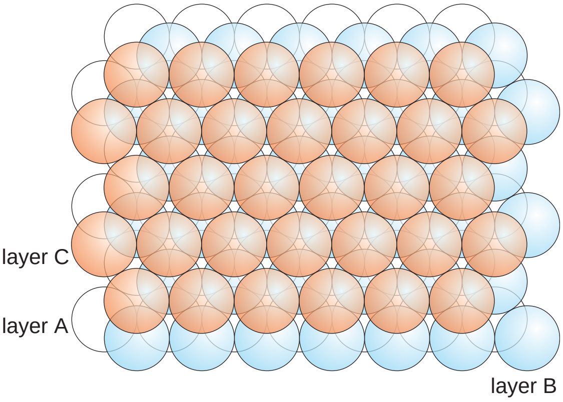


Figure 5.28 Cubic closest packing of spheres. The relative position of the layers follows the sequence . . . ABCABC . . . .

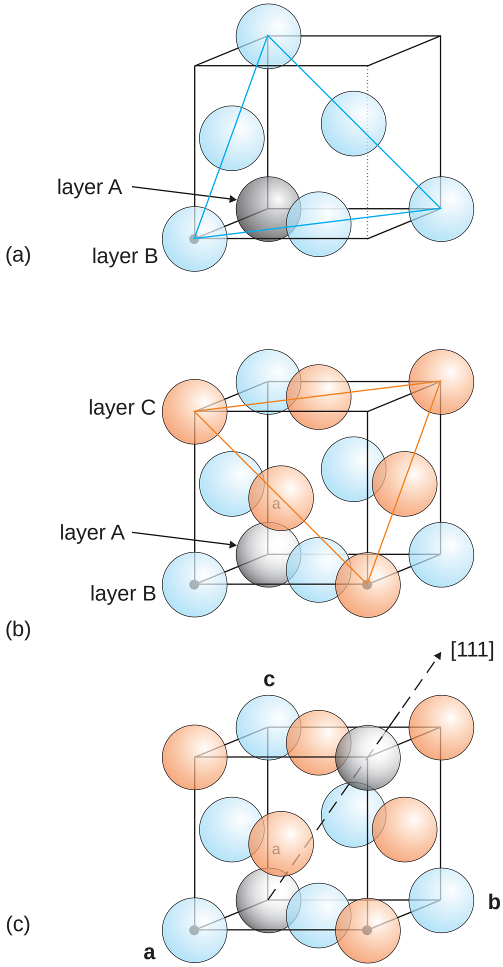


Figure 5.29 The cubic A1 structure in terms of cubic closest packing: (a) the first two layers, A and B; (b) the first three layers, A, B, C; (c) the fourth layer is completed by an atom at the top corner which is identical to the position (0,0,0) in the unit cell, and so is part of an A layer. The close-packed layers lie perpendicular to the [111] direction and form (111) planes.

the spheres just touch, the relationship between the sphere radius, r, and the lattice parameter a is:

$$r = \frac{a}{\sqrt{8}}$$

The relationship between the spacing of the close packed planes of spheres, d, the cell parameter a and r is therefore:

$$d = \frac{a}{\sqrt{3}} = \frac{\sqrt{8}r}{\sqrt{3}} \approx 1.633r$$

which is identical to that in hexagonal closest packing. The cubic close-packed structure is identical to the A1 structure.

Hexagonal and cubic close-packing 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 sulphide.

5.4.2 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 5.3). In such arrangements, small holes occur in layers between the sheets of anions. These holes, which are called *interstices*, *interstitial sites* or *interstitial positions*, are of two types (Figure 5.30). 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.

In the two close-packed sequences ABCABC and ABAB, there are $2N$ tetrahedral interstices and N octahedral interstices for every N anions. Structures

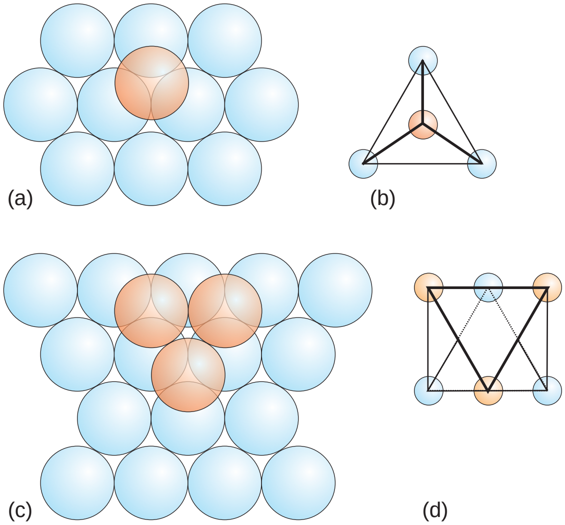


Figure 5.30 Tetrahedral and octahedral sites between closest packed arrays of spheres: (a) a tetrahedral site; (b) the same site drawn as a polyhedron; (c) an octahedral site; (d) the same site drawn as a polyhedron.

Table 5.3 Structures in terms of anion packing

Fraction of tetrahedral sites occupied	Fraction of octahedral sites occupied	Sequence of anion layers	
		. . . ABAB . . .	. . . ABCABC . . .
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)

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 ion of each sort present.

Compounds with halite (B1) 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 phase 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 of LiF, NaF, KF, and so on, 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, NiO and CoO.

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 the nicolite (NiAs) structure, which is adopted by a number of alloys and metallic sulphides, including 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 2/3 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 derives from a cubic close-packed array of oxygen ions with an ordered filling of half the available octahedral sites and an eighth 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.

5.4.3 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–nonmetal coordination polyhedra. These are 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. 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.

The complex families of silicates are best 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 5.31). 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 5.32 shows the way in which the [SiO_4] tetrahedra are linked in the commonest form of silica (SiO_2), quartz.

Octahedral coordination is frequently adopted by the important 3d transition-metal ions. Each cation is surrounded by six anions, to form an octahedral [MO_6] group (Figure 5.33). The structure of rhenium trioxide, ReO_3, in terms of linked [ReO_6] octahedra, has the appearance of a three-dimensional chessboard (Figure 5.34a). This structure is similar to that

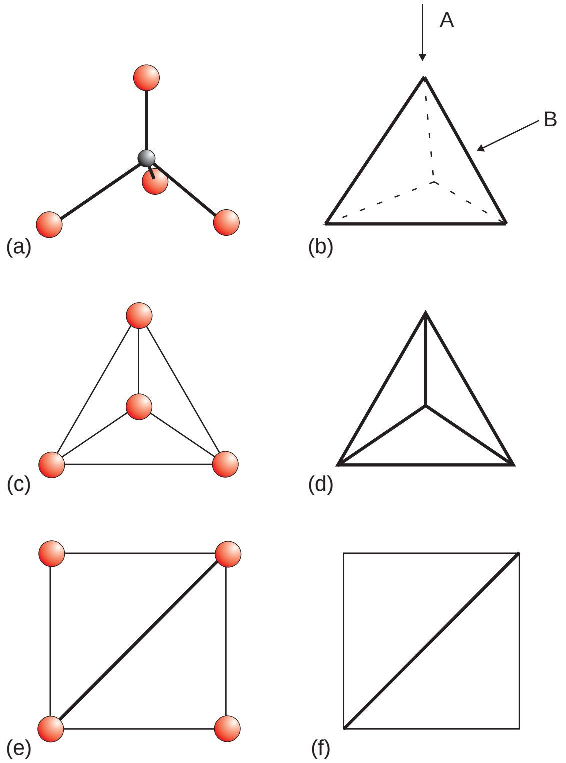


Figure 5.31 Representations of tetrahedra found in crystal structure diagrams: (a, b) show conventional tetrahedra; (c, d) show representations viewed down A in (a, b); (e, f) show representations viewed down B in (a, b).

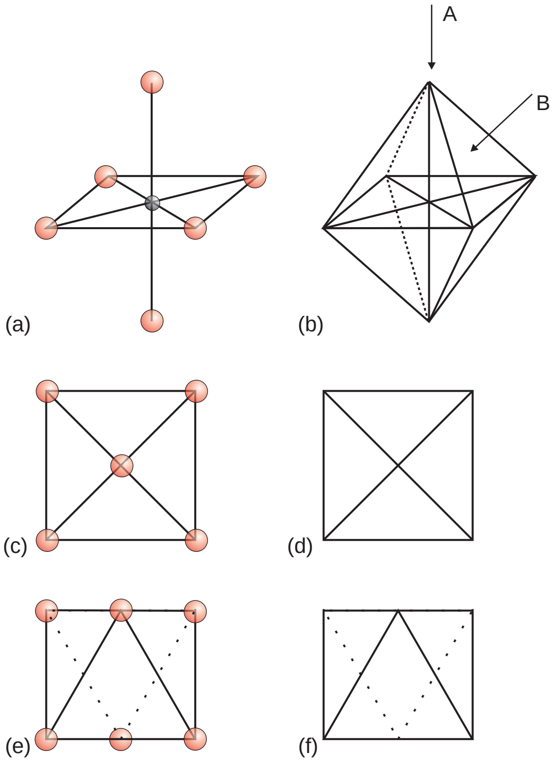


Figure 5.33 Representations of octahedra found in crystal structure diagrams: (a, b) show conventional octahedra; (c, d) show representations viewed down A in (a, b); (e, f) show representations viewed down B in (a, b).

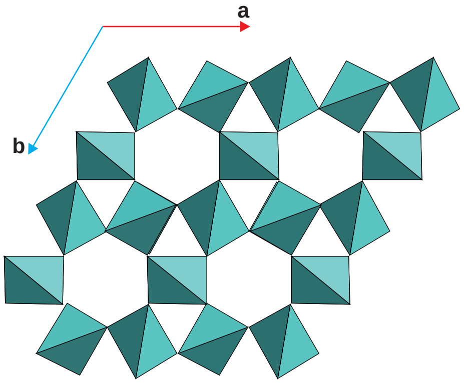


Figure 5.32 The structure of the high-temperature form of SiO_2, β-quartz, drawn as corner-shared tetrahedra projected down the hexagonal **c**-axis. This projection obscures the fact that the tetrahedra form three-dimensional spirals, not rings.

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 is similar, but has the large A cation in the centre of the cage of $[BO_6]$ octahedra (Figure 5.34b), although the cubic perovskite unit cell is generally depicted with the large A cations at the cell corners (Figure 5.34c). Most real perovskites are built of slightly distorted $[BO_6]$ octahedra, which reduce the symmetry from cubic to orthorhombic or monoclinic. These distortions, though, are often small, and relationships between the various perovskite phases can often be fruitfully described in terms of a pseudocubic unit cell (sections 8.3, 8.4 and 11.3.7).

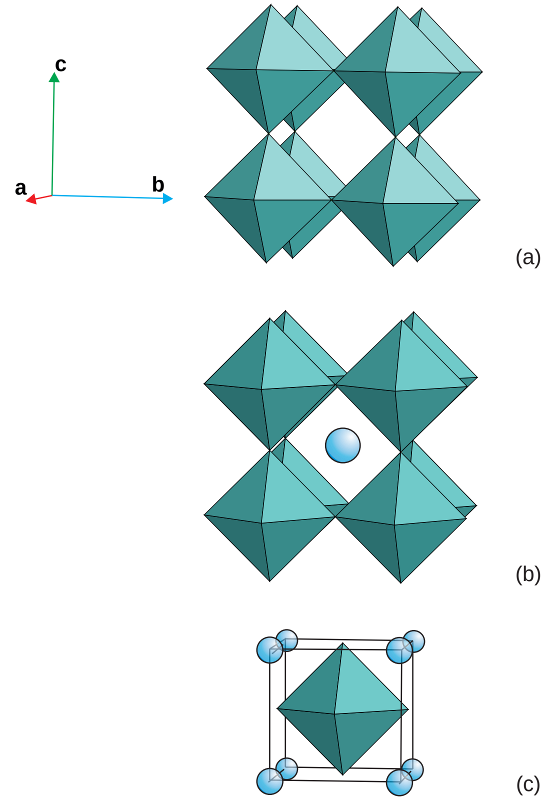


Figure 5.34 (a) The cubic ReO_3 structure represented as corner-shared $[ReO_6]$ octahedra. (b) The idealised cubic perovskite ABO_3 structure, where A is a large cation, typically Ca^{2+}, and B is a medium-sized cation, typically Ti^{4+}. The framework is identical to that in (a), and consists of corner-shared BO_6 octahedra containing an A cation in the central cage site. (c) The conventional unit cell of the cubic perovskite structure, with A cations at the cell corners and the $[BO_6]$ octahedron at the cell centre.

Further reading

Bloss, F.D. (1971) *Crystallography and Crystal Chemistry*. Holt Rinehart and Winston, New York.

Giacovazzo, C., Monaco, H.L., Artioli, G., *et al.* (2002) *Fundamentals of Crystallography*, 2nd edn. International Union of Crystallography, Oxford University Press, Oxford.

Megaw, H.D. (1973) *Crystal Structures*. W.B. Saunders, Philadelphia, PA.

O'Keeffe, M. (1977) On the arrangement of ions in crystals. *Acta Crystallographica* **A33**: 924.

O'Keeffe, M. and Hyde, B.G. (1985) An alternative approach to non-molecular crystal structures. *Structure and Bonding*, **61**: 77–144.

Smith, J.V. (1982) *Geometrical and Structural Crystallography*. John Wiley & Sons, Ltd., New York.

Tilley, R.J.D. (2006) *Crystals and Crystal Structures*. John Wiley & Sons, Ltd., Chichester.

Wells, A.F. (1984) *Structural Inorganic Chemistry*, 5th edn. Oxford University Press, Oxford.

Problems and exercises

Quick quiz

1 The basis vectors in a lattice define:
(a) The unit cell.
(b) The crystal structure.
(c) The atom positions.

2 The number of Bravais lattices is:
(a) 12.
(b) 13.
(c) 14.

3 A face-centred (F) unit cell contains:
(a) One lattice point.
(b) Two lattice points.
(c) Four lattice points.

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.

5 A crystal system is:
(a) A set of axes.
(b) A lattice.
(c) A unit cell.

6 A tetragonal unit cell has:
(a) $a = b = c$.
(b) $a = b \neq c$.
(c) $a \neq b \neq c$.

7 A crystal class represents:
(a) The internal symmetry of a crystal.
(b) The unit cell of a crystal.
(c) The crystal lattice.

8 A point group is identical to:
(a) A crystal structure.
(b) A crystal lattice.
(c) A crystal class.

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.

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.

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.

12 {*hkl*} means:
(a) A set of parallel planes.
(b) A group of non-equivalent planes.
(c) A family of symmetry-related planes.

13 Miller-Bravais indices (*hkil*) are used with:
(a) All non-cubic crystals.
(b) Hexagonal crystals.
(c) Primitive crystals.

14 A direction in a crystal structure is represented by:
(a) {uvw}.
(b) [uvw].
(c) <uvw>.

15 The atom coordinates $^1/_2$, $^1/_2$, $^1/_2$ represent an atom at:
(a) The centre of a unit cell.
(b) The face centres of a unit cell.
(c) The middle of the edges of a unit cell.

16 In a cubic close-packed (A1) unit cell there are:
(a) One atom.
(b) Two atoms.
(c) Four atoms.

17 In a body-centred cubic (A2) unit cell there are:
(a) One atom.
(b) Two atoms.
(c) Four atoms.

18 In a hexagonal (A3) unit cell there are:
(a) One atom.
(b) Two atoms.
(c) Four atoms.

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.

20 The sphere packing giving rise to a hexagonal structure is:
(a) . . . ABABAB . . .
(b) . . . ABCABC . . .
(c) . . . ABACABAC . . .

21 In a cubic close-packed array of *N* spheres there are:
(a) N tetrahedral interstices.
(b) 2N tetrahedral interstices.
(c) 4N tetrahedral interstices.

22 In a hexagonal close-packed array of *N* spheres there are:
(a) N octahedral interstices.
(b) 2N octahedral interstices.
(c) 4N octahedral interstices.

23 A tetrahedron is a polyhedron with:
(a) Four triangular faces.
(b) Six triangular faces.
(c) Eight triangular faces.

24 An octahedron is a polyhedron with:
(a) Four triangular faces.
(b) Six triangular faces.
(c) Eight triangular faces.

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.

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.

27 The wurtzite structure of zinc sulphide has the sulphur atoms in a hexagonal close-packed array while the zinc atoms occupy half of the tetrahedral positions. The formula of the sulphide is:

(a) Zn_2S.

(b) ZnS.

(c) ZnS_2.

Calculations and questions

5.1 Index the planes (a)–(d) shown in Figure 5.35A.

5.2 Index the planes (a)–(d) shown in Figure 5.35B.

5.3 Index the planes (a)–(d) shown in Figure 5.35C.

5.4 Index the planes (a)–(d) shown in Figure 5.35D.

5.5 Sketch the (111) and $(1\bar{1}1)$ planes in a cubic crystal.

5.6 List the planes belonging to the {110} set in a cubic crystal.

5.7 List the planes belonging to the {111} set in a cubic crystal.

5.8 List the planes belonging to the {hh0} set in a cubic crystal.

5.9 List the planes belonging to the {hk0} set in a cubic crystal.

5.10 Index the directions (a)–(e) shown in Figure 5.36A.

5.11 Index the directions (a)–(e) shown in Figure 5.36B.

5.12 Index the directions (a)–(e) shown in Figure 5.36C.

5.13 Index the directions (a)–(e) shown in Figure 5.36D.

5.14 List the directions belonging to the $\langle 100 \rangle$ set.

5.15 List the directions belonging to the $\langle 110 \rangle$ set.

5.16 What is the angle between (110) and [110] in a cubic crystal?

5.17 What is the angle between (132) and [132] in a cubic crystal?

5.18 Sketch the reciprocal lattice of a cubic crystal with $a = 5$ nm.

5.19 Nickel has the A1 (face-centred cubic) structure, with $a = 0.352$ nm. A powder sample is irradiated with X-rays with a wavelength of 0.1542 nm. What angles would the diffracted beams from the (111), (220) and (400) planes make with the incident beam direction?

5.20 Tantalum has the A2 (body-centred cubic) structure, with $a = 0.3303$ nm. A powder sample is irradiated with X-rays with a wavelength of 0.1542 nm. What angles would the diffracted beams from the (110), (211) and (310) planes make with the incident beam direction?

5.21 A sample of the cubic alloy β-brass (an alloy of copper and zinc) gives an X-ray powder pattern in which the first three reflections are (100), $\theta = 22.9°$; (110), $\theta = 33.35°$; (111), $\theta = 42.35°$; when the X-ray wavelength is 0.229 nm. Calculate the lattice parameters of the brass.

5.22 A sample of the cubic spinel $CuAl_2O_4$ gives an X-ray powder pattern in which the first three reflections are (111), $\theta = 9.51°$; (200), $\theta = 11.00°$; (220), $\theta = 15.65°$; when the X-ray wavelength is 0.1541 nm. Calculate the lattice parameters of the spinel.

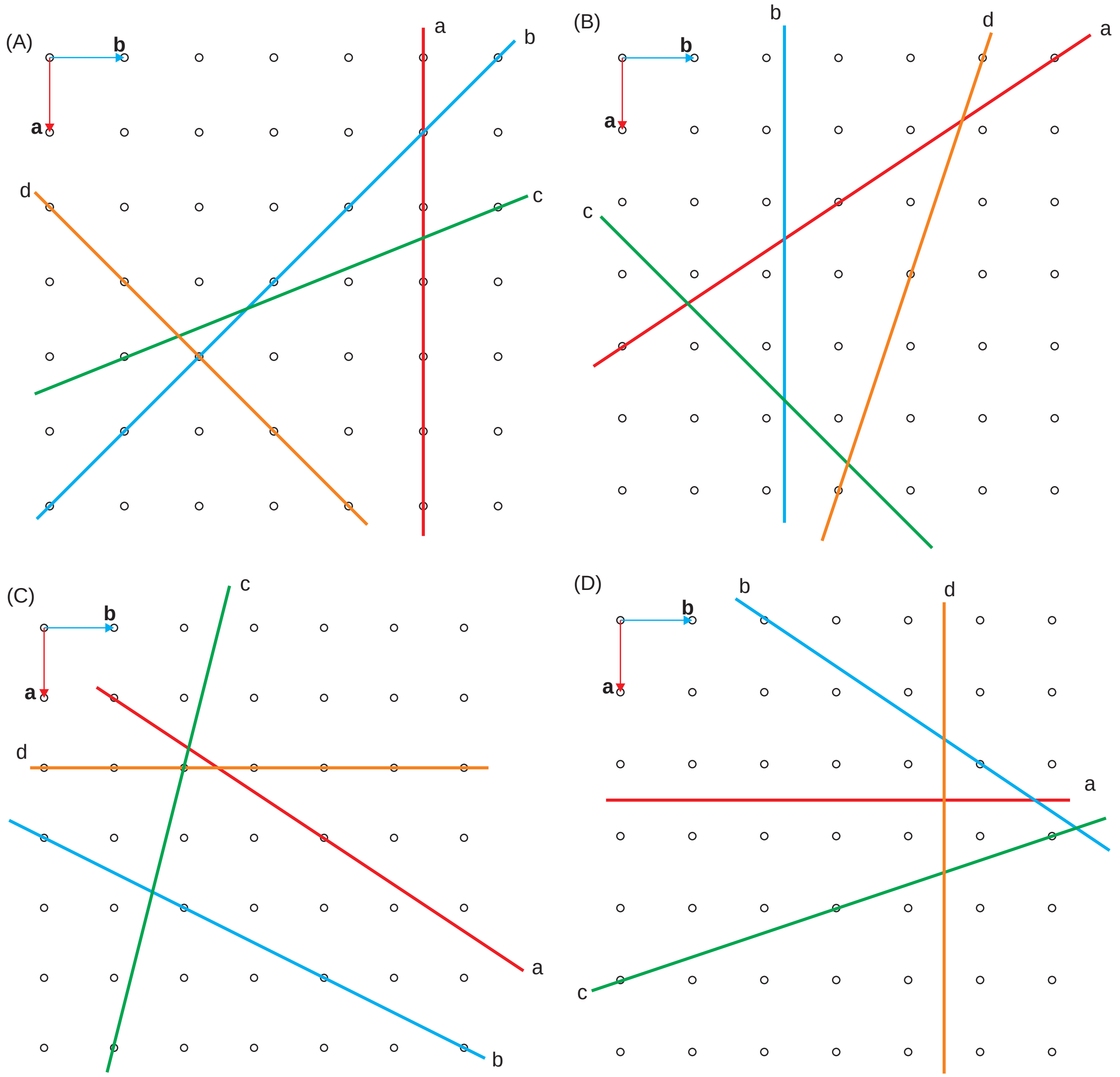


Figure 5.35 (A) to (D) The lines represent planes in a cubic crystal parallel to the **c**-axis (normal to the plane of the page). The circles mark the corners of the cubic unit cell.

5.23 A mineral sample contains a crystalline oxide with a formula $NiAl_2O_x$, where x is uncertain. The crystals gave an X-ray powder pattern with reflections characteristic of a cubic material. The first reflection, (111), was at $\theta = 9.55°$ when the X-ray wavelength was 0.1541 nm. Confirm that this is consistent with the spinel nickel aluminate, $NiAl_2O_4$, for which $a = 0.8048$ nm.

5.24 The unit cell size of CaO is 0.48105 nm and that of SrO is 0.51602 nm. Both adopt the halite (B1) structure-type. Estimate the composition of a crystal of formula

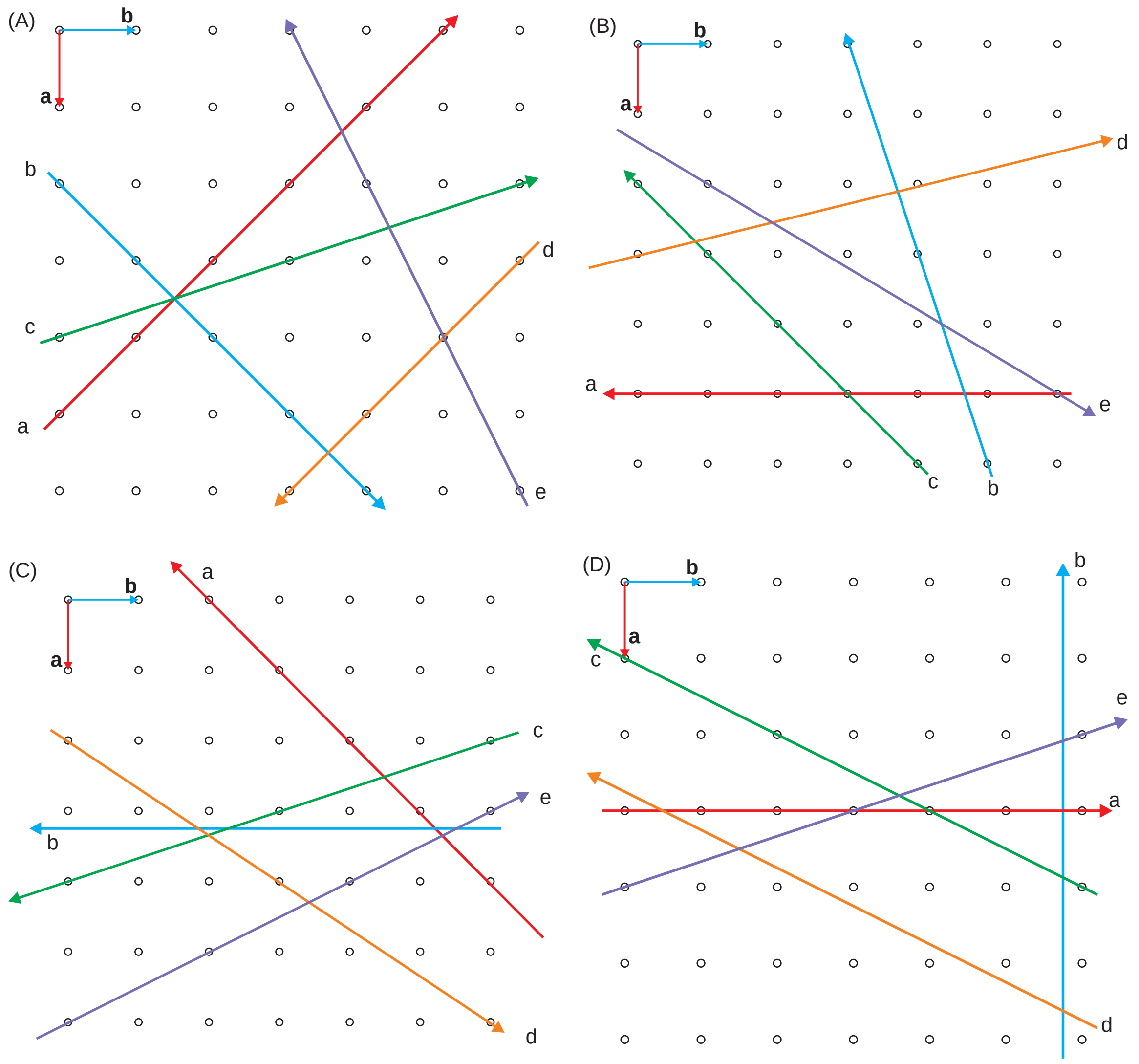


Figure 5.36 (A) to (D) The lines represent directions (in the plane of the page) in a cubic crystal. The circles mark the corners of the cubic unit cell.

$Sr_xCa_{1-x}O$, which was found to have a unit cell of 0.5003 nm.

5.25 A mixed cubic spinel $ZnAl_{2-x}Ga_xO_4$, is made up by heating together $ZnAl_2O_4$ ($a = 0.8086$ nm) and $ZnGa_2O_4$ ($a = 0.8328$ nm). The X-ray powder pattern, taken using radiation of wavelength 0.1541 nm, gave the first reflection, (111), at a position $\theta = 9.435°$. Estimate the value of x.

5.26 The cubic unit cell of iridium is drawn in Figure 5.37a. What are the atomic coordinates? What is the unit cell type?

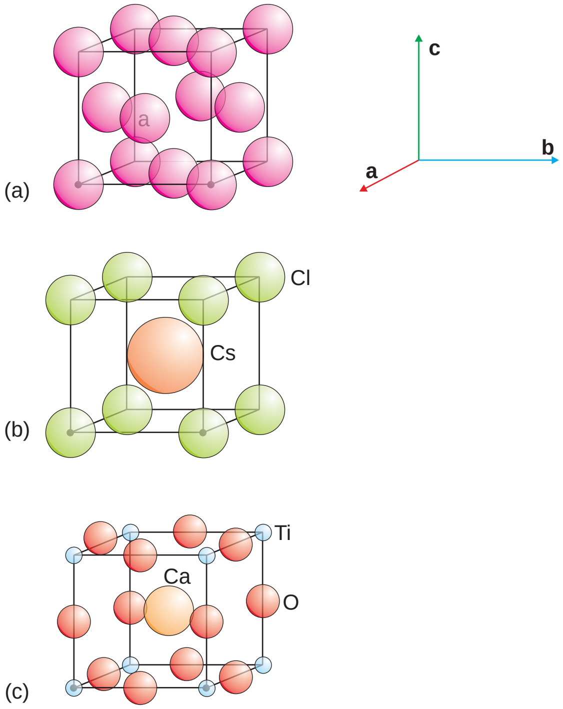


Figure 5.37 The cubic unit cells of (a) iridium, (b) CsCl, and (c) $CaTiO_3$.

5.27 The cubic unit cell of CsCl is drawn in Figure 5.37b. What are the atomic coordinates of each ion?

5.28 The cubic unit cell of perovskite, $CaTiO_3$, is drawn in Figure 5.37c. What are the atomic coordinates of the atoms?

5.29 The atom positions in cubic nickel oxide are:

Ni: 0,0,0 $^1/_2$,$^1/_2$,0 0,$^1/_2$,$^1/_2$ $^1/_2$,0,$^1/_2$
O: $^1/_2$,$^1/_2$,$^1/_2$ 0,0,$^1/_2$ $^1/_2$,0,0 0,$^1/_2$,0

Sketch the unit cell. What is the formula of the oxide? What is the structure type?

5.30 A copper–gold alloy has a cubic structure. The atom positions are:

Au: 0,0,0
Cu: 0,$^1/_2$,$^1/_2$ $^1/_2$,0,$^1/_2$ $^1/_2$,$^1/_2$,0

Sketch the unit cell and determine the formula of the alloy.

5.31 Aluminium has the cubic A1 structure with a lattice parameter of 0.4050 nm. Estimate the density of the metal.

5.32 Tungsten has the cubic A2 structure with a lattice parameter of 0.31651 nm. Estimate the density of the metal.

5.33 Magnesium has the A3 structure with hexagonal lattice parameters of $a = 0.320$ nm, $c = 0.520$ nm. Estimate the density of the metal.

5.34 Copper has the cubic A1 structure and a density of 8.96×10^3 kg m^{-3}. What is the length of the unit cell edge?

5.35 A sample of calcia-stabilised zirconia is 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 4 metal positions and 8 non-metal positions, to give an overall composition of MX_2.)

5.36 The unit cell of a zirconium sulphide, with a measured composition of 77 Zr: 100 S, is of the halite (B1) 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 B1 structure, and give an opinion on the defect structure of the real material.

5.37 Will the density of a crystal go up or down if it contains: (a) Schottky defects; (b) Frenkel defects; (c) vacancies; (d) interstitials?

5.38 An iron titanium oxide has the anions in a hexagonal close-packed array. The Fe and Ti atoms each occupy a third 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?

PART 2

Classes of Materials

6

Metals, ceramics, polymers and composites

- Why are alloys usually stronger than pure metals?
- How is strengthened glass made?
- Why are plastic bags hard to degrade?

Traditionally, materials have been divided into three major groups: metals, ceramics and polymers. Metallic materials are made up of pure metals, for example, titanium, iron or copper, together with vast numbers of alloys, among which are the historically important materials bronze, brass and steel. Ceramics bring to mind porcelain, silicon carbide, glass, and synthetic gemstones such as ruby and zirconia. Polymers are mainly compounds of carbon and include the familiar materials poly(vinyl chloride), polyethylene and nylon, as well as important biological molecules such as DNA.

In addition to these major divisions, two others should be mentioned: composites and biomaterials. Composites are combinations of materials from more than one of the groups listed above and have superior properties to the separate compounds. For example, glass-fibre (ceramic) reinforced epoxy resin (polymer) has mechanical properties superior to either of the separate components. One of the most important of composites is concrete, which is a composite of cement and stony material called aggregate.

Biomaterials are naturally occurring materials with important properties, such as wood, silk and bone. They are invariably composites. Because of the superior properties of many biomaterials, much effort is placed into trying to recreate these materials synthetically, as biomimetics.

At first sight, metals, ceramics and polymers have little in common. This is because of two main factors: the chemical bonding holding the atoms together and the microstructures of the solids themselves. Both of these are quite different in representative examples of each material. However, the difference is illusory. Many ceramics can be considered as metals, for example the ceramic superconductors. Many polymers show electronic conductivity greater than metals, and have use in lightweight batteries and electronic devices. The material in this and later chapters will allow these apparent anomalies to be understood.

6.1 Metals

Roughly speaking, about three-quarters of the elements can be regarded as metallic. Because of the variation in the outer electron configuration that this implies, one might expect that a large variety of metallic structures would form, and that these would vary in a predictable way across the periodic table.

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

It is rather surprising, therefore, to find that the majority of metallic elements possess one of only three structures. This fact arises because the outer electrons of metals are distributed throughout the crystal structure and the core that remains is, to a very good approximation, spherical. The crystal structures of many metals can then be approximated to sphere packings. Alloys, materials made up of two or more metallic elements, lose this simplicity and show a much greater variety of structures.

6.1.1 The crystal structures of pure metals

Most pure metals mainly adopt one of three crystal structures: A1, copper structure, cubic close-packed; A2, tungsten structure, body-centred cubic; or A3, magnesium structure, hexagonal close-packed (Figure 6.1). The difference in energy between these structures is small, and changes are commonly induced by variation of temperature and pressure. The different forms are called *allotropes* (Table 6.1).

It is surprising, in view of the many structures that are derived from either the hexagonal (ABAB) or cubic (ABCABC) close-packing (Section 5.4), that so few complex arrangements occur. Cobalt is one metal that shows this behaviour. Below about 435 °C the structure is a disordered random stacking of A, B and C planes of metal atoms. It can be transformed into the A3 structure by careful annealing at lower temperatures, and this transforms to the A1 structure above 435 °C. The metals lanthanum, praseodymium and neodymium adopt mixed close-packing that has an ABAC repeat. Samarium has a packing repeat of BABCAC.

At the right-hand side of Figure 6.1, simple structures are no longer found. These elements were once called the *semi-metals*. In them, the outer electrons are not completely lost to the structure and the shapes of the electron orbitals begin to influence bonding. This first becomes noticeable in the anomalous metal, mercury. The structure can be thought of as the A1 structure compressed along one body diagonal, to become rhombohedral. Similarly, indium has a slightly distorted A1 structure.

Stronger bonding effects are found within the carbon group. At normal temperatures and pressures,

Li A2 0.3509	Be A3 *a* 0.2286 *c* 0.3585											B	C
Na A2 0.4291	Mg A3 *a* 0.3209 *c* 0.5211											Al A1 0.4050	Si
K A2 0.5321	Ca A1 0.5588	Sc A3 *a* 0.3309 *c* 0.5268	Ti A3 *a* 0.2951 *c* 0.4686	V A2 0.3024	Cr A2 0.2885	Mn	Fe A2 0.2867	Co	Ni A1 0.3524	Cu A1 0.3615	Zn A3 a 0.2665 c 0.4947	Ga	Ge
Rb A2 0.5705	Sr A1 0.6084	Y A3 *a* 0.3648 *c* 0.5732	Zr A3 *a* 0.3232 *c* 0.5148	Nb A2 0.3300	Mo A2 0.3147	Tc A3 *a* 0.2738 *c* 0.4393	Ru A3 *a* 0.2706 *c* 0.4282	Rh A1 0.3803	Pd A1 0.3890	Ag A1 0.4086	Cd A3 a 0.2979 c 0.5620	In	Sn
Cs A2 0.6141	Ba A2 0.5023	La	Hf A3 *a* 0.3195 *c* 0.5051	Ta A2 0.3303	W A2 0.3165	Re A3 *a* 0.2761 *c* 0.4458	Os A3 *a* 0.2734 *c* 0.4392	Ir A1 0.3839	Pt A1 0.3924	Au A1 0.4078	Hg	Tl A3 *a* 0.3457 *c* 0.5525	Pb A1 0.4950

Figure 6.1 The crystal structures of the metallic elements at room temperature (25 °C) and atmospheric pressure. Unit cell parameters (*a* for cubic structures) are in nm. A1: copper (cubic close-packed). A2: tungsten (body-centred cubic). A3: magnesium (hexagonal close-packed).

Table 6.1 Allotropic structures of some metals

Element	Room temperature structure	High-temperature structure	Transition temperature/°C
Calcium, Ca	A1	A2	445
Strontium, Sr	A1	A2	527
Scandium, Sc	A3	A2	1337
Titanium, Ti	A3	A2	883
Zirconium, Zr	A3	A2	868
Hafnium. Hf	A3	A2	1742
Ytterbium, Yb	A3	A2	1481
Iron, Fe	A2	A1	912
Cobalt, Co	(A3)	A1	435

the bonding in carbon (graphite) is a mixture of sp^2 and weaker van der Waals bonding. At high pressures, graphite transforms to the diamond structure in which the atoms are linked by sp^3-hybrid bonds arranged tetrahedrally. The diamond structure is adopted by silicon and germanium at normal temperatures and pressures. Tin is a borderline solid from the point of view of bonding effects. At temperatures below 13.2 °C, the allotrope α-tin (grey tin) is stable. This has the diamond structure built with sp^3-hybrid bonding. At temperatures above 13.2 °C the stable structure is β-tin (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, the structure is complex and not simply related to the A1, A2 or A3 structures, revealing the importance of bonding effects. With lead, the increased atomic size leads to extensive outer electron delocalisation. The solid is metallic and the structure is the A1 type.

In the semimetals antimony, arsenic and bismuth, bonding effects are more pronounced, and the structures are not related to the structures of most metals. Bismuth, the heaviest, is the most metallic, and phosphorus, lying above antimony in the periodic table, is not considered even to be a semimetal.

6.1.2 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 6.2). The relationship between the cell edge, a for cubic crystals, a, c, for hexagonal crystals, and the radius of the component atoms, r, for the three common metallic structures is given below.

A1, copper structure, face-centred cubic The atoms are in contact along a cube face diagonal, so that:

$$r = \frac{a}{2\sqrt{2}} = \frac{a}{\sqrt{8}}$$

The separation of the close-packed (111) atom planes, (along a cube body diagonal), is:

$$d_{111} = \frac{a}{\sqrt{3}} = \frac{\sqrt{8}r}{\sqrt{3}} \approx 1.633r$$

Each atom has 12 nearest neighbours.

A2, tungsten structure, body-centred cubic The atoms are in contact along a cube body diagonal, which is equal to $4r$. This distance is also equal to 3 d_{111}, i.e. $3a/\sqrt{3}$, so that:

$$r = \frac{\sqrt{3}a}{4}$$

Each atom has 8 nearest neighbours.

Li 0.1562	Be 0.1128											B	C
Na 0.1911	Mg 0.1602											Al 0.1432	Si
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	Ge
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	Tl 0.1716	Pb 0.1750

Figure 6.2 Metallic radii, nm, for 12-coordinated elements.

A3, magnesium structure, hexagonal close packed The atoms are in contact along the **a**-axis, hence:

$$r = \frac{a}{2}$$

The separation of the close packed atom planes is $c/2$. The ratio of c/a in an ideal close-packed structure is $\sqrt{8}/\sqrt{3} \approx 1.633$.
Each atom has 12 nearest neighbours.

As in the case of ionic radii, the radius determined experimentally is found to depend upon the number of nearest neighbours (the coordination number, CN, of the atom in question). Atoms in both the A1 and A3 structures have 12 nearest neighbours (CN 12), and the radius determined will be appropriate to that coordination. Atoms in the 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:

$$\text{Radius (CN 12)} = 1.032\, r\,(\text{CN 8}) - 0.0006$$

when the radii are measured in nm.

The metallic radii of a few important elements, notably Mn, Ga and Sn, have complex structures that do not allow for the application of the simple rules above. The radii for these latter elements are derived from a comparison of the interatomic distances in many 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 passes along the period, and these generally increase with atomic number going down a group. The same is true for the lanthanoids and actinoids.

6.1.3 Alloy solid solutions

Alloys are important because they often show superior mechanical properties compared with 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 to those of the pure metals. *Substitutional solid solutions* have a structure identical to one of the metals involved, called the *parent structure*. The alloy-forming (or *foreign*) atoms simply occupy positions in the structure normally occupied by the *parent atoms*. *Interstitial solid solutions* are formed when very small foreign atoms enter the parent structure and sit in normally unoccupied interstitial positions. As these names imply, the foreign atoms in both of these examples are often regarded as in *solution* in the matrix of the parent metal.

6.1.3.1 Substitutional solid solutions

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 20th 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 (Section 4.2.3) 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 A1 (face-centred cubic) structure (Figure 6.3a). 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 6.3b), 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 6.3c) has alternating (100) planes composed of either copper or gold atoms. This structure is called the CuAu I

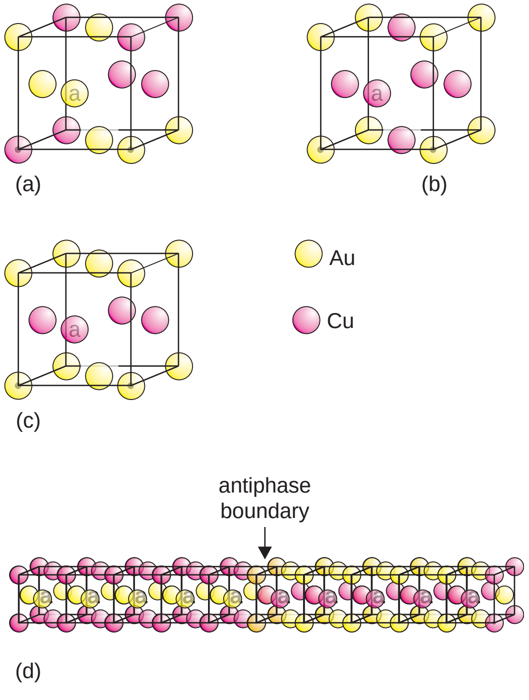


Figure 6.3 The crystal structures of: (a) disordered CuAu; (b) ordered Cu_3Au; (c) ordered CuAu I; (d) ordered CuAu II.

structure. Arrangements that are more complex also occur. For example, in the CuAu II structure, (Figure 6.3d), 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.

Ordered structures formed by annealing are commonplace and play a large part in determining the physical, and especially mechanical, properties of substitutional alloys.

6.1.3.2 Interstitial solid solutions

Just as atoms should ideally be of similar sizes to form extensive substitutional solid solutions, so to form an interstitial solid solution 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. This consists of carbon atoms distributed at random in interstitial sites within the face-centred cubic structure of iron to form the phase *austenite*. 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 A1 (face-centred cubic) or A3 (hexagonal) structures there are twice as many tetrahedral sites and the same number of octahedral sites present as metal atoms. In a unit cell of the A1 structure, containing four metal atoms, the octahedral sites lie at the midpoints of each of the cell edges, with a further site at the cell centre (Figure 6.4a). Each site on a cell edge is shared by four cells, and there are 12 cell edges, hence each cell contains $^{1}/_{4} \times 12$ 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 6.4b). Two metal atoms are found per unit cell of the A3 structure, and hence four tetrahedral sites and two octahedral sites per unit cell occur (Figure 6.5a, b). In both structures, all of the tetrahedral sites and all of the octahedral sites have identical geometries.

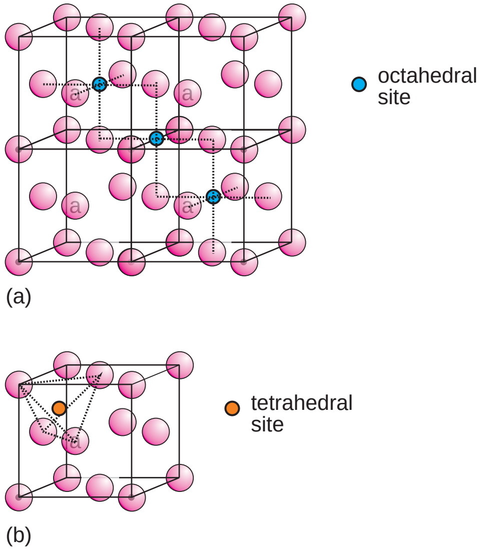


Figure 6.4 (a) Octahedral and (b) tetrahedral sites in the A1 structure. Four unit cells are drawn in (a). Not all equivalent sites are marked.

The A2 (body-centred cubic) structure also has 12 tetrahedral and 6 octahedral sites available. The octahedral sites lie on all of the cube faces (Figure 6.6a), and the tetrahedral sites, also on the cube face, are slightly below them (Figure 6.6b). These octahedral and tetrahedral sites are of slightly different geometry than those in the A1 and A3 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.

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

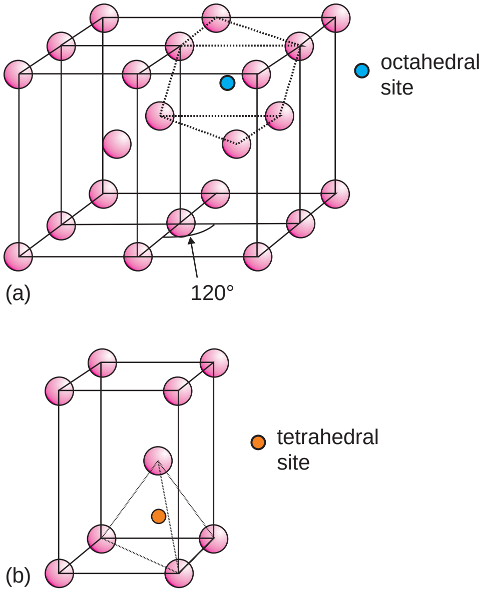


Figure 6.5 (a) Octahedral and (b) tetrahedral sites in the A3 structure. Four unit cells are drawn in (a). Not all equivalent sites are marked.

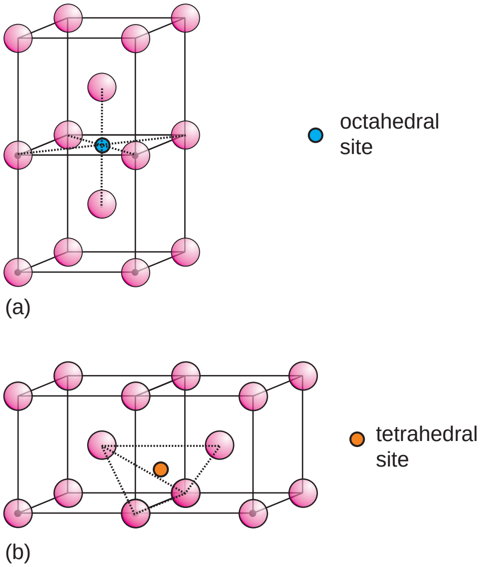


Figure 6.6 (a) Octahedral and (b) tetrahedral sites in the A2 structure. Two unit cells are drawn. Not all equivalent sites are marked.

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.

6.1.4 Metallic glasses

In order to make a metallic glass by supercooling, 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 several 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, (SM) is one or more semimetals such as Si or Ge, and (NM) is one or more nonmetals 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 Vitreloy (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.

6.1.5 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. At a more subtle level, 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. As with all materials, the observed properties arise via the interaction of bonding, crystal structure and microstructure, all of which are consequences of the metallic bond. The free electrons in this bonding model can move throughout the metal under the imposition of only a very low thermal or electrical driving force. The relationship between the resulting effects, thermal conductivity and electrical conductivity, is given by the *Wiedemann-Franz law*:

$$\frac{\text{thermal conductivity}}{\text{electrical conductivity}} = \frac{3Tk_B^2}{e^2}$$

where T is the temperature (K), k_B is Boltzmann's constant and e is the charge on the electron.

The shiny appearance of metals is also due to the free electrons. When light photons strike the metal surface, those near to the Fermi surface can absorb the photons, as plenty of empty energy states lie nearby (see Section 2.3.2). 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.

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 inferior electrical 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 that defects play in crystals. Dislocations 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 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, that 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 nano-scale crystallites, as these can offer enhanced mechanical properties compared with similar alloys with larger grains.

In this short section, only the principal properties of metals and a simple appreciation of the origins of these properties are outlined. Far more detail will be found in later chapters of this book.

6.2 Ceramics

Ceramics are inorganic materials fabricated by a high-temperature chemical reaction. Most ceramics are oxides, but the term is also used for silicides, nitrides, oxy-nitrides, hydrides and other inorganic materials. Ceramics are generally regarded as chemically inert materials that are hard, brittle, thermal and electronic insulators, but important exceptions exist, such as ceramics that show metallic conductivity.

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*, 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, diamond); 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, diamond). Electroceramics are very high purity materials that possess unique electronic properties, varying from insulating to superconducting. Electroceramics form the active element in many gas sensors, temperature sensors, batteries and fuel cells. Ceramic magnets are widely used in motors. Ceramics are also important in fluorescent lighting and as components of computer displays.

6.2.1 Bonding and structure of silicate ceramics

The Earth's crust is composed very largely of silicates, as are the majority of semiprecious gemstones. Natural silicates are minerals that were formed from a complex molten magma and it is therefore not surprising that they are of variable composition. To each mineral is ascribed an *ideal* composition: the composition that it would have if it were homogeneous. *Isomorphous replacement*, in which some cations are replaced by others of similar size, although not necessarily of the same charge, is common in minerals. Thus, the cations Na^+, Mg^{2+}, Ca^{2+}, Mn^{2+} and Fe^{3+} are readily interchangeable, as are the anions O^{2-}, F^- and OH^-. Aluminium, which is to the left of silicon in the periodic table, occupies a special role in silicate chemistry. Aluminium can replace silicon in silicates and does so in a random manner and to an indefinite extent. Isomorphous replacement produces *substitutional* defects in the crystal structure. The mineral *hornblende* provides an illustrative example of isomorphous replacement. The ideal composition of this silicate is $Ca_2Mg_2(Si_4O_{11})_2(OH)_2$. A typical analysis of a naturally occurring sample might well show that up to a quarter of the silicon is replaced by aluminium; most of the Mg^{2+} is replaced by Fe^{2+}, together with smaller amounts of Fe^{3+}, Mn^{2+} and Ti^{4+}; and about a third of the Ca^{2+} is replaced by a mixture of Na^+ and K^+.

Silicon is a small atom with an electronic structure [Ne] $3s^2$ $3p^2$. Silicon lies below carbon in the periodic table, and like carbon, makes use of sp^3-hybrid bonds, which are arranged tetrahedrally. In silicates, each silicon atom is usually linked to oxygen to form a $[SiO_4]$ tetrahedron (Figures 5.31, 5.32). The bonds are very strong and silicon–oxygen tetrahedra are stable and vary little in size. The Si-O distances are always close to 0.162 nm and the 0-0 distances close to 0.27 nm. In terms of an ionic model, the $[SiO_4]$ tetrahedral group has an overall charge of -4, and is written $[SiO_4]^{4-}$.

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 networks. In all these structures, only the vertices of the tetrahedra are shared, never edges or faces.

As a first approximation, silicates can be divided into three groups, described in the following sections.

6.2.1.1 Isolated silicate groups

Three-dimensional silicates that contain isolated silicate groups are often called *ionic silicates* (Table 6.2). The $[SiO_4]$ tetrahedra can form: (a) isolated $[SiO_4]^{4-}$ units; (b) pairs, $[Si_2O_7]^{6-}$; (c) three-membered rings, $[SiO_9]^{6-}$; (d) four-membered rings, $[Si_4O_{12}]^{8-}$; (e) six-membered rings, $[Si_6O_{18}]^{12-}$ (Figure 6.7).[1]

6.2.1.2 Chain or sheet silicates

Silicates that contain chains or sheets of silicate tetrahedra are often called *extended anion silicates*. There are thousands of these compounds, many of which are valuable minerals. Even the simplest structure, a single chain of tetrahedra linked by corners (Figure 6.8a), can adopt several different configurations. The formula 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 6.8b). 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 6.9a,b), and linking the free vertices that lie at the apex of each tetrahedron will lead to double silicate layers (Figure 6.9c).

These latter structures are rather rare in mineralogical terms, but 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 6.10a, 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 6.10c) or two silicate layers to sandwich a hydroxide layer (Figure 6.10d). The single silicate plus hydroxyl layer is found in clays such as *kaolinite* and minerals such as *chrysotile asbestos* (Figure 6.11a). Double silicate sandwich layers enclosing a hydroxyl layer are found in minerals such as *talc* (Figure 6.11b). 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 6.11c).

[1] The figures in this chapter section 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.

Table 6.2 A summary of silicate structures

Structure	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), dmisteinbergite
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	$A1_2(OH)_2Si_4O_{10}$, pyrophyllite (*clays*) $Mg_3(OH)_2Si_4O_{10}$, talc (*clays*)
	$[(Si,Al)_4O_{10}]^{n-} + [OH]^-$	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
Aluminosilicate	$[(Si,Al)_4O_8]$	7–5	$KAlSi_3O_8$, orthoclase (*feldspars*)

6.2.1.3 Network silicates

In these minerals, the $[SiO_4]$ tetrahedra link by all vertices to form a three-dimensional covalent network. This can adopt a number of different conformations, typified by the polymorphs of silica, SiO_2, the commonest of which is quartz (Figure 5.32). When aluminium replaces some of the silicon, the framework takes on a negative charge, compensated for by the insertion of cations into the network. This produces the large group of minerals the *aluminosilicates*, which includes the important *feldspars*, which make up many strong rocks, and the *zeolites* and *ultramarines*, all of which are of considerable industrial importance.

6.2.2 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 structure (Section 5.3.8). 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, able to withstand high temperatures, is used in laboratory furnace-ware. A number of other oxides, including Fe_2O_3 and Cr_2O_3, adopt the same structure. This is most easily described as a hexagonal close-packed array of oxygen ions, with Al^{3+} ions distributed in an ordered fashion over two-thirds of the available octahedral sites (Figure 6.12). 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+}.

Zirconia, ZrO_2, is an important ceramic because it is 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 6.13). In this, each Zr^{4+} ion is surrounded by eight oxide

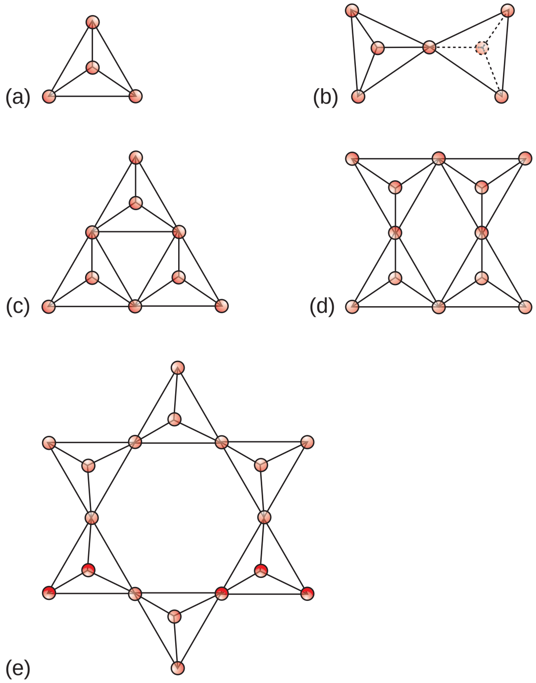


Figure 6.7 Corner-linked $[SiO_4]$ units found in ionic silicates: (a) isolated (SiO_4); (b) (Si_2O_7); (c) (Si_3O_9); (d) (Si_4O_{12}); (e) (Si_6O_{18}).

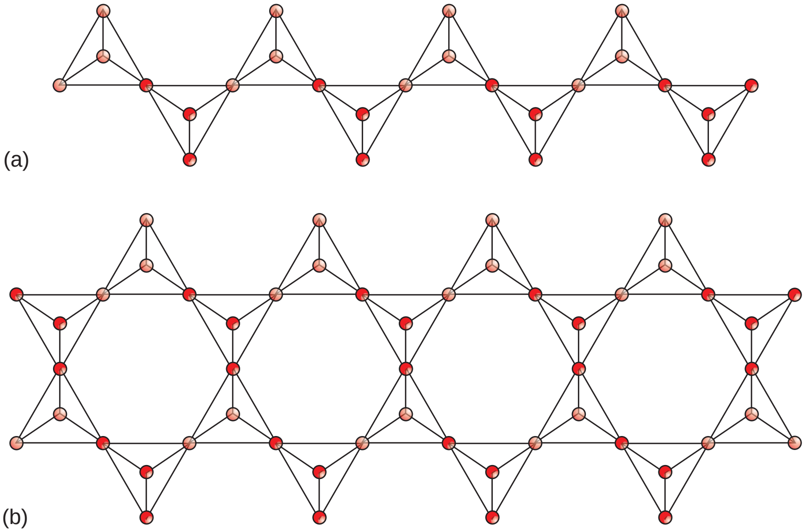


Figure 6.8 (a) Single chain (SiO_3) strings, found in pyroxenes. (b) Double-chain (Si_4O_{11}) strings found in amphiboles.

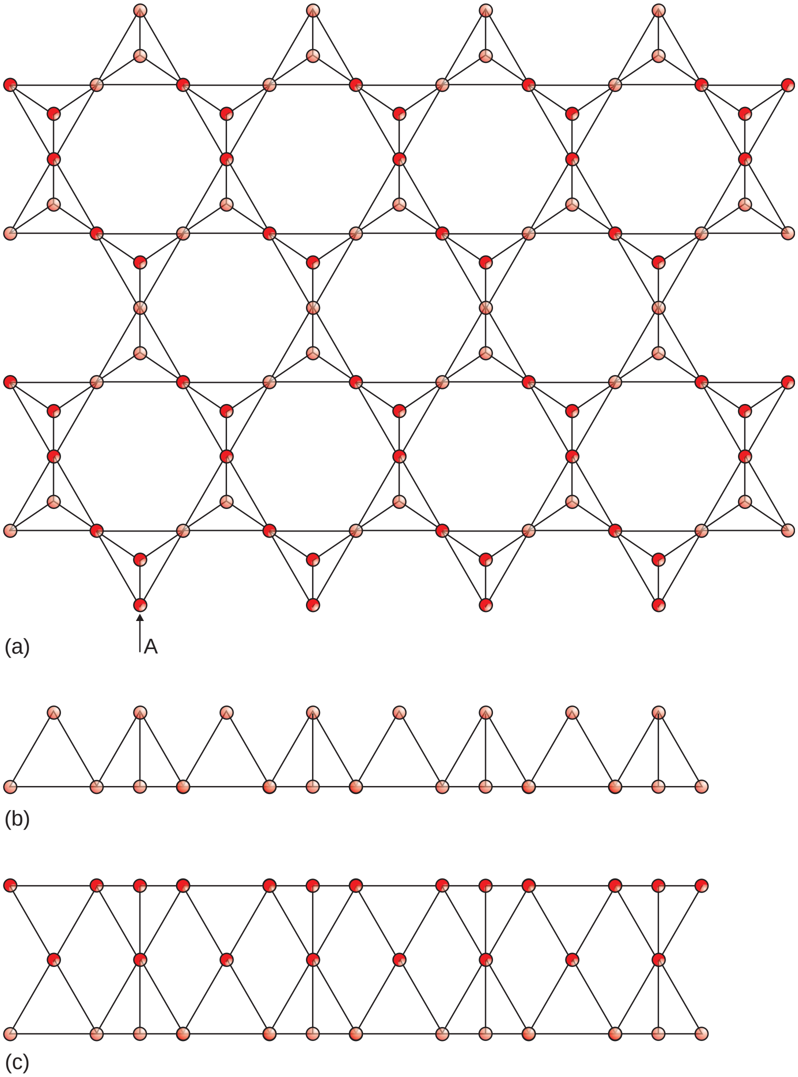


Figure 6.9 (a) A single sheet of corner-linked ((SiO_4)) tetrahedra. (b) The same sheet viewed along A. (c) Double sheets formed by joining two layers of the type shown in (a) by the free tetrahedral vertices, one over the other, viewed along A.

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 three-dimensional checkerboard 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 (Section 3.4.5). The resulting *calcia-stabilised zirconia* remains cubic from room temperature to the melting point, above 2300 °C, allowing problem-free high-temperature use.

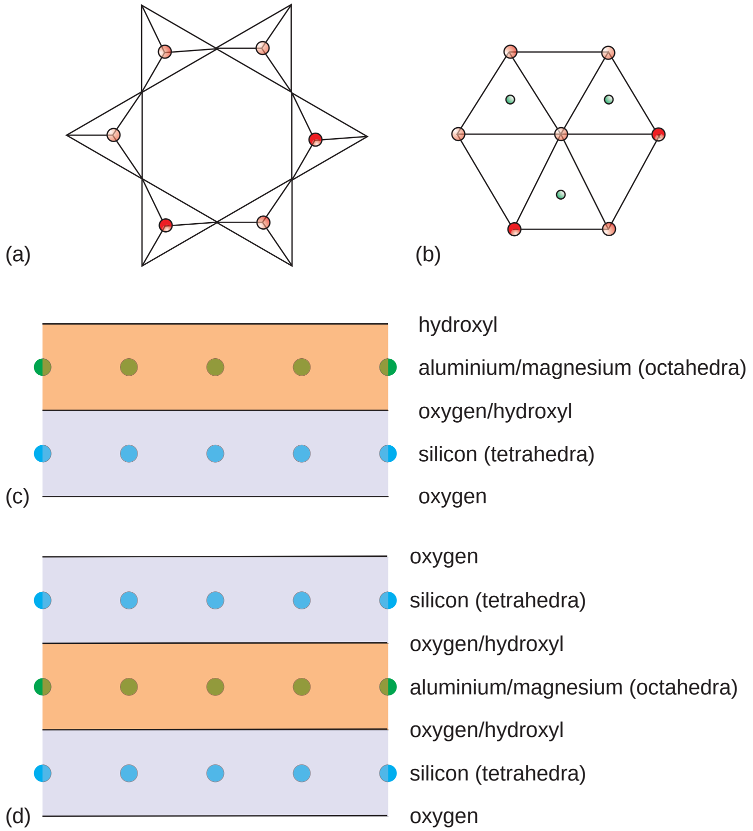


Figure 6.10 The structure of clays and related minerals: (a) a single layer of corner-linked ((SiO_4)) tetrahedra (the upper layer of apical oxygen atoms are drawn as spheres); (b) a basal close-packed layer of oxygen atoms in $Mg(OH)_2$ or $Al(OH)_3$ (large spheres), with cations in the octahedral sites above the layer (small spheres); (c) a composite layer formed by uniting a single silicate layer as in (a) and a single hydroxide layer as in (b); (d) a composite sandwich structure formed by uniting two silicate layers as in (a), with a single hydroxide layer as in (b).

6.2.3 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 and 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. The use of these three major components has led to the name *triaxial whitewares* for these materials.

Clay-based traditional ceramics are first formed into the shape desired, using traditional potter's 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

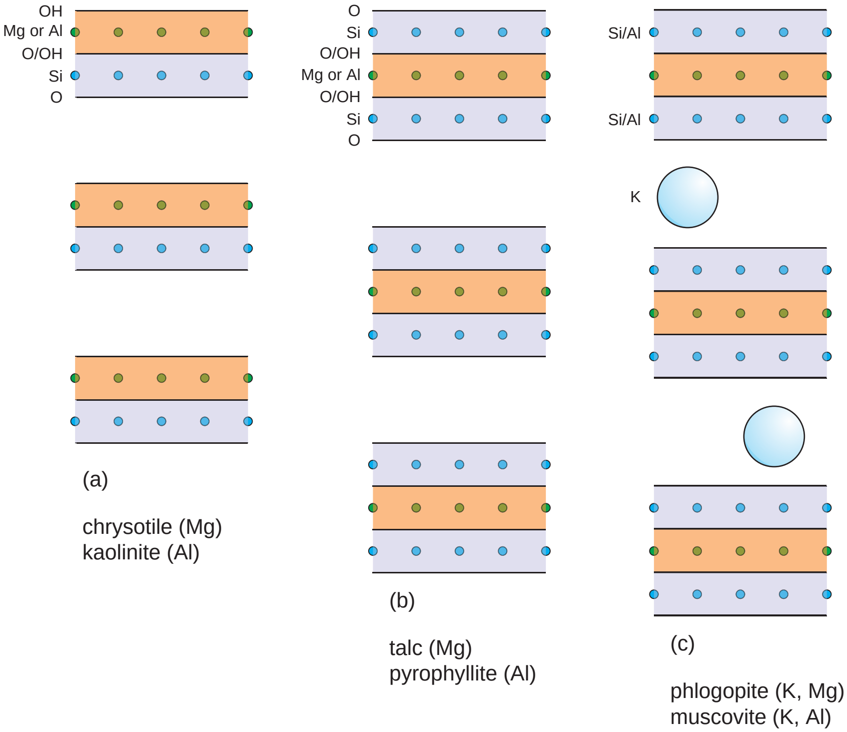


Figure 6.11 Structures of some clay and mica-related minerals, formed from composite silicate–hydroxide layers.

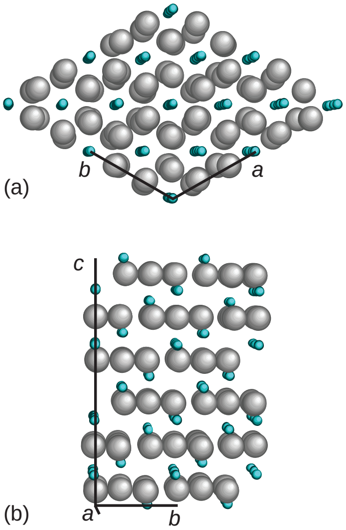


Figure 6.12 The structure of Al_2O_3: (a) projection close to [001]; (b) projection close to [100].

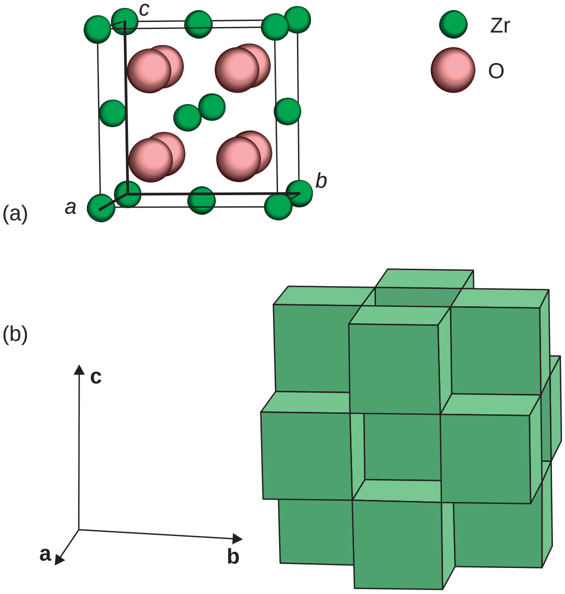


Figure 6.13 The cubic ZrO_2 structure: (a) a single unit cell; (b) the structure as a stacking of (ZrO_8) cubes linked by edge-sharing.

reactions, and the production of an object with a precise size and shape is difficult.

Engineering ceramics and electroceramics are usually pure single phases. Many are made by milling components to produce a fine powder, pressing it 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 (Section 8.1). 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.

6.2.4 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 10.3.1). 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 a source 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.

6.3 Silicate glasses

Glasses are non-crystalline materials, and while most glasses are oxides, specialised glasses such as metallic glasses (Section 6.1.4) are becoming increasingly important. In this section, glass is understood to mean a hard, transparent, fairly strong, corrosion-resistant material, in which the main component is silica, SiO_2 – the *silicate glasses* (Table 6.3). 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.

Silicate glass production marks 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.

Table 6.3 Some silicate glasses

Name	Typical composition	Important property	Principal uses
Soda glass	15%Na_2O : 85% SiO_2	Cheap	Window glazing
Soda-lime glass	72% SiO_2 : 14% Na_2O : 14% CaO	Cheap	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	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

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 19th century, and then with the development of optical fibre-based communication systems towards the end of the 20th century.

6.3.1 Bonding and structure of silicate glasses

The main structural unit in silicate glasses is the [SiO_4] group (Figures 5.31, 5.32, 6.7). The covalent bonds between the central silicon atom and the oxygen atoms are very strong, and both the liquid and solid states of silica and silicates contain large numbers of [SiO_4] tetrahedra. These tetrahedra link to one another by sharing corner oxygen atoms to form discrete or interpenetrating chains and rings. In the solid these are locked in place, while in the liquid they continually change orientation.

To form a crystalline solid, a liquid must first form crystal nuclei. In the case of metals this is extremely easy, as the spherical atoms can quickly pack into arrays. In the case of silicates, the entangled chains are difficult to reorganize into an ordered crystalline arrangement. This process often needs bonds to be broken as well as a rearrangement of tetrahedra. Because of this, nucleation is very slow, and cooling a melt at even slow rates can result in the formation of a glass. The 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 6.14).

A silicate glass is often described as a *supercooled liquid*. However, it is important to keep in

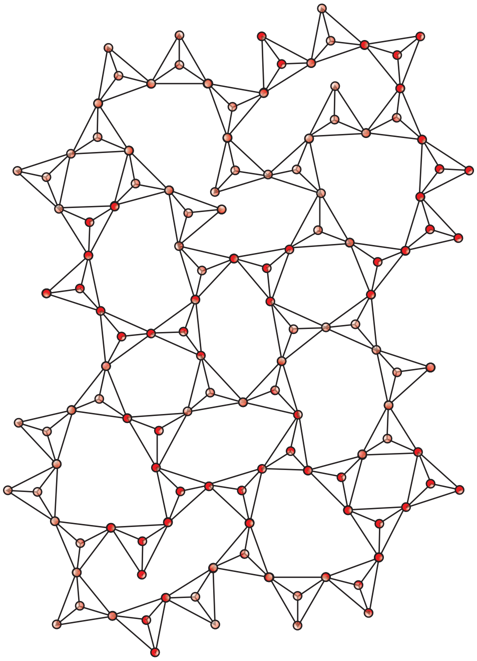


Figure 6.14 The random network model of corner-linked (SiO_4) tetrahedra in a silicate glass.

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 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 which 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 6.15). 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.

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*.

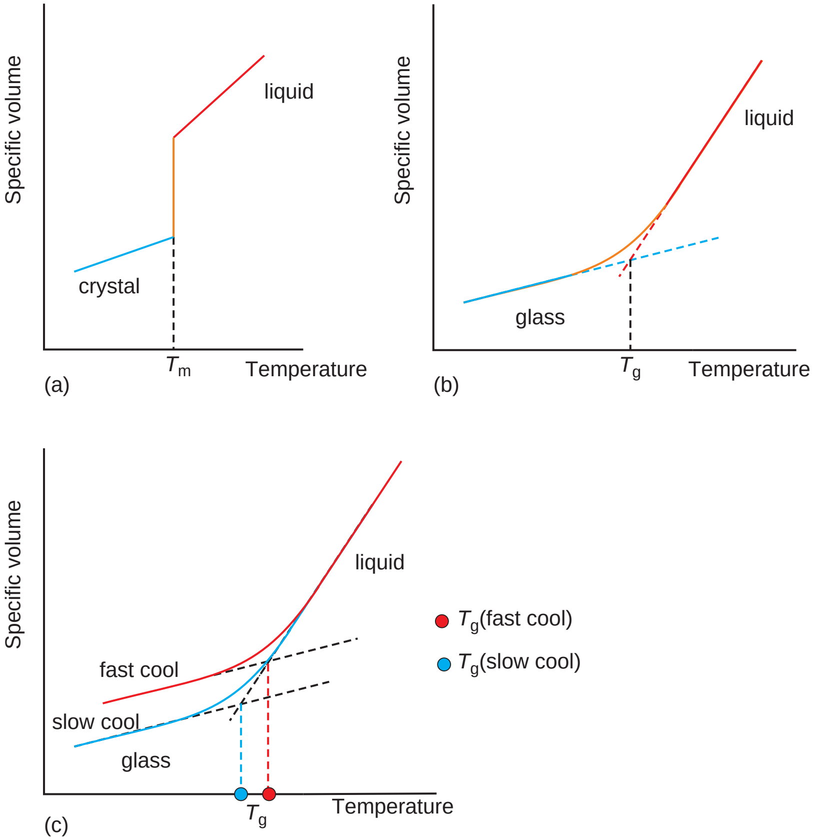


Figure 6.15 Specific volume versus temperature curves: (a) a crystalline solid, with melting point T_m; (b) a glass, with glass transition temperature T_g; (c) the effects of cooling rate on glass transition temperature.

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 usually of 6, are called *intermediates*. Typical examples are Ti^{4+}, Cu^{2+} and Zn^{2+}. Although intermediate ions 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.

6.3.2 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 upon cooling.

The viscosity of a glass, especially the variation of viscosity with temperature (Figure 6.16a) is an important physical parameter, as it defines the temperature regimes over which the glass can be worked (Table 6.4). The *working range* of a glass is the range of viscosities over which normal processing takes place, usually between viscosity[2] 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.

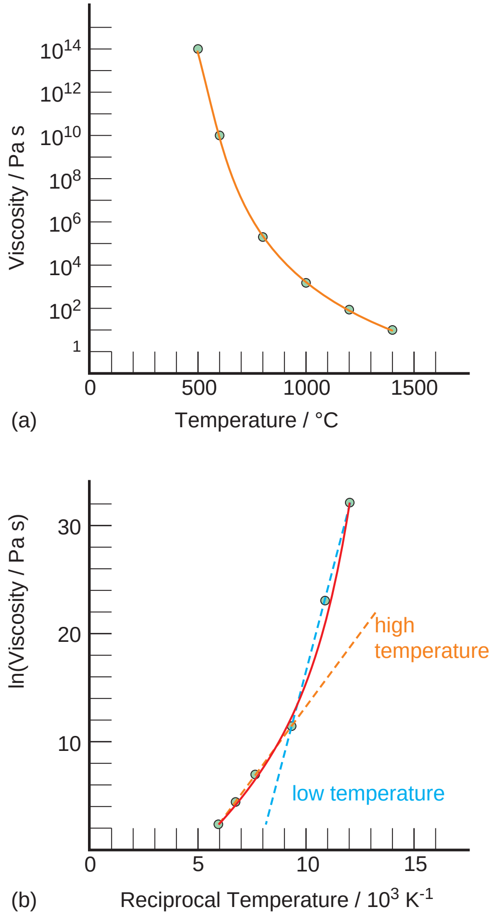


Figure 6.16 (a) Viscosity versus temperature for a typical soda-lime glass. (b) Plot of ln (viscosity) versus reciprocal temperature for the same glass.

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 of these is an Arrhenius equation:

$$\eta = \eta_0 \exp\left(\frac{E}{RT}\right)$$

where η is the viscosity, η_o is a constant, E is the energy for viscous flow, often called the *activation energy*, R is the gas constant and T the temperature (K). This equation describes activated processes in which transformation is possible only when the participants overcome an energy barrier. In such reactions, Arrhenius behaviour is confirmed by a

[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.

Table 6.4 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

plot of ln η versus $1/T$:

$$\ln \eta = \ln \eta_0 + \frac{E}{RT} \qquad (6.1)$$

The plot should be linear with a slope of E/R. Materials that conform to equation (6.1) are said to exhibit *Arrhenian behaviour*. This is not generally found to occur for glasses, which show *non-Arrhenian* behaviour. The plot of ln η versus $1/T$ is frequently a smooth curve, although the low- and high-temperature parts of the graph taken separately may offer reasonable fits to equation (6.1) (Figure 6.16b). In this case the high-temperature slope is usually lower than that of the low temperature part.

Several other equations have been suggested to overcome the shortcomings of the Arrhenius equation. The most widely used is the *Vogel-Fulcher-Tamman equation* (also called the *Vogel-Fulcher* equation):

$$\eta = A \exp\left(\frac{B}{T - T_0}\right)$$

where A, B and T_0 are empirical constants, and T is the temperature (K). This leads to the relation:

$$\ln \eta = \ln A + \left(\frac{B}{T - T_0}\right)$$

Viscosity is often represented on an *Angell plot* (Figure 6.17) of log η versus T_g/T, which is simply a scaled version of an Arrhenius plot (Figure 6.16b). 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 interatomic interactions present contributing to the viscosity, including weak bonding such as van der Waals and hydrogen bonding, but is not simply an

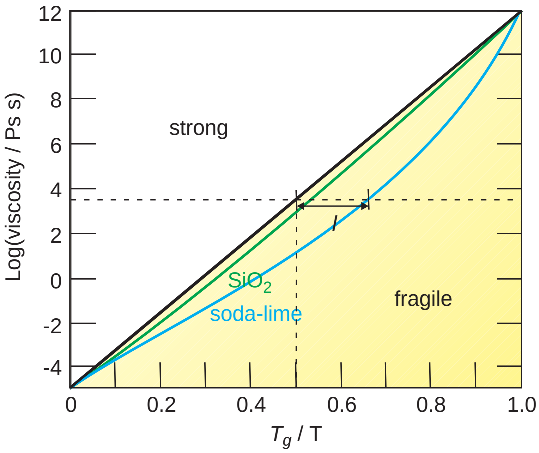


Figure 6.17 Angell plot: strong liquids are to the upper left and fragile liquids to the lower right; curves are for silica and a soda-lime glass. The line l is a measure of the departure of a glass (here soda-lime glass) from Arrhenian behaviour, used in determination of the kinetic fragility, $F_{½}$.

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.

The fragility has been quantified by measuring the slope of the Arrhenius plot as the temperature tends towards the glass transition temperature, T_g:

$$m = \left(\frac{\mathrm{d} \log \eta}{\mathrm{d}(T_g/T)} \right)_{T \approx T_g}$$

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:

$$F_{1/2} = 2\left(\frac{T_g}{T} - 0.5 \right)$$

where T_g/T is the value where a horizontal line drawn halfway up the log η cuts the requisite glass curve. The quantity in brackets corresponds to the value l (for soda-lime glass, Figure 6.17) 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.

6.3.3 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 *Griffiths flaws* (Section 10.3). Under tension the flaws generate cracks allowing the material to fracture. Many of the 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 sheets are being pulled in and so are 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 can be seen as coloured shapes in the windshield of a car driven in sunlight when the driver wears sunglasses with polarising lenses.)

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 principal 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.

6.3.4 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 6.18). In addition, the crystallisation must occur in a glass that is viscous enough not to sag or distort during the transformation.

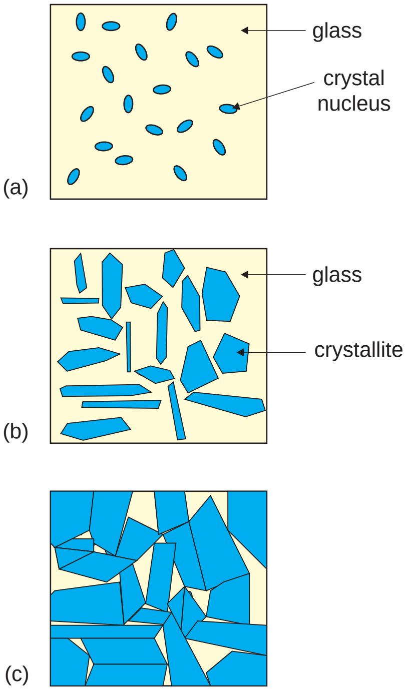


Figure 6.18 Crystal nucleation (a) and crystal growth (b, c) in a glass ceramic.

The two most important factors in glass-ceramic production are the composition of the melt and the microstructure of the final product. These are interrelated, 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 in 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 °C and 870 °C, will grow. The presence of 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.

6.4 Polymers

Polymers are long chain-like macro (giant) molecules made by the linkage of large numbers of small repeating molecules called *monomers*. Short chain lengths formed in the course of synthesis or degradation of polymers are called *oligomers*. The majority of polymers, and the only ones considered here, are compounds of carbon. Within this group the carbon backbone of the molecule can be linear, branched or cross-linked, and a great variety of chemical groups can attached to the chain spine. Because of this, polymer nomenclature can be complicated. Initially polymeric materials were characterised using common names, such as Bakelite, named after the inventor, Baekeland. As the number and complexity of synthetic polymers increased, this *ad hoc* method became confusing and now two main polymer-naming systems are in place. These are *structure-based* and *source-based* names. Structure-based names use the formal naming procedure of organic chemistry to specify the chain structure unambiguously. Unfortunately these are complex and mean little to the non-professional. For example, nylon 66 (see later) is formally named poly [imino (1,6-dioxo-1,6-hexanediyl) amino-1,6-hexanediyl]. Source-based names describe the materials that are used in polymer formation, for instance poly(vinyl acetate), making it clear that the source material is vinyl acetate. This nomenclature is more familiar, and as the polymers discussed here are well known, source-based names will be used, as well as common names such as neoprene when these are adequate.

Polymers are very widespread and can be synthetic, such as nylon, or natural, for example, rubber. They form vital components of living organisms, and the most important molecule, DNA, is a polymer of amino acids. Colloquially, polymers are often called plastics. More precisely, plastics are sometimes defined as polymers that can be easily formed at low temperatures, and sometimes as a pure polymer together with a non-polymeric additive, which may be solid, liquid or gas.

There are three main divisions of polymeric materials. *Thermoplastic materials* can be melted and reformed a number of times. *Thermosetting materials* can only be formed once; they cannot be remelted. *Elastomers* can be deformed a considerable amount and return to their original size rapidly when the force is removed.

The properties of polymers depend both upon the details of the carbon chain of the polymer molecule and upon the way in which these chains fit together. The chain form can be remarkably complex when branching and substituents are taken into account. In addition, the chains may be ordered to form crystals, or tangled in an amorphous mass. Amorphous polymers tend not to have a sharp melting point, but soften gradually. These materials are characterised by a glass transition temperature, T_g, and in a pure state are often transparent.

Although polymers are associated with electrically insulating behaviour, the increasing ability to control both the fabrication and constitution of polymers has lead to the development of polymers that show metallic conductivity superior to that of copper (Section 13.4), and to polymers that can conduct ions well enough to serve as polymer electrolytes in batteries.

6.4.1 Polymer formation

Polymers were once mainly grouped in terms of the overall chemistry of the polymerisation reaction. Molecules that simply added together were called *addition polymers*, and those that joined and at the same time eliminated one or more small molecules were called *condensation polymers*. These designations, which derived from organic chemistry, have now largely been replaced by groupings that reflect the *mechanism* of the polymerisation rather than the overall chemical reaction.

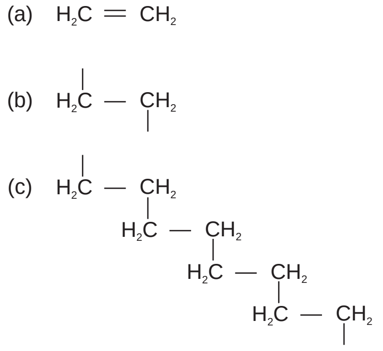
In order to link a large number of monomers it is necessary for each end of the molecules to be made chemically reactive. Generally this involves breaking chemical bonds to yield two reactive half-bonds. The simplest starting point for a discussion of this process is the monomer molecule ethene (ethylene):

$$CH_2 = CH_2$$

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* (Scheme 6.1). The chemical formula of the resulting polymer, called polyethylene or polythene, is:

$$[CH_2\text{-}CH_2]_n$$

where n takes a value of several thousand. (Note that the industrial preparation of polythene, and of all the polymers described here, is quite different from the scheme illustrated.) The polymer chain is constructed from (CH_2) units. In these, the carbon atoms are bonded to two hydrogen atoms and two carbon atoms using strong sp^3-hybrid bonds (Figure 6.19). 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.



Scheme 6.1 The polymerisation of ethene (ethylene) schematic: (a) a single monomer molecule; (b) double bond opening; (c) monomer linkage to form the polymer chain.

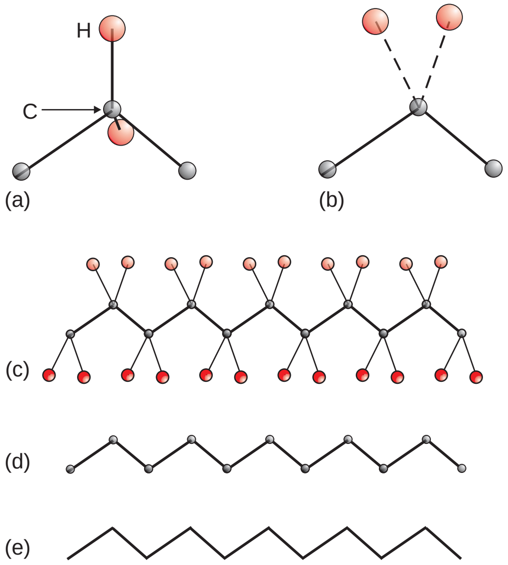


Figure 6.19 (a, b) The structure of the CH_2 unit in a polymer chain, in two orientations. The four bonds arising at the carbon atom are arranged tetrahedrally. (c) A chain of linked CH_2 units. (d, e) Representations of the chains with H atoms omitted.

Table 6.5 Addition polymers

Monomer source	Polymer formula	Names	Uses
$CH_2{=}CH_2$ (ethene, ethylene)	$(CH_2\text{-}CH_2)_n$	Polyethylene, polythene, PE	Squeeze bottles, food bags, dishes, insulation, coatings
$CH_2{=}CH\ Cl$ (vinyl chloride)	$(CH_2\text{-}CHCl)_n$	Poly(vinyl chloride), PVC	Pipes, floor covering, insulation, adhesives, films, credit cards
$CH_2{=}CCl_2$ (vinylidene chloride)	$(CH_2\text{-}CCl_2)_n$	Poly(vinylidene chloride)	Food wraps, fibres, cling film
$CH_2{=}CH\ CH_3$ (propylene)	$(CH_2\text{-}CH\ (CH_3))_n$	Polypropylene, PP	Pipes, valves, carpets
$CH_2{=}CH\ C_6H_5$ (styrene)	$(CH_2\text{-}CH\ (C_6H_5))_n$	Polystyrene, PS	Jugs, cups, packaging, styrofoam, appliance parts
$CH_2{=}CH\ CN$ (acrylonitrile)	$(CH_2\text{-}CH\ (CN))_n$	Polyacrylonitrile, PAN, Orlon, Acrilan	Fabrics, carpets, high-impact plastics
$CH_2{=}CH\ COOCH_3$ (vinyl acetate)	$(CH_2\text{-}CH\ (CH_3\ COO))_n$	Poly(vinyl acetate), PVA	Wood adhesives, paper coatings, latex paints
$CF_2{=}CF_2$ (tetrafluoroethene)	$(CF_2\text{-}CF_2)_n$	Poly(tetrafluoroethylene), PTFE, Teflon	Non-stick coating, electrical insulation, bearings
$CH_2{=}C(CH_3)COOCH_3$ (methyl methacrylate)	$(CH_2\text{-}C(CH_3)COOCH_3)_n$	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	Tyres, hoses, pond liners

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 6.5). 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:

$$R - COOH + R' - OH \rightarrow R - COO - R' + H_2O$$

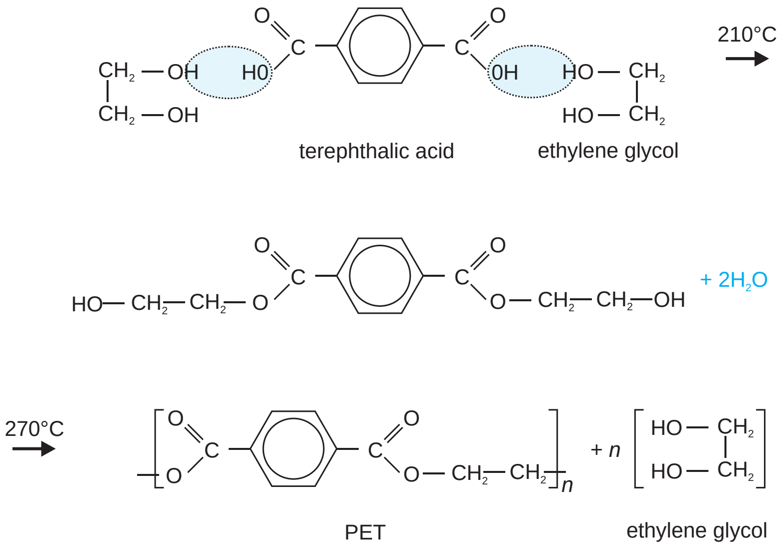
where R and R′ represent different carbon chains. In molecules with only one acid or hydroxyl group, as written, the reaction stops after the first step. In order to create a polymer, the monomers of condensation polymers need *two* reactive groups on each monomer.

The polyesters, which generally start from terephthalic acid and the alcohol ethylene glycol, are an important group of polymers made by reaction between acid groups and hydroxyl groups (Scheme 6.2). The product, a polyester called poly (ethylene terephthalate), or PET, is widely used for shatterproof bottles.

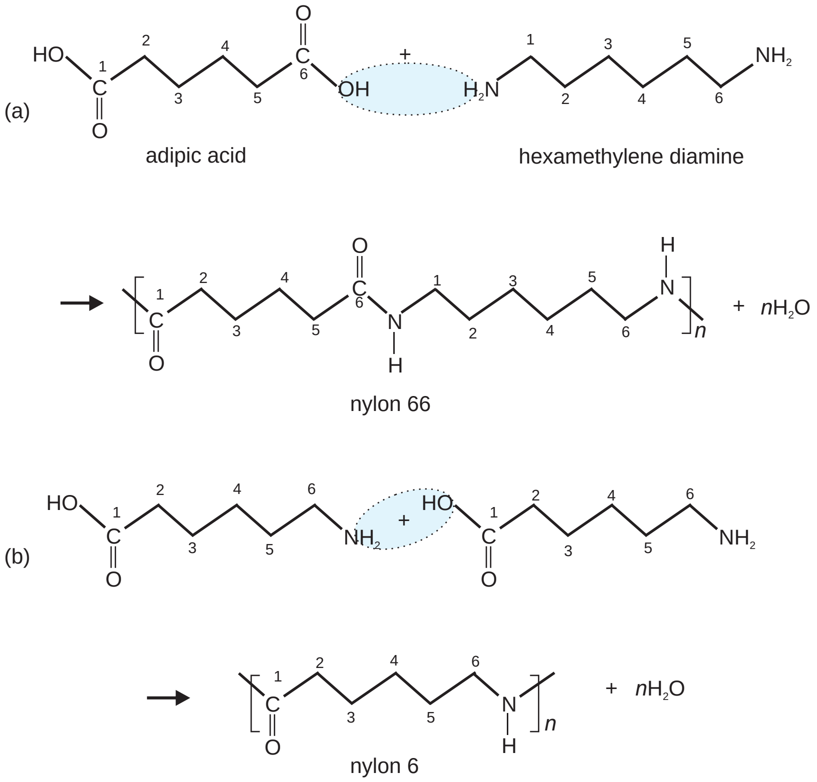
A group of polymers that form in a similar way are the thermoplastic polyamides, better known as *nylons*. The principal reaction is between an acid group, —COOH, and an amide group, $—NH_2$:

$$R - COOH + R' - NH_2 \rightarrow R - CON - R' + H_2O$$

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 of the commonest type of nylon, nylon 66, is an example (Scheme 6.3a). The name, nylon 66 indicates that



Scheme 6.2 The production of the polyester PET schematic. The linking of the terephthalic acid and ethylene glycol molecules occurs at 210 °C, and polymerisation and regeneration of ethylene glycol at 270 °C.



Scheme 6.3 Formation of nylon schematic: (a) nylon 66; (b) nylon 6. The carbon skeleton of the polymer chain is drawn as a zigzag and the number of carbon atoms in the polymer chain between nitrogen atoms is indicated.

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 (Scheme 6.3b). (Note that nylon 6 is not actually made from the amino acid shown, but from ε-caprolactam.)

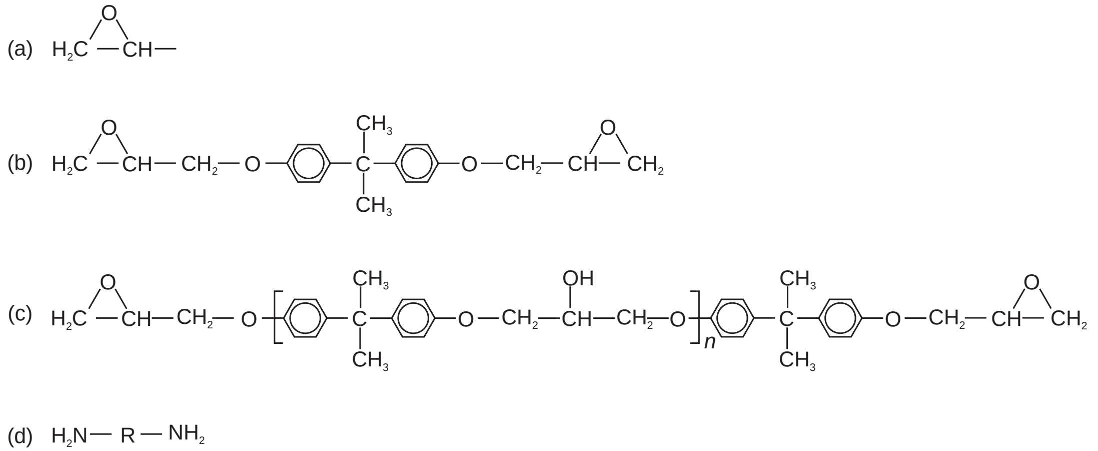
Nylon 66 is an example of an *even–even* nylon, sometimes just abbreviated to *even*. It is easy to envision other even–even polymers, such as nylon 44. Similarly, it is possible to think of *odd–odd*, or just *odd*, nylons, such as nylon 55. Although these nylons have very similar chemical properties, they differ in important electrical aspects and the nature of the chain determines whether these plastics can be used to make piezoelectric components (Section 11.2.3).

A rather similar chemical reaction produces widely-used thermosetting polymers typified by epoxy resins. In this group of materials the reactive *epoxide* group is opened and used to link monomers (Scheme 6.4a). Epoxy resin adhesives normally come as two-part mixes. The resin component contains small and medium-sized molecules with an epoxy group at each end, called di-epoxy molecules (Scheme 6.4b,c). The resin is set by adding a cross-linking agent or *catalyst* which is a diamine (Scheme 6.4d) to join 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.

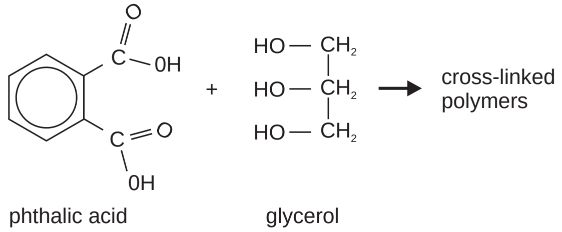
All cross-linked polymers tend to be difficult to disrupt, and they can be regarded as one giant molecule. This makes them of use when durability is needed. Many of the polymers mentioned above can be modified so as to form extensive cross-linking by increasing the number of reactive groups on the monomers. For example, the polyester formed by phthalic acid and glycerol (Scheme 6.5) can cross-link to form a polymer that is used in bake-on car paints.

6.4.2 Microstructures of polymers

Although the chemical make-up of polymers influences properties, by far the most important aspect is that of microstructure. The most significant features



Scheme 6.4 Epoxy resins: (a) the epoxy group; (b) a small diepoxy molecule; (c) a small polymer molecule (with *n* up to about 25) found in the resin part of a two-part epoxy adhesive mix; (d) a diamine linking group (catalyst). R represents a short chain of CH_2 groups.



Scheme 6.5 Reaction of molecules with several active groups can give rise to cross-linked polymers.

of polymer microstructure are chain length, chain branching, chain side-groups (which contribute to chain stiffness), and the 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.

6.4.2.1 Molar mass

Polymers consist of long chains of varying length that *cannot* be characterised by a constant molar mass. Indeed, cross-linked polymers can be thought of as a single molecule, so that the molar mass is the total mass of the object. 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 6.20a). Generally, the distribution departs from a simple bell shape (Figure 6.20b) and the form of the curve is often characteristic of a particular reaction mechanism.

In order to quantify the chain length, the molar mass must be defined statistically. The *number average molar mass*, M_n, is given by:

$$\begin{aligned} M_n &= \sum_i x_i M_i \\ x_i &= \frac{n_i}{\sum_i n_i} \end{aligned}$$

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.

$$\begin{aligned} M_w &= \sum_i w_i M_i \\ w_i &= \frac{n_i M_i}{\sum_i n_i M_i} \end{aligned}$$

where w_i is the mass fraction of each type of molecule present within the chosen molar mass range.

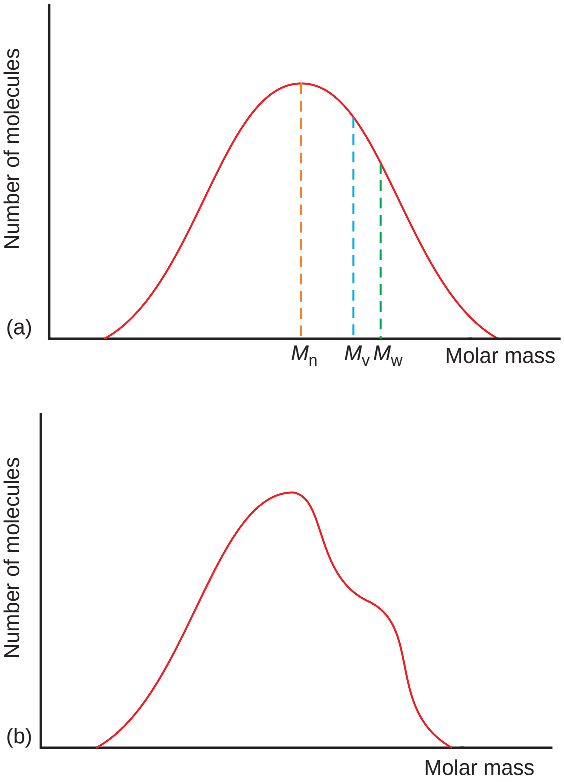


Figure 6.20 Molecular mass of a polymer. (a) Idealised distribution: M_n is the number average molecular mass, M_v is the viscosity molecular mass and M_w the weight average molecular mass. (b) Example of a real mass distribution.

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 (Figure 6.20a).

The degree of polymerisation, N, is the number of monomer units in an average chain. It 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.

$$N_n = \frac{M_n}{m} \quad N_w = \frac{M_w}{m} \quad N_v = \frac{M_v}{m}$$

6.4.2.2 Chain structure

Polymerisation involves the breaking and reforming of large numbers of chemical bonds, and this process results in a number of different molecular structures (Figure 6.21). 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.

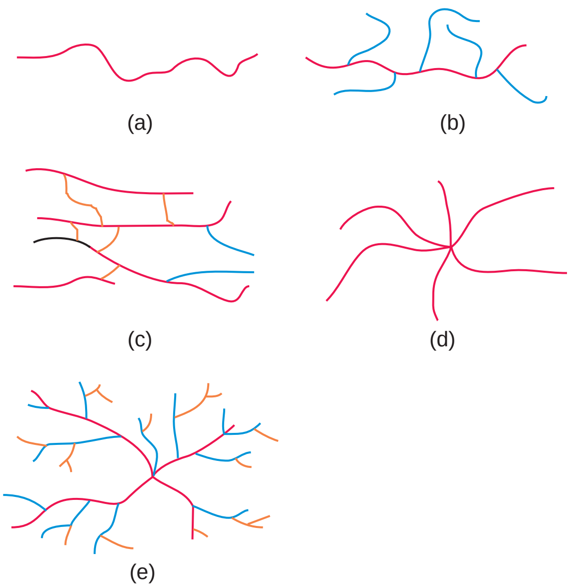


Figure 6.21 Polymer chain geometries: (a) linear; (b) branched; (c) cross-linked; (d) star; (e) dendrimer.

Polyethylene (polythene) is a long-chain polymer of 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.

6.4.2.3 Crystal structure

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 (Figure 6.22a). The folded chains aggregate into blocks that have a regular structure similar to a crystal. This unit of microstructure is called a *lamella* (Figure 6.22b). The lamellae are not usually made from a single folded chain, but are formed by a variety of neighbouring chains (Figure 6.22c). 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 6.23).

The degree to which a polymer can crystallise depends upon the details of the chain. For example,

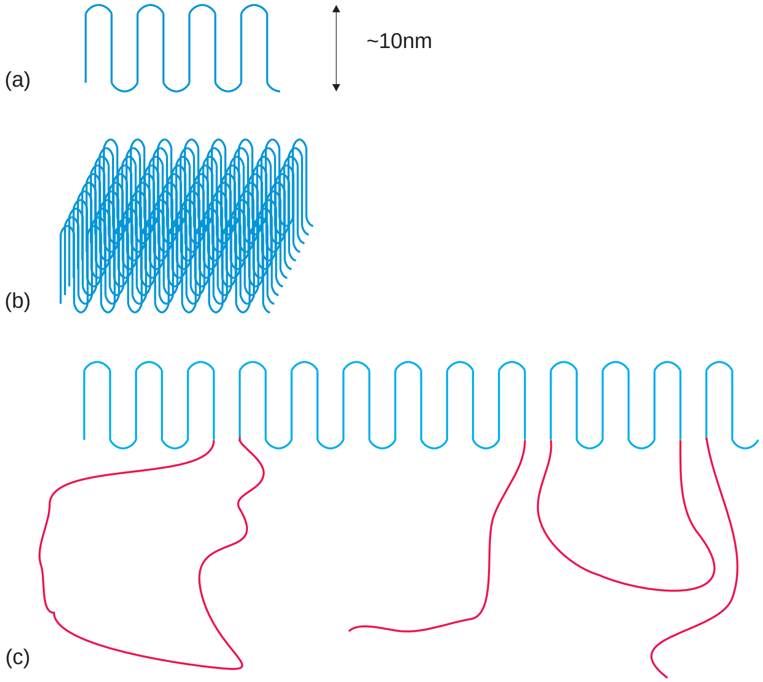


Figure 6.22 (a) Polymer chains in polyethylene fold back on themselves approximately every 10 nm. (b) Folded chains aggregate to form a lamella. (c) Lamellae can contain more than one polymer chain, or a chain folded back into the arrangement.

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.

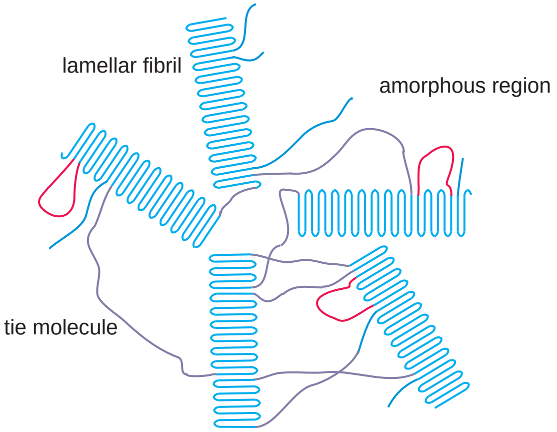


Figure 6.23 Schematic structure of a spherulite in a polymer such as polyethylene.

The degree of crystallinity of a 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 5.2.2). The fraction of crystalline polymer, x_c, is:

$$x_c = \frac{\rho_s - \rho_a}{\rho_c - \rho_a}$$

where ρ_s is the density of the sample, ρ_c is the density of the crystals and ρ_a the density of completely amorphous polymer.

6.4.2.4 *Tacticity*

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 the 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 6.24). 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.

6.4.2.5 Cross-linking

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 as epoxy resins are heavily cross-linked into a hard mass. Cross-linking is the key to elastomeric properties (Section 6.4.4).

6.4.2.6 Copolymers

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

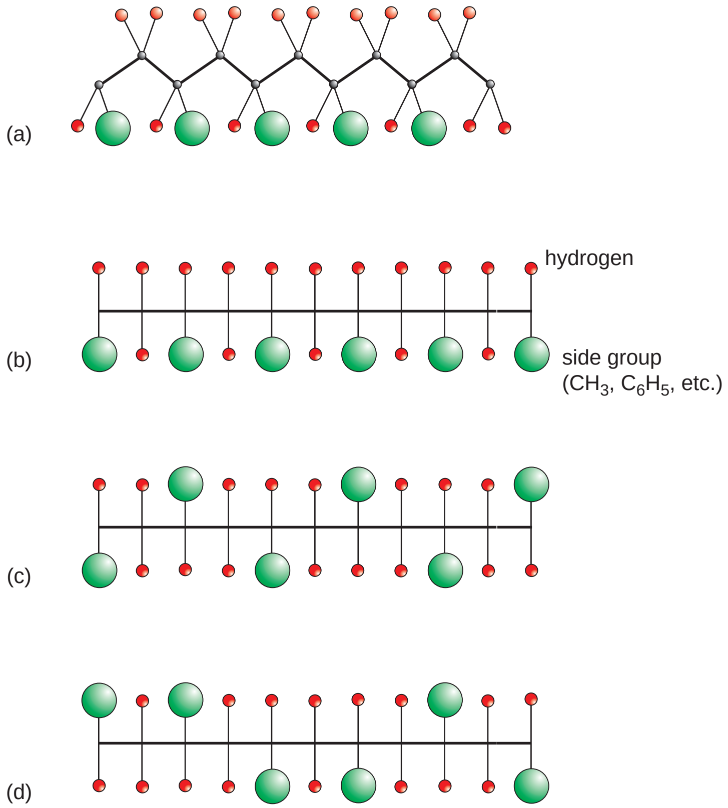


Figure 6.24 Polymer chains: (a) perspective view of an isotactic polymer chain; (b) projection of (a) from above; (c) a syndiotactic polymer chain depicted as in (b); (d) an atactic polymer chain depicted as in (b).

extensive blocks of just one or other polymer, or one polymer can attach as branches to the main chain of the other (Figure 6.25). 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.

6.4.3 Production of polymers

Because the properties of polymers depend so critically upon microstructures, the manufacture of polymers is a highly skilled undertaking. Much of the evolution in the properties of plastics can be attributed to the improvements in preparation methods. 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 (Scheme 6.6). 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.

6.4.3.1 Initiation, propagation and termination

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

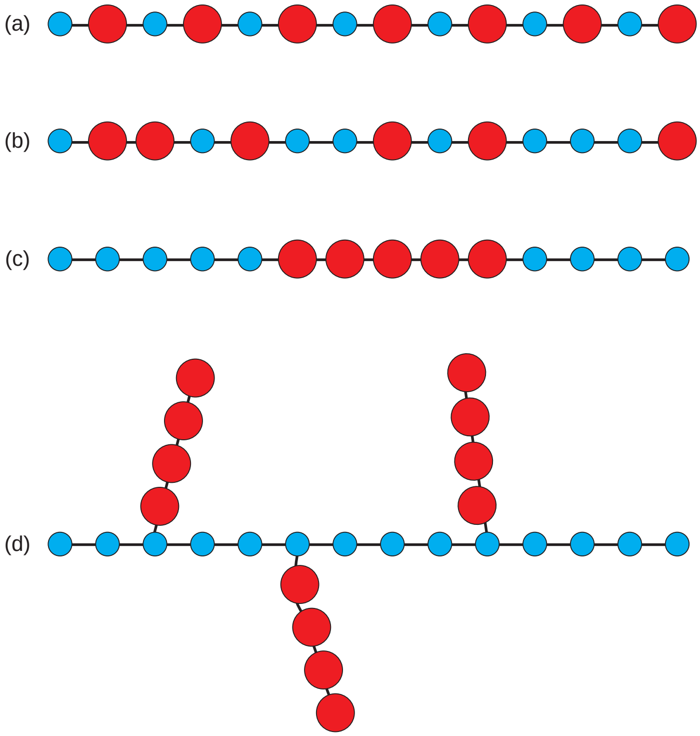
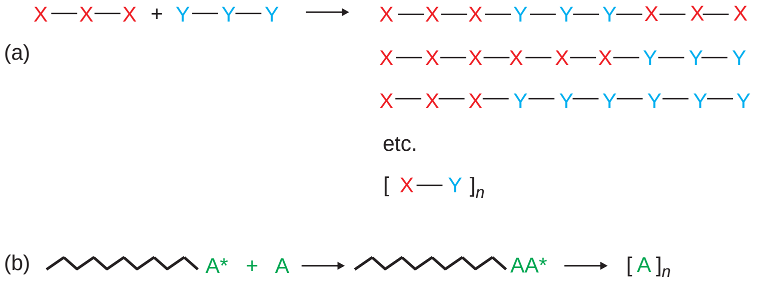


Figure 6.25 Schematic polymer geometry: (a) alternating copolymer; (b) random copolymer; (c) block copolymer; (d) graft copolymer.



Scheme 6.6 Polymerisation mechanisms: (a) in step growth, short chains link to give a variety of sequences; (b) in chain growth, new molecules add to one end of a growing chain. A^* represents a free radical or similar active centre.

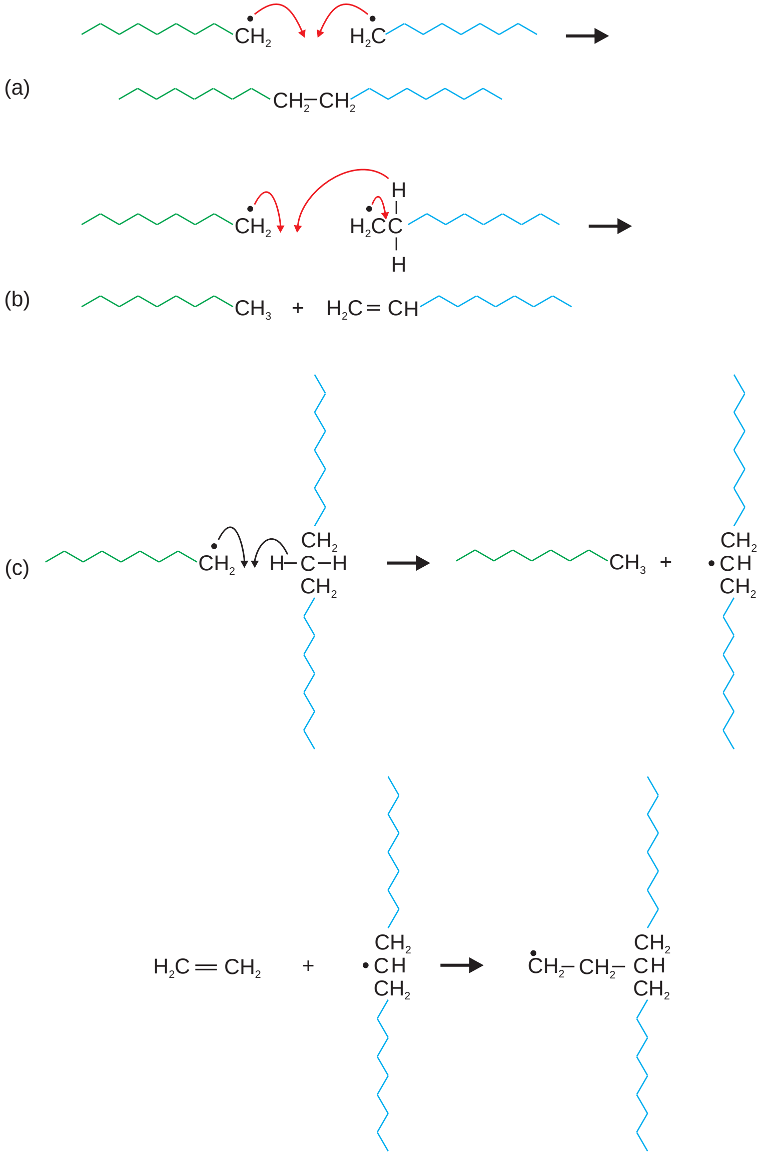
or high-energy radiation such as ultraviolet light. These processes, which are not reproducible enough for industrial production, contribute to the degradation of polymers in normal use. Industrially, the initiation stage is achieved by mixing a wide variety of active molecules with the monomers.

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. 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 (Section 6.4.3.2).

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 (Scheme 6.7a). A second mechanism, *disproportionation*, allows one growing chain to take a hydrogen atom from another chain. The result is that one chain terminates with a $–CH_3$ group and one in a double bond (Scheme 6.7b). 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* (Scheme 6.7c). This is a very common occurrence in polyethylene polymerisation and results in the highly branched low-density (LD) polythene described above. In order to make much stronger non-branched high-density (HD) polythene, different preparation methods need to be used.

6.4.3.2 Metallocene catalysis

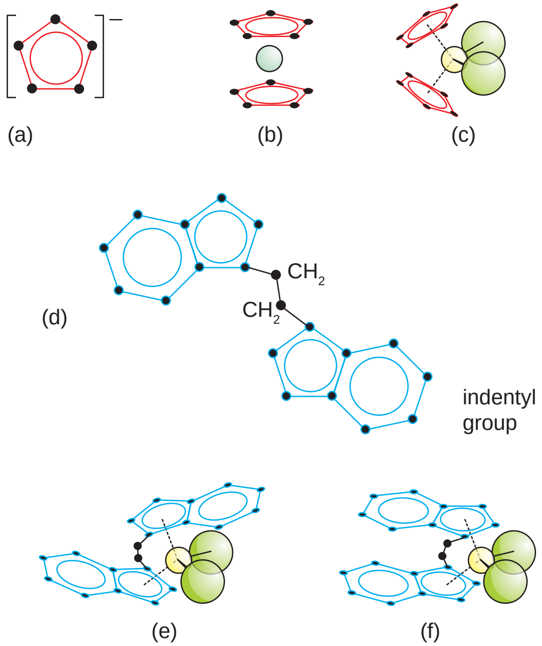
Metallocenes are derived from the cyclopentadiene anion, $(C_5H_5)^-$ (Scheme 6.8a). These are stable units, in which a delocalised orbital lies above and below the plane of the pentagon of carbon atoms, allowing strong chemical bonds to form with metal cations. The first metallocene investigated, *ferrocene*, contained Fe^{2+} sandwiched between the anions (Scheme 6.8b). At present, catalysts for polymer production are derived from a molecule called *zirconocene*, containing Zr^{4+} cations (Scheme 6.8c). 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



Scheme 6.7 Chain termination: (a) two chains meeting; (b) disproportionation; (c) disproportionation leading to branching.

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.

In the molecule used to produce polypropylene (Scheme 6.8d), 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 (Scheme 6.8e), 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



Scheme 6.8 Metallocene catalysis: (a) carbon atom framework of $[C_5H_5]^-$; delocalised π molecular orbitals above and below the plane of the carbon atoms are represented by the circle; (b) ferrocene, in which Fe^{2+} is sandwiched between two $[C_5H_5]^-$ anions; (c) zirconocene, in which Zr^{4+} is bonded to two $[C_5H_5]^-$ anions and two Cl^- anions; (d) two identyl groups linked by two CH_2 groups; (e) molecular geometry required to produce isotactic polypropylene; (f) molecular geometry required to produce atactic polypropylene.

(Scheme 6.8e,f), the approach of the monomers is variable, and atactic polypropylene is produced.

6.4.3.3 Free radical polymerisation

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 (Scheme 6.9a,b). These decompose to form extremely reactive free radicals, which attack double bonds, thus initiating polymerisation (Scheme 6.9c). 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 (Scheme 6.9d,e). The reactions can be written as:

$$R^\bullet + CH_2 = CHX \rightarrow R - CH_2 - CHX^\bullet$$

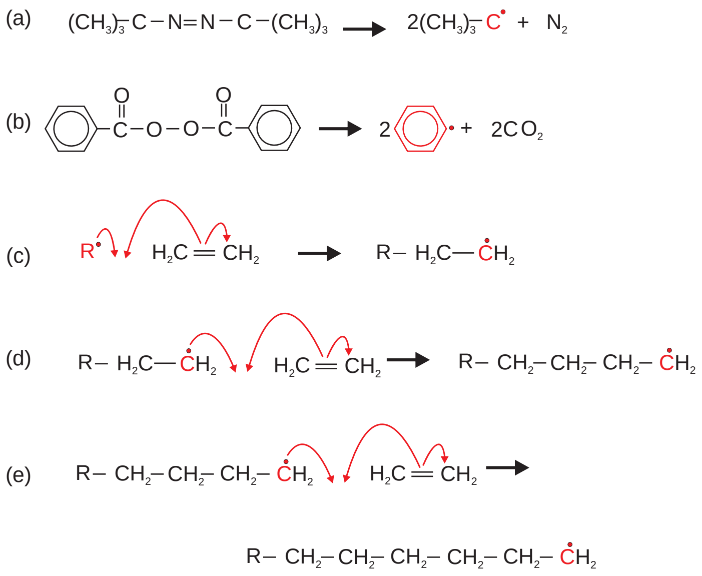
$$R - CH_2 - CHX^\bullet + CH_2 = CHX \rightarrow R - CH_2 - CHX - CH_2 - CHX^\bullet$$

where $R^\bullet$ 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.

6.4.4 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 of being 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 6.26a). On being stretched, the chains tend to partly align parallel to one another (Figure 6.26b). 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



Scheme 6.9 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.

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.

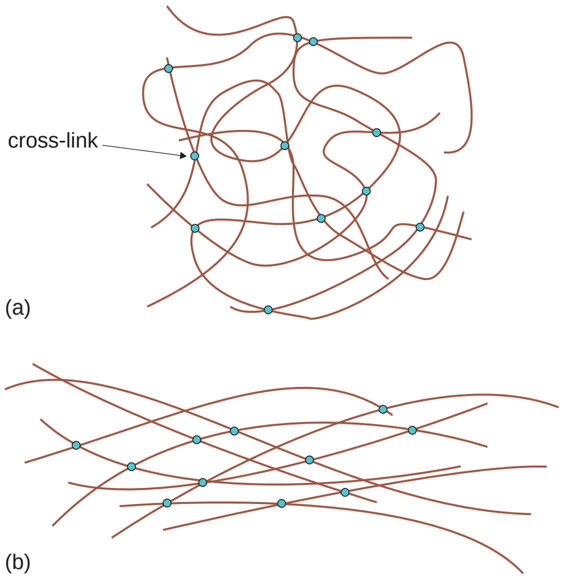
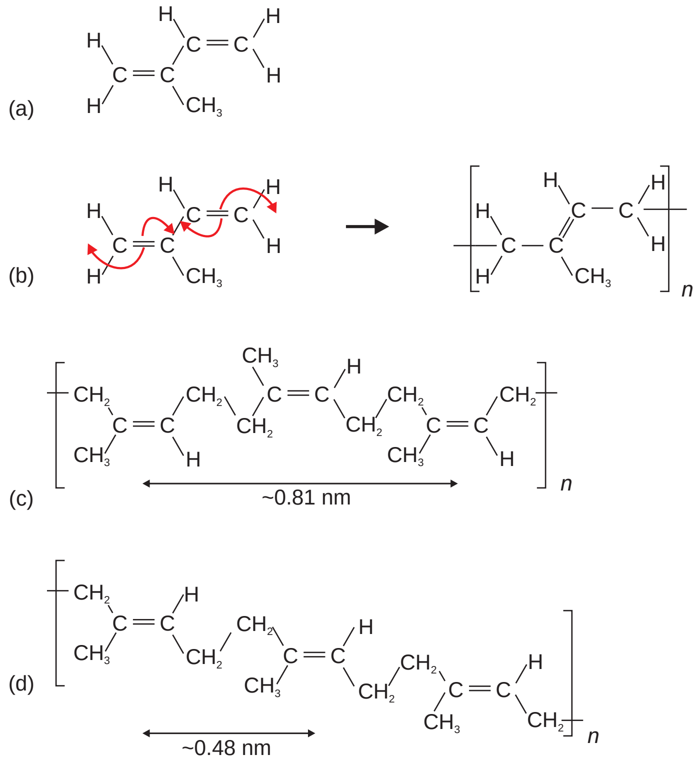


Figure 6.26 Elastomer (schematic) in (a) an unstretched, and (b) a stretched state.

The best-known elastomer is natural rubber, polyisoprene. Isoprene (Scheme 6.10a) is a liquid at room temperature, which polymerises readily to give the elastomer polyisoprene (Scheme 6.10b). 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 (Scheme 6.10c), 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* (Scheme 6.10d), is also found in nature. It is harder and finds use in golf balls and dentistry.

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 posses the typical elastomer properties of snap-back when stretched.



Scheme 6.10 Rubber: (a) the structure of isoprene; (b) bond redistribution and polymerisation to form poly(isoprene); (c) natural rubber, *all-cis*-poly(isoprene); (d) gutta-percha, *all-trans*-poly(isoprene).

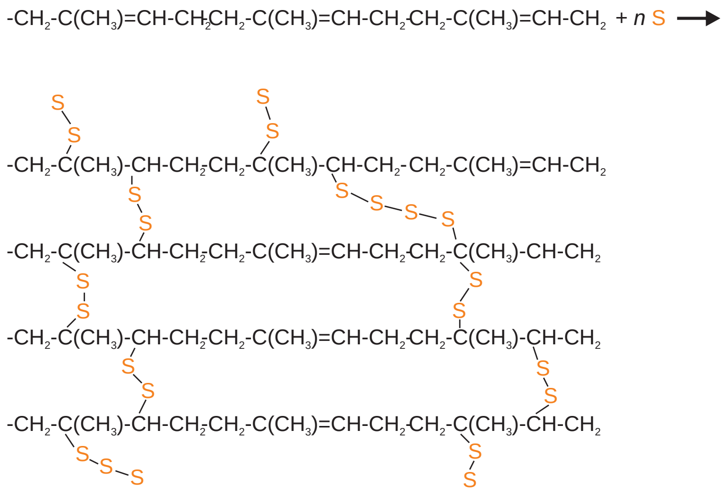
The practice of cross-linking the polyisoprene chains in natural rubber to form a usable elastomer was discovered by Goodyear, in 1839. He heated natural rubber latex with sulphur, a process called *vulcanisation*. This transforms the sticky, runny natural material into a product in which the elastic properties are retained while the stickiness is lost. The cross-linking (Scheme 6.11) utilises the remaining double bonds in the elastomer chains. These are opened by the sulphur molecules to form the ties.

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, (Scheme 6.12a), a rubber resistant to organic solvents prepared from a chlorinated hydrocarbon precursor, and nitrile rubber (Scheme 6.12b), which is a copolymer of butadiene ($CH_2{=}CH{-}CH{=}CH_2$) and propenenitrile ($CH_2{=}CH{-}CN$), containing a nitrile (cyanide, —CN) group.

6.4.5 The principal properties of polymers

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,



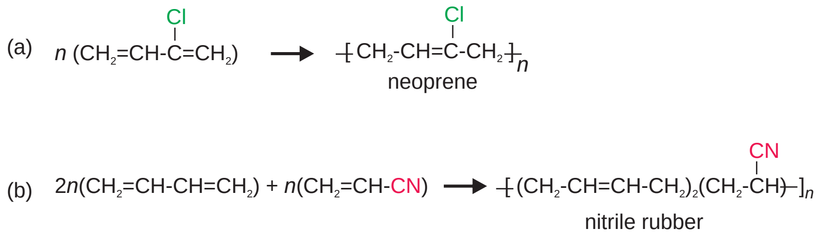
Scheme 6.11 The formation of sulphur cross-links between rubber molecules, produced during vulcanisation.

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 $\rangle$NH . . . O=C$\langle$ 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 good electronic and ionic conductors of electricity, allowing these materials to be used as electrolytes and current collectors in lightweight 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.



Scheme 6.12 Synthetic elastomers: (a) neoprene; (b) nitrile rubber.

One of the principal properties shown by polymers relates to this aspect of chemical stability. Polymers do not degrade rapidly, present an eyesore in many parts of the world, 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, as the many large dumps of used vehicle tyres indicate. The heart of the problem is that the carbon–carbon backbone of polymers is extremely resistant to chemical attack since 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.

6.5 Composite 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.

Manmade 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 below. The mechanical properties of composites are outlined in Section 10.6.

6.5.1 Fibre-reinforced plastics

The main polymers used as matrices in polymer composites are thermosetting resins, especially polyester and epoxy resins. 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 as 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 geometric factors. These include the amount of fibre added and the length and orientation of the fibres (Figure 6.27), 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 6.28).

6.5.2 Metal-matrix composites

Metals are frequently reinforced with continuous fibres to give improvement in strength. The fibres

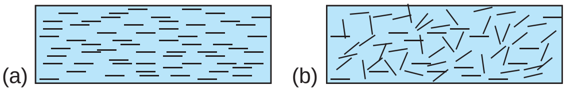


Figure 6.27 (a) Aligned fibre-reinforced composite. (b) Random fibre-reinforced composite.

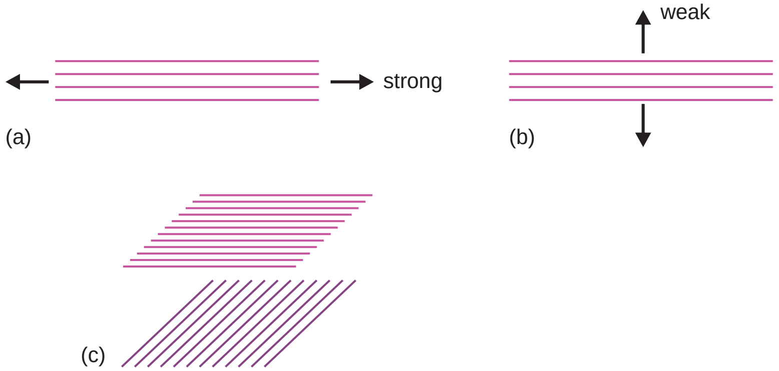


Figure 6.28 Reinforcing fibres tend to be strong in tension (a), but are weak reinforcing materials when subjected to a transverse force (b). (c) Laminates in which the fibres are aligned in differing orientations offset this disadvantage.

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 tried later, the cobalt–tungsten carbide composite remained the best for most purposes.

6.5.3 Ceramic-matrix composites

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 bulletproof glass.

6.5.4 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, pre-date Roman times. The material was made from limestone (impure calcium carbonate) which was heated or *burned* to give *quicklime* or *burnt lime* (calcium oxide).

$$\mathrm{CaCO_3(s) \rightarrow CaO(s) + CO_2(g)}$$

Quicklime reacts with water to release considerable heat, a process called *slaking*, to give *slaked lime* (i.e. calcium hydroxide).

$$\mathrm{CaO(s) + H_2O(l) \rightarrow Ca(OH)_2(s)}$$

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.

6.5.4.1 Portland cement

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 Joseph 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, amongst other projects, the London Underground system, were 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 ($CaSO_4.2H_2O$) 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 6.6). There are also traces of other impurities in ordinary cement powder, which may have a significant effect on the final strength and durability of the concrete.

6.5.4.2 Hardening of Cement

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

$$\begin{aligned}&\text{alite} + \text{water} \rightarrow \text{calcium hydroxide}\\&\qquad\qquad\qquad + \text{calcium silicate hydrate}\\&Ca_3SiO_4 + H_2O \rightarrow Ca(OH)_2 + (CaO)_x(SiO_2)_y{\cdot}aq.\end{aligned}$$

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

Table 6.6 Typical Portland cement constituents

Chemical name	Mineral name	Chemical formula	Shorthand notation	Typical 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

of the calcium silicate hydrate phase is $Ca_2SiO_4.aq$. The reaction is rapid, and continues for up to approximately 20 days. Considerable heat is evolved – of the order of 500 J per gram 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 6.29).

Reaction of Belite and Water

$$\begin{aligned} \text{belite} + \text{water} &\rightarrow \text{calcium hydroxide} \\ &\quad + \text{calcium silicate hydrate} \\ Ca_2SiO_5 + H_2O &\rightarrow Ca(OH)_2 \\ &\quad + (CaO)_x(SiO_2)_y \cdot aq. \end{aligned}$$

The reaction, like that of alite, is imprecise, and the chemical equation is not balanced. Belite reacts slowly, taking about a year to harden, and is responsible for the long-term strength of concrete (Figure 6.29). The heat liberated, approximately 250 J per gram 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.

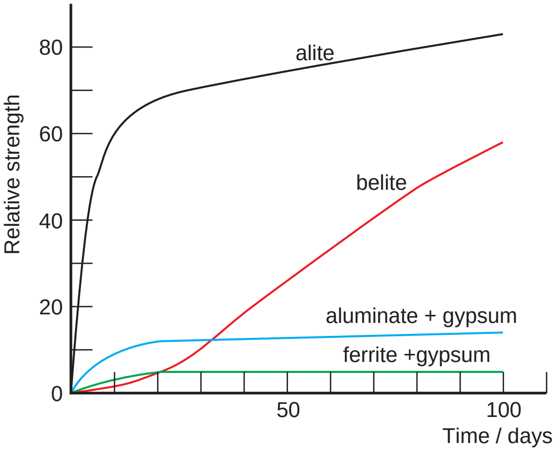


Figure 6.29 The approximate relative strengths of the components of Portland cement after hydration, as a function of time elapsed.

Reaction of Aluminate and Water

$$\begin{aligned} &\text{aluminate} + \text{water} \rightarrow \text{calcium aluminate hydrate} \\ &Ca_3Al_2O_6 + 6H_2O \rightarrow Ca_3Al_2O_6{\cdot}6H_2O \end{aligned}$$

This is a very rapid reaction, and is completed in minutes. It is also a very energetic reaction, and about 900 J per gram 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 per gram. 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

$$\text{gypsum} + \text{aluminate} + \text{water} \rightarrow \text{etringite}$$

$$Ca_3Al_2O_6 + 3CaSO_4{\cdot}2H_2O + 30\,H_2O$$

$$\rightarrow Ca_6Al_2S_3O_{18}{\cdot}32H_2O$$

Gypsum is important as it reacts with aluminate to give *etringite* (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.

6.5.4.3 Heat of hydration

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:

$$H = A\,x_{\mathrm{C3S}} + B\,x_{\mathrm{C2S}} + C\,x_{\mathrm{C3A}} + D\,x_{\mathrm{C4AF}}$$

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 days), and after 1 year, H(1 year), is:

$$\begin{aligned} H(3\ \text{days}) &= 240\,x_{\mathrm{C3S}} + 50\,x_{\mathrm{C2S}} + 880\,x_{\mathrm{C3A}} + 290\,x_{\mathrm{C4AF}} \\ H(1\ \text{year}) &= 490\,x_{\mathrm{C3S}} + 225\,x_{\mathrm{C2S}} + 1160\,x_{\mathrm{C3A}} \\ &\quad +375\,x_{\mathrm{C4AF}} \end{aligned}$$

6.5.4.4 Microstructures of cement and concrete

The microstructures that form as the 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 etringite, 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.

Further reading

General texts:

Callister, W.D. (2007) *Materials Science and Engineering*, 7th edn. John Wiley and Sons, Ltd., New York.

Smallman, R.E. and Ngan, A.H.W. (2007) *Physical Metallurgy and Advanced Materials*, 7th edn. Elsevier, Oxford.

Ceramics and glass:

Beall, G.H. (1992) Synthesis and design of glass ceramics. *J. Materials Educ.*, **14**: 315.

Boyd, D.C. and Thompson, D.A. (1980) Glass, in *Kirk-Othmer Encyclopedia of Chemical Technology*, 3rd edn, Vol. 11. John Wiley and Sons, Ltd., New York.

Carter, C.B. and Norton, M.G. (2007) *Ceramic Materials: Science and Engineering*. Springer Science+Business Media, New York.

For information on glass fragility see:

Angell, C.A. (2002) *Chem. Rev.*, **102**: 2627–50. and the references therein.

Martinez, L.-M. and Angell, C.A. (2001) *Nature*, **410**: 663–7.

Xia, X. and Wolynes, P.G. (2000) *Proc. Natl. Acad. Science*, **97**: 2990.

Bulk metallic glasses:

Johnson, W.L. (1999) *Materials Research Society Bulletin*, **24** (October): 42.

Polymers:

Bates, F.S., *et al.* (2012) Multiblock polymers: panacea or Pandora's box? *Science*, **36**: 434–60.

Ebewele, R.O. (2000) *Polymer Science and Technology*. Chapman and Hall, CRC Press, Boca Raton, FL.

Ewen, J.A. (1997) *Scientific American*, **276** (May): 60.

Young, R. and Lovell, P. (2011) *Introduction to Polymers*. CRC Press, Boca Raton, FL.

For information on polymer nomenclature see:

Maréchal, E. and Wilks, E.S. (2001) *Pure Appl. Chem.*, **73**: 1511–19.

Maréchal, E. and Wilks, E.S. (2002) *Pure Appl. Chem.*, **74**: 2019.

Wilks, E.S. (2000) *Prog. Polymer Sci.*, **25**: 9–100.

Composites:

Barbero, E.J. (2011) *Introduction to Composite Materials Design*, 2nd edn. Taylor and Francis, CRC Press, Boca Raton, FL.

Mindess, S. (1983) Concrete materials. *J. Materials Education*, **5**: 983.

Problems and exercises

Quick quiz

1 In the A1 (face-centred cubic) structure the atoms are in contact along:
(a) A cube body diagonal.
(b) A cube face diagonal.
(c) A cube edge.

2 Ordered alloy structures such as Cu_3Au are formed by:
(a) Rapidly cooling the alloy.
(b) Rapidly heating the alloy.
(c) Annealling the alloy.

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 In each unit cell of the A1 (face-centred cubic) structure there are:
(a) 4 octahedral sites.
(b) 8 octahedral sites.
(c) 12 octahedral sites.

5 In each unit cell of the A2 (body-centred cubic) structure there are:
(a) 4 octahedral sites.
(b) 6 octahedral sites.
(c) 8 octahedral sites.

6 The tetrahedral sites in interstitial alloys can be occupied by:
(a) Hydrogen.
(b) Nitrogen.
(c) Both hydrogen and nitrogen.

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.

8 Isomorphous replacement in silicate ceramics creates:
(a) Substitutional defects.
(b) Interstitial defects.
(c) No defects.

9 Silicates are stable because of strong bonds between:
(a) Oxygen and oxygen.
(b) Metal atoms and silicon.
(c) Silicon and oxygen.

10 Ionic silicates contain isolated:
(a) Silicate groups.
(b) Silicate chains.
(c) Silicate sheets.

11 Clays are silicates containing:
(a) Chains of $[SiO_4] +$ hydroxyl.
(b) Networks of $[SiO_4] +$ hydroxyl.
(c) Sheets of $[SiO_4] +$ hydroxyl.

12 Ruby is a gemstone consisting of aluminium oxide containing small amounts of:
(a) Titanium impurity.
(b) Chromium impurity.
(c) Titanium and iron impurities.

13 The formation of a glass during ceramic production is called:
(a) Vitrification.
(b) Sintering.
(c) Glass–ceramic formation.

14 A refractory ceramic is one that is:
(a) Difficult to process.

(b) Particularly hard.

(c) Able to withstand high temperatures.

15 Pyrex® glass is also known as:
(a) Flint glass.
(b) Borosilicate glass.
(c) Soda-lime glass.

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.

17 In glass technology, small ions such as P and B are known as:
(a) Network modifiers.
(b) Network formers.
(c) Intermediates.

18 A glass ceramic is:
(a) A glass processed at high temperatures.
(b) A transparent ceramic.
(c) A ceramic containing both crystals and glass.

19 Polystyrene is an example of:
(a) An addition polymer.
(b) A condensation polymer.
(c) An elastomer.

20 Nylon 66 is an example of:
(a) An addition polymer.
(b) A condensation polymer.
(c) An elastomer.

21 Nylons are:
(a) Polyamides.
(b) Polyesters.
(c) Polycarbonates.

22 The molar mass of a polymer:
(a) Is a fixed number.
(b) Varies between narrow limits.
(c) Varies between wide limits.

23 The molar mass of a monomer:
(a) Is a fixed number.
(b) Varies between narrow limits.
(c) Varies between wide limits.

24 A lamella in a polymer is:
(a) A spherulite.
(b) A chain.
(c) A crystalline region.

25 A polymer in which the side groups lie on alternate sides of the polymer chain backbone is:
(a) Isotactic.
(b) Syndiotactic.
(c) Atactic.

26 The growth of polymer chains by the joining of existing chains is called:
(a) Chain growth.
(b) Step growth.
(c) Link growth.

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.

28 Ceramic matrix composites are designed to overcome:
(a) The weight of ceramics.
(b) The brittle nature of ceramics.
(c) The inertness of ceramics.

29 The initial hardening of Portland cement is attributed to:
(a) Alite.
(b) Belite.
(c) Gypsum.

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.

31 The long-term hardening of Portland cement is attributed to the presence of:
  (a) Tricalcium silicate.
  (b) Dicalcium silicate.
  (c) Tricalcium aluminate.

Calculations and questions

6.1 The radius of a gold atom is 0.144 nm. Gold adopts the A1 structure. Estimate the unit cell parameter of gold crystals.

6.2 The radius of a lead atom is 0.175 nm. Lead adopts the A1 structure. Estimate the unit cell parameter of lead crystals.

6.3 The radius of a palladium atom is 0.138 nm. Palladium adopts the A1 structure. Estimate the unit cell parameter of palladium crystals.

6.4 The radius of an iridium atom is 0.136 nm. Iridium adopts the A1 structure. Estimate the unit cell parameter of iridium crystals.

6.5 Nickel adopts the A1 structure, with a lattice parameter of 0.3524 nm. Estimate the metallic radius of nickel in this structure.

6.6 Rhodium adopts the A1 structure, with a lattice parameter of 0.3803 nm. Estimate the metallic radius of rhodium in this structure.

6.7 The radius of a barium atom is 0.224 nm. Barium adopts the A2 structure. Estimate the unit cell parameter of barium crystals.

6.8 The radius of a niobium atom is 0.147 nm. Niobium adopts the A2 structure. Estimate the unit cell parameter of niobium crystals.

6.9 Vanadium adopts the A2 structure, with a lattice parameter of 0.3024 nm. Estimate the metallic radius of vanadium in this structure.

6.10 Potassium adopts the A2 structure, with a lattice parameter of 0.5321 nm. Estimate the metallic radius of potassium in this structure.

6.11 At room temperature, iron adopts the 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 A1 allotrope that exists above 912 °C.

6.12 At room temperature, calcium adopts the 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.

6.13 At room temperature, strontium adopts the A1 structure. The metallic radius of strontium in this structure is 0.215 nm. (a) Estimate the lattice parameter of the A1 structure of strontium. (b) Ignoring thermal expansion, estimate the lattice parameter of the A2 allotrope that exists above 527 °C.

6.14 The metallic radius of magnesium, which adopts the A3 structure, is 0.160 nm. Estimate the value of the lattice parameters a and c (ideal)

6.15 The metallic radius of rhenium, which adopts the A3 structure, is 0.138 nm. Estimate the value of the lattice parameters a and c (ideal)

6.16 The a lattice parameter of titanium, which adopts the A3 structure, is 0.2951 nm. Estimate the radius of titanium in this structure and the ideal value of the parameter c.

6.17 The a lattice parameter of beryllium, which adopts the A3 structure, is 0.2286 nm. Estimate the radius of beryllium in this structure and the ideal value of the parameter c.

6.18 At room temperature, hafnium adopts the 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 A2 allotrope that exists above 1742 °C.

6.19 At room temperature, yttrium adopts the 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 A2 allotrope that exists above 1481 °C.

6.20 The rapidly cooled form of the alloy CuAu has the 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.

6.21 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 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.

6.22 Which of the following metals would be expected to form the most extensive substitutional solid solution with nickel, A1 structure, metallic radius 0.1246 nm, electronegativity 1.8?

(a) Cobalt, A3 structure, metallic radius 0.125 nm, electronegativity 1.7.

(b) Chromium, A2 structure, metallic radius 0.128 nm, electronegativity 1.4.

(c) Platinum, A1 structure, metallic radius 0.139 nm, electronegativity 2.1.

(d) Silver, A1 structure, metallic radius 0.145 nm, electronegativity 1.8.

6.23 Which of the following metals would be expected to form the most extensive substitutional solid solution with copper, A1 structure, metallic radius 0.1278 nm, electronegativity 1.8?

(a) Aluminium, A1 structure, metallic radius 0.143 nm, electronegativity 1.5.

(b) Palladium, A1 structure, metallic radius 0.138 nm, electronegativity 2.0.

(c) Vanadium, A2 structure, metallic radius 0.135 nm, electronegativity 1.4.

(d) Titanium, A3 structure, metallic radius 0.146 nm, electronegativity 1.6.

6.24 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

6.25 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}.

6.26 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.

6.27 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.

6.28 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.

6.29 The viscosity of a borosilicate glass is drawn in Figure 6.30a. Estimate the activation energy of the viscosity and comment on the form of the Arrhenius plot.

6.30 The viscosity of a high-silica glass is drawn in Figure 6.30b. Estimate the activation energy of the viscosity and comment on the form of the Arrhenius plot.

6.31 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? The number average molar mass of the polymer is 250 000 g mol^{-1}. What is the degree of polymerisation?

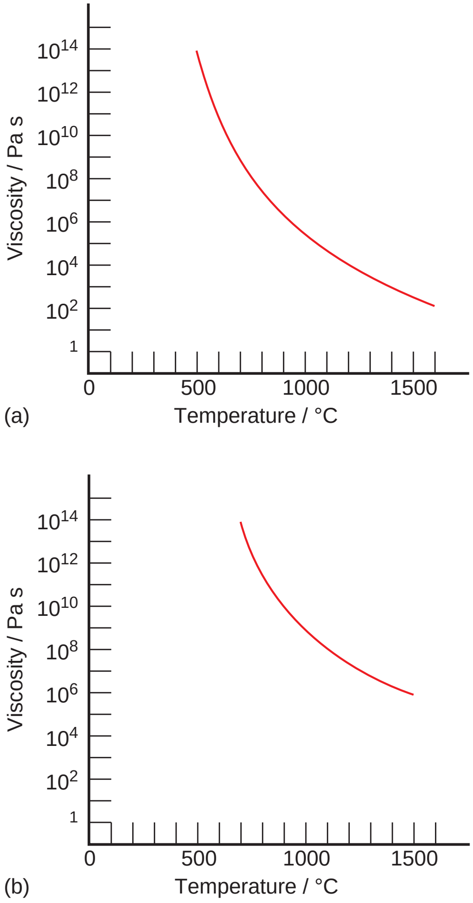


Figure 6.30 Plot of viscosity against temperature for (a) a borosilicate glass, Question 6.29; (b) a high-silica glass, Question 6.30.

6.32 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?

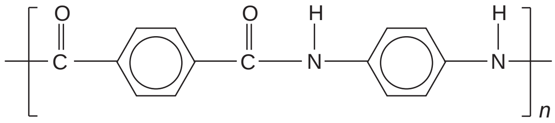
6.33 Nitrile rubber is a copolymer of butadiene ($CH_2{=}CH{-}CH{=}CH_2$) and acrylonitrile (propene nitrile) ($CH_2{=}CH{-}CN$). Write the reaction equation between these two molecules to produce a dimer. (a) What masses of the reactants are needed to give 100 kg of polymer? (b) How many molecules of each reactant is this?

6.34 PET is a polymer of terephthalic acid and ethylene glycol (see Scheme 6.2). Write the hypothetical reaction equation between these two molecules to produce a dimer. (a) What masses of the reactants are needed to give 100 kg of polymer? (b) How many molecules of each reactant is this?

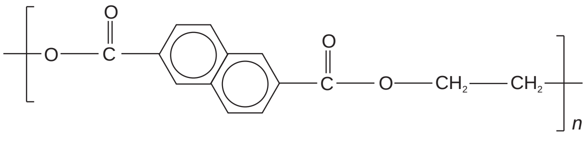
6.35 The structural formula of the aramid polymer Kevlar, used in bulletproof vests, is drawn in Scheme 6.13. Write the formulae of the two monomers used.

6.36 The structural formula of the polycarbonate PEN [poly(ethelylene naphthalate)], used in recyclable jars and bottles, is drawn in Scheme 6.14. Write the formulae of the two monomers used.

6.37 Compute (a) the number average molar mass and (b) the degree of polymerisation for polypropylene from the following data.



Scheme 6.13 Kevlar.



Scheme 6.14 Poly(ethylene naphthalate) (PEN).

Molar mass range/g mol^{-1}	Mean molar mass in range/g mol^{-1}	Fraction of molecules in range
5000–10,000	7500	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

6.38 Compute (a) the weight average molar mass and (b) 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
5000–10 000	7500	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

6.39 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?

6.40 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 the previous question.)

6.41 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.

PART 3

Reactions and transformations

7

Diffusion and ionic conductivity

- What is steady-state diffusion?
- What are diffusion coefficients?
- What is a correlation factor?

Diffusion originally described the way in which heat (believed to be a fluid) flowed through a solid. Later the same ideas were applied to describe the way in which a gas would spread out to fill the available volume. In solids, diffusion refers to the transport of atoms, ions or molecules under the influence of a driving force that is usually a concentration gradient. Diffusion takes place in solids at a much slower rate than in gases or liquids, and in the main it is a high-temperature process. However, this is not always so, and in some solids the rate of diffusion at room temperature is considerable.

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 all other crystals, the rate of diffusion depends upon the direction taken, and is *anisotropic*. Moreover, atoms can also 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 pathways is collectively referred to as *short-circuit* diffusion.

Diffusion is studied by measuring the concentration of the atoms at different distances from the release point after a given time has elapsed. Raw experimental data thus consists of concentration and distance values. The speed at which atoms or ions move through a solid is usually expressed in terms of a *diffusion coefficient*, *D*, which is obtained from experimental data by use of two *diffusion equations*. In general, it has been found that *D* depends on position and concentration, and hence *D* varies throughout the solid. In this chapter, the focus is upon volume diffusion along a single direction in an isotropic medium. Moreover, attention is confined to diffusion when the concentration of the diffusing species is very small, so that concentration effects are not important, or where the diffusion coefficient does not depend upon concentration or position. This is equivalent to stipulating that the diffusion coefficient is a constant.

7.1 Self-diffusion, tracer diffusion and tracer impurity diffusion

When atoms in a pure crystal diffuse under no concentration gradient or other driving force, the

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

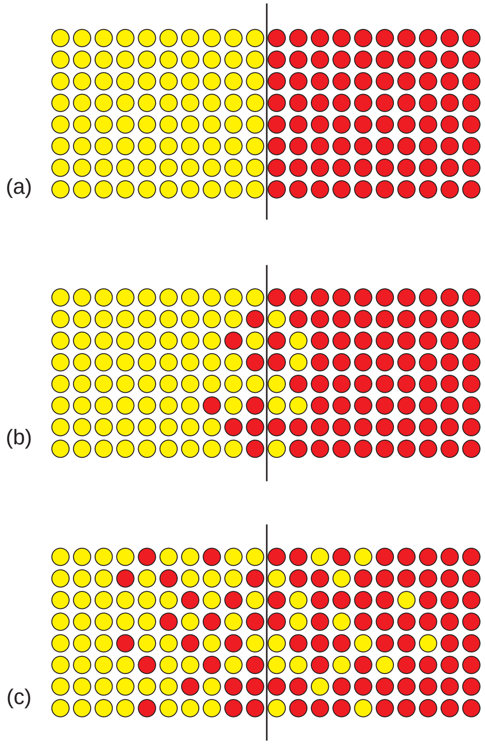


Figure 7.1 The diffusion of atoms: (a) initial configuration; (b) after heating for shorter times; (c) after heating for longer times.

process is called *self-diffusion*. In such a case, the atomic movements are random, with motion in one direction just as likely as in another (Figure 7.1). The relevant diffusion coefficient is called the *self-diffusion coefficient* and is given the symbol D_s.

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.

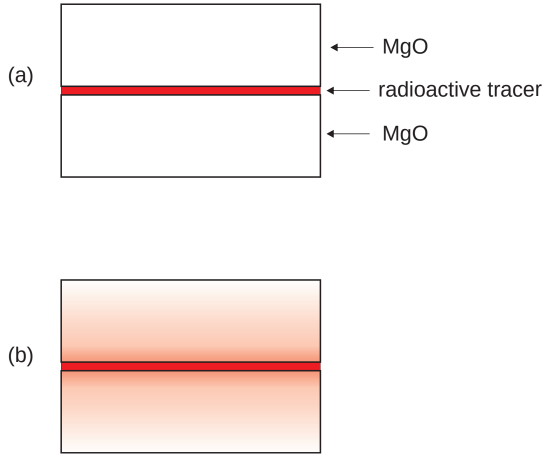


Figure 7.2 A diffusion couple formed by two crystals of MgO separated by radioactive material: (a) initially; (b) after heating.

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 7.2). 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 7.3). Note that the concentration of the tracer at any point changes over time.

The tracer diffusion coefficient is obtained from such profiles via the *diffusion equation*, also called *Fick's second law*:

$$\frac{dc_x}{dt} = D^* \frac{d^2 c_x}{dx^2} \tag{7.1}$$

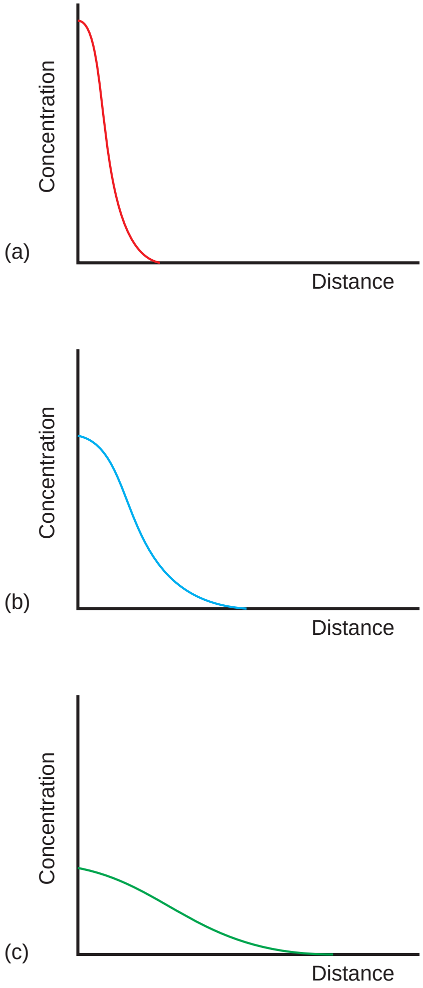


Figure 7.3 Diffusion profiles: (a), (b) and (c) represent gradually increasing heating times.

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 7.2, the solution is:

$$c_x = \frac{c_0}{2(\pi D^* t)^{1/2}} \exp\left(\frac{-x^2}{4D^* t}\right) \qquad (7.2)$$

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 of this equation to give:

$$\ln c_x = \ln\left(\frac{c_0}{2(\pi D^* t)^{1/2}}\right) - \frac{x^2}{4D^* t} \qquad (7.3)$$

This has the form:

$$\ln c_x = \text{constant} - \frac{x^2}{4D^* t}$$

and a plot of ln c_x versus x^2 will have a gradient of $(-1/4D^* t)$ (Figure 7.4). 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.

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.

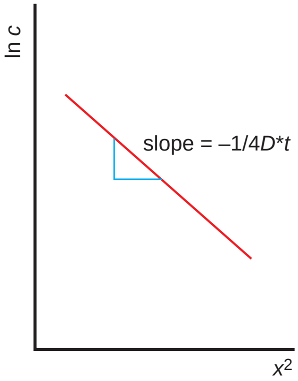


Figure 7.4 A straight-line graph of ln c versus x^2 from a diffusion experiment. The slope of the graph allows a value for the tracer diffusion coefficient, D*, to be determined.

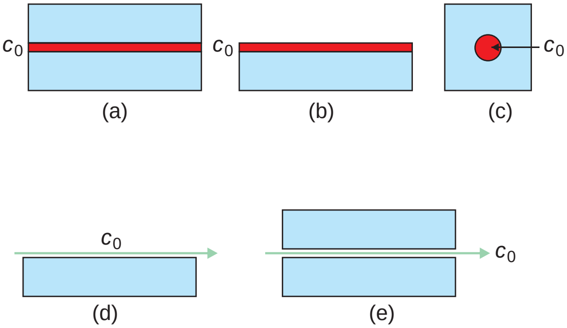


Figure 7.5 Geometries for non-steady state diffusion: (a)–(c), the concentration of diffusant is unreplenished; (d)–(e) the concentration of diffusant is maintained at a constant value, c_0, by gas or liquid flow.

Table 7.1 Solutions of the diffusion equation

Experimental arrangement	Solution[*]
Initial concentration, c_0, unreplenished	
Thin film planar sandwich	$c_x = \dfrac{c_0}{2(\pi Dt)^{1/2}} \exp\left(\dfrac{-x^2}{4Dt}\right)$
Open planar thin film	$c_x = \dfrac{c_0}{(\pi Dt)^{1/2}} \exp\left(\dfrac{-x^2}{4Dt}\right)$
Small spherical precipitate	$c_r = \dfrac{c_0}{8(\pi Dt)^{3/2}} \exp\left(\dfrac{-r^2}{4Dt}\right)$
Initial concentration, c_0, maintained constant	
Open plate	$\dfrac{c_x - c_0}{c_s - c_0} = 1 - \operatorname{erf}\left(\dfrac{x}{2(Dt)^{1/2}}\right)$
Sandwich plate	$\dfrac{c_x - c_0}{c_s - c_0} = \dfrac{1}{2}\left[1 - \operatorname{erf}\left(\dfrac{x}{2(Dt)^{1/2}}\right)\right]$

[*] Derived assuming diffusion occurs in infinite solids with constant D.

7.2 Non-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 above. 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 7.5 and Table 7.1).

In some experimental arrangements (Figure 7.5a, b,c), 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, gas molecules diffuse into a solid and the gas supply is constantly replenished (Figure 7.5d,e). The solution for diffusion in these cases involves the *error function*, in the form erf $[x/2\ (Dt)^{1/2}]$. The error function cannot generally be evaluated analytically, but numerical values can readily be calculated via many computer routines. An abbreviated list is given in Table 7.2.

Table 7.2 Values of the error function[*]

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

[*] erf (z) is equal to 0 when z is equal to 0, and equal to 1 when z is equal to 2.8.
erf $(-z) = [-\text{erf}\ (z)]$.

7.3 Steady-state diffusion

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 7.6). Steady-state diffusion can occur when a gas permeates through a metal foil or thin-walled tube. Hydrogen gas, for example, can be purified by allowing it to diffuse through a palladium thimble. The same steady-state conditions can arise when oxygen diffuses through a plastic wrapping film.

Under steady-state conditions, the diffusion coefficient is obtained using *Fick's first law*. This is written:

$$J_\mathrm{i} = -D_\mathrm{i}\frac{dc_\mathrm{i}}{dx} \tag{7.4}$$

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) $\mathrm{m}^{-2}\,\mathrm{s}^{-1}$. When the steady state has been reached, the diffusion coefficient across the foil, D, will be given by:

$$D = \frac{J\,l}{c_1 - c_2}$$

where the concentrations on each side of the foil are c_1 and c_2 and the foil thickness is l.

7.4 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*:

$$D = D_0 \exp\left(\frac{-E}{RT}\right) \tag{7.5}$$

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.

$$\ln D = \ln D_0 - \frac{E}{RT}$$

The activation energy can be determined from the gradient of a plot of ln D versus $1/T$ (Figure 7.7). Such graphs are known as *Arrhenius plots*. Diffusion coefficients are usually expressed in terms of D_0 and E (Table 7.3).

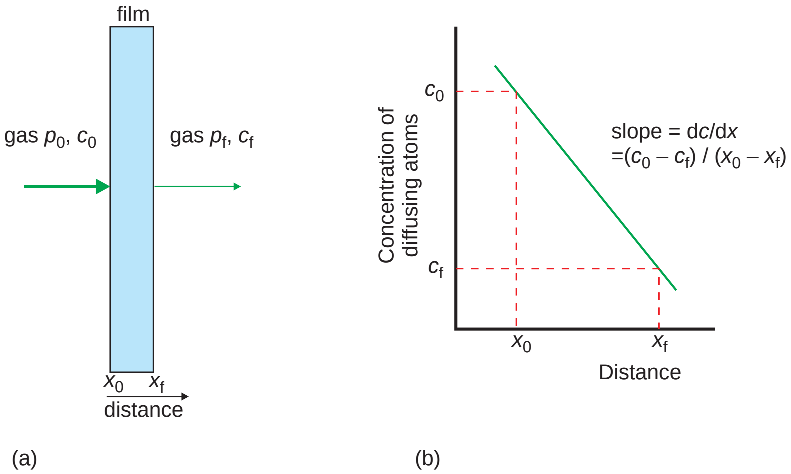

film
gas p_0, c_0
gas p_f, c_f
x_0
x_f
distance
(a)
Concentration of diffusing atoms
c_0
c_f
slope = dc/dx
=(c_0 – c_f) / (x_0 – x_f)
x_0
x_f
Distance
(b)

Figure 7.6 Steady-state diffusion: (a) a gas diffusing through a thin film or foil: (b) resultant diffusion profile.

Table 7.3 Some representative values for self-diffusion coefficients*

Atom	Matrix**	$D_o/\mathrm{m^2\,s^{-1}}$***	$E/\mathrm{kJ\,mol^{-1}}$
Metals and semiconductors			
Cu	Cu (A1)	2.0×10^{-5}	200
γ-Fe	Fe (A1)	2.0×10^{-5}	270
α-Fe	Fe (A2)	2.0×10^{-4}	240
Na	Na (A2)	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
Glasses			
Na	Soda-lime	1.0×10^{-6}	85
Si	Silica	3.0×10^{-2}	580
Na	Soda	1.0×10^{-6}	80

*Literature values for self-diffusion coefficients vary widely, indicating the difficulty of making reliable measurements. The values here are meant to be representative only.

**The structure type is shown in parentheses.

***The values of diffusion coefficients in the literature are mostly given in $\mathrm{cm^2\,s^{-1}}$. To convert the values given here to $\mathrm{cm^2\,s^{-1}}$, multiply by 10^4.

7.5 The effect of impurities

Although Arrhenius plots for the majority of materials resemble Figure 7.7, plots obtained from very pure materials often consist of two straight-line parts with differing slopes (Figure 7.8). 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. This shows that impurities play a significant role in diffusion: an observation explained by consideration of diffusion mechanisms.

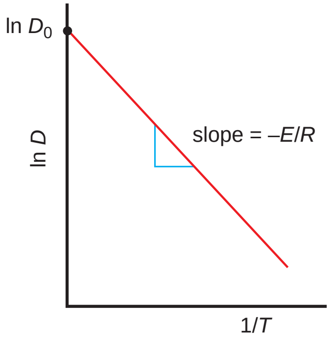


Figure 7.7 An Arrhenius plot of diffusion data, ln D versus 1/T. D, diffusion coefficient; T temperature (K); D_0, pre-exponential factor.

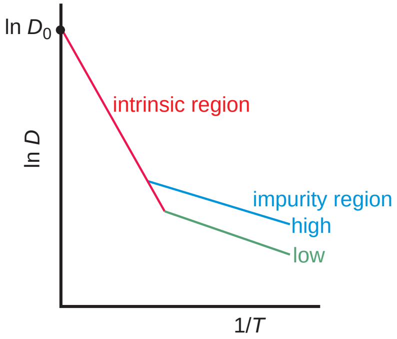


Figure 7.8 A form of Arrhenius plot found for almost pure crystals with low impurity concentrations. D, diffusion coefficient; T, temperature (K); D_0, pre-exponential factor.

7.6 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*. 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 and a diffusion coefficient for this mechanism can be derived.

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. 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.:

$$x = x_1 + x_2 + x_3 \ldots = \sum x_i$$

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.

To remove the uncertainty inherent in this result, the jump distances are squared so as to eliminate negative quantities.

$$\begin{aligned} x^2 &= (x_1 + x_2 + x_3 \ldots x_n)(x_1 + x_2 + x_3 \ldots x_n) \\ &= (x_1x_1 + x_1x_2 + x_1x_3 \ldots x_1x_n) \\ &\quad + (x_2x_1 + x_2x_2 + x_2x_3 \ldots x_2x_n) \\ &\quad + \ldots \\ &\quad + (x_nx_1 + x_nx_2 + x_nx_3 \ldots x_nx_n) \end{aligned}$$

This can be written:

$$\begin{aligned} x^2 &= \sum x_i^2 + 2\sum x_ix_{i+1} + 2\sum x_ix_{i+2} + \ldots \\ &= \sum x_i^2 + 2\sum\sum x_ix_{i+j} \end{aligned}$$

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 $-a$,

$$\begin{aligned} \langle x^2 \rangle &= x_1^2 + x_2^2 + x_3^2 \ldots + x_n^2 \\ &= a^2 + a^2 + a^2 \ldots\ldots + a^2 \\ &= na^2 \end{aligned}$$

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:

$$\langle x^2 \rangle = na^2 = \Gamma ta^2$$

It is possible to use this equation to define a diffusion coefficient for the random walk. To agree with Fick's first law, this is:

$$D_{\mathrm{r}} = \frac{1}{2}\Gamma a^2$$

so that

$$\langle x^2 \rangle = 2D_{\mathrm{r}}t$$

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^*. The statistics of the random walk show that, 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 $\pm\sqrt{\langle x^2\rangle} = \pm\sqrt{(2D_r t)}$ 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 can 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:

$$x_p = \sqrt{(2D^* t)} \tag{7.6}$$

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:

$$x_p = 2\sqrt{(2D^* t)}$$

or as the diffusion length or diffusion distance, L, where:

$$L = \sqrt{(D^* t)}$$

The factor of 2 in equation (7.6) 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:

$$\langle x^2\rangle = 6D_r t$$

7.7 Diffusion in solids

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. Nevertheless, volume diffusion does occur. This conflict can be overcome by assuming the presence of point defects in the crystal.

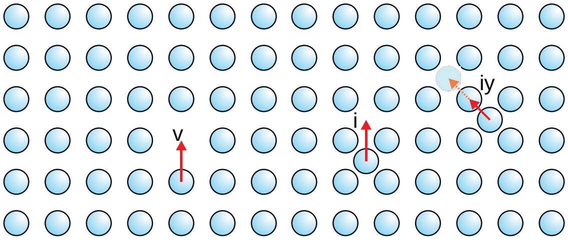


Figure 7.9 Self-diffusion mechanisms; v, vacancy; i, interstitial; iy, interstitialcy.

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 7.9). 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 (Figure 7.9). 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*.

Substitutional impurity defects can move by way of three mechanisms (Figure 7.10). As well as vacancy diffusion, an impurity can swap places with a neighbouring normal atom – *exchange diffusion* – while in *ring diffusion* cooperation 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

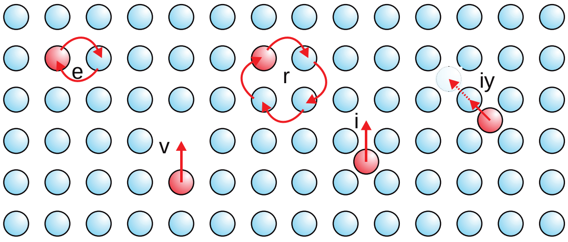


Figure 7.10 Impurity diffusion mechanisms; v, vacancy; e, exchange; r, ring; i, interstitial; iy, interstitialcy.

can move by interstitial and interstitialcy jumps similar to those described above (Figure 7.10).

When Schottky defects are present in a crystal, vacancies are found on both the cation and anion sublattices, allowing both cation and anion diffusion to occur (Figure 7.11a). In the case of Frenkel defects, interstitials and vacancies occur, allowing interstitial, interstitialcy and vacancy diffusion to take place in the same crystal (Figure 7.11b).

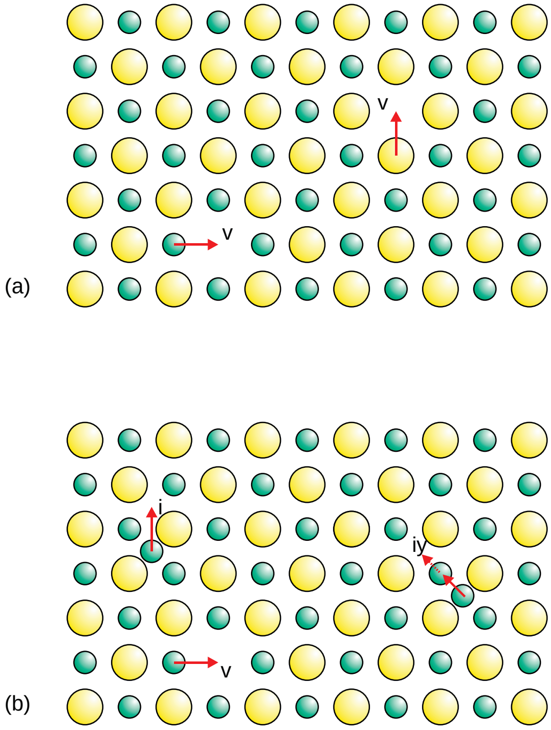


Figure 7.11 (a) Vacancy diffusion, v, in a crystal containing Schottky defects; (b) vacancy diffusion, v; interstitial diffusion, i; and interstitialcy diffusion, iy; in crystals containing Frenkel defects.

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 in the previous section, is a good indicator of the spatial extent of diffusion.

7.8 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 which 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 7.12a). 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 7.12b).

Obviously, 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:

$$p = \exp\left(\frac{-E}{k_B T}\right) \qquad (7.7)$$

where k_B is the Boltzmann constant and T the temperature (K). Equation (7.7) indicates that if E is very

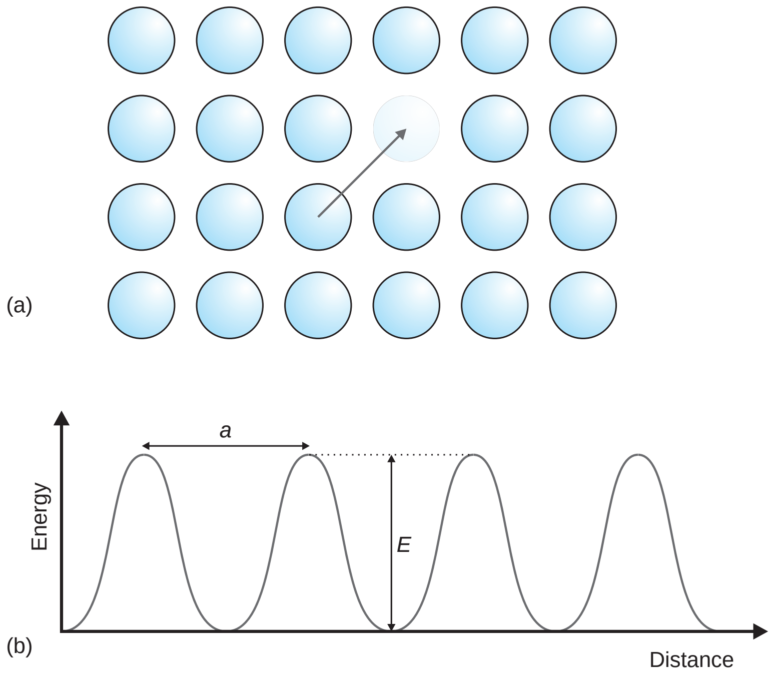


Figure 7.12 (a) An atom migrating from one stable position to another, separated by a distance a. (b) The energy barrier encountered has a periodicity equal to a, and is a maximum when the diffusing atom has to pass a bottleneck between two stationary atoms.

small, the probability that the atom will clear the barrier approaches 1.0; if E is equal to $k_{\mathrm{B}}T$, the probability for a successful jump is about one third, and if E increases above $k_{\mathrm{B}}T$ 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.:

$$\Gamma = \nu \exp\left(\frac{-E}{k_{\mathrm{B}}T}\right) \tag{7.8}$$

This flow of atoms, say along the x-direction, is related to the diffusion coefficient by Fick's first law, equation 7.4:

$$J = -D\frac{dc}{dx}$$

To relate this to equation (7.8), take 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 7.13). If Γ_{12} is the frequency with which an atom moves from

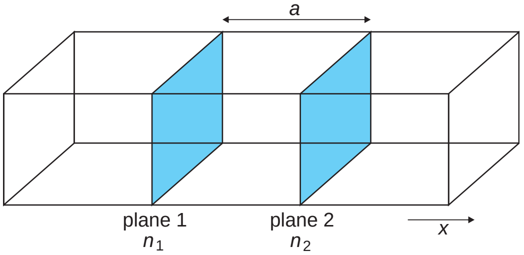


Figure 7.13 Two adjacent planes in a crystal, 1 and 2, containing n_1 and n_2 diffusing atoms, separated by the jump distance of the diffusing ion, a.

plane 1 to plane 2, then the numbers of atoms moving from plane 1 to 2 per second is j_{12}, where:

$$j_{12} = n_1 \Gamma_{12}$$

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:

$$\Gamma_{12} = \Gamma_{21} = \frac{1}{2}\,\Gamma$$

where Γ represents the overall jump frequency of the diffusing atoms. Thus:

$$J = \frac{1}{2}\,(n_1 - n_2)\,\Gamma$$

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:

$$\mathrm{J} = \frac{1}{2}\,a(c_1 - c_2)\,\Gamma$$

The concentration gradient, dc/dx, is given by the change in concentration between planes 1 and 2 divided by the distance between planes 1 and 2, that is:

$$\frac{-dc}{dx} = \frac{c_1 - c_2}{a}$$

where the minus sign is introduced as the concentration falls on moving from plane 1 to plane 2. Hence:

$$c_1 - c_2 = -a\frac{dc}{dx}$$

and:

$$J = -\frac{1}{2}\Gamma\, a^2 \frac{dc}{dx}$$

A comparison with Fick's First Law shows that:

$$D = \frac{1}{2}\,\Gamma a^2$$

Substituting the expression for the jump frequency, Γ, in terms of the barrier height to be negotiated, E, yields:

$$D = \frac{1}{2}a^2\,\nu \exp\left(\frac{-E}{k_\mathrm{B}T}\right)$$

or, in molar quantities:

$$D = \frac{1}{2}a^2\,\nu \exp\left(\frac{-E}{RT}\right)$$

7.9 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 the equation for D so that it becomes:

$$D = g\,a^2\,\nu \exp\left(\frac{-E}{RT}\right)$$

In the one-dimensional case, the factor $^1\!/_2$ is a geometric 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 1/6 is appropriate.

In the foregoing discussion, every possible atom jump is allowed. This may not be true in real crystals. Equation (7.8) ignores this, and should contain a term, p_J, that expresses the probability that the jump is possible from this structural point of view:

$$\Gamma = p_J \, \nu \exp\left(\frac{-E}{RT}\right)$$

The diffusion coefficient is then given by:

$$D = p_J \, g \, a^2 \, \nu \exp\left(\frac{-E}{RT}\right) \tag{7.9}$$

For example, in the A1 (face-centred cubic) structure of magnesium, 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.

7.10 The Arrhenius equation and point defects

A comparison of equation (7.9) with the Arrhenius equation, equation (7.5), 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 7.14).

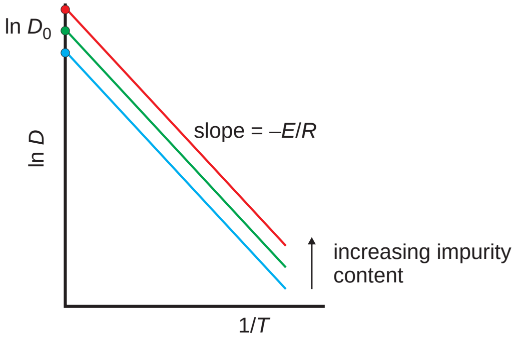


Figure 7.14 The variation of Arrhenius plots with impurity content. D, diffusion coefficient; T, temperature (K); D_0, pre-exponential factor.

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 equation (3.3):

$$n_S = N \exp\left(\frac{-\Delta H_S}{2RT}\right)$$

At high temperatures the appropriate form of equation (7.9) becomes:

$$D = g \, \nu \, a^2 \exp\left(\frac{-E_v}{RT}\right) \exp\left(\frac{-\Delta H_S}{2RT}\right)$$

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 equation (3.6):

$$n_F = \sqrt{(NN^*)}\exp\left(\frac{-\Delta H_\mathrm{F}}{RT}\right)$$

In these circumstances, the probability factor, p_J, will be proportional to the number of Frenkel defects present, and it is possible to write equation (7.9) in a form:

$$D = g\,\nu\,a^2\exp\left(\frac{-E_\mathrm{i}}{RT}\right)\exp\left(\frac{-\Delta H_\mathrm{F}}{2RT}\right)$$

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:

$$D = D_0\exp\left(\frac{-E}{RT}\right)$$

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:

$$E = E_\mathrm{v} + \frac{\Delta H_\mathrm{S}}{2}$$

and for Frenkel defects:

$$E = E_\mathrm{i} + \frac{\Delta H_\mathrm{F}}{2}$$

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 7.8). 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 7.4.

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.

Table 7.4 Approximate enthalpy values for the formation and movement of vacancies in alkali halide crystals

Schottky defects			
Material	ΔH_s/kJ mol^{-1}	E_v (cation)/kJ mol^{-1}	E_v (anion)/kJ mol^{-1}
NaCl	192	84	109
NaBr	163	84	113
KCl	230	75	172
KBr	192	64	46
Frenkel defects			
Material	ΔH_f/kJ mol^{-1}	E_i (interstitial)/kJ mol^{-1}	E_v (vacancy)/kJ mol^{-1}
AgCl	155	13	36
AgBr	117	11	23

7.11 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 (or 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 7.15a). 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 7.15b). 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 7.15c). The tracer can only jump to a new position after the vacancy has diffused to an alternative neighbouring position.

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 accurate, and a correlation factor of 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

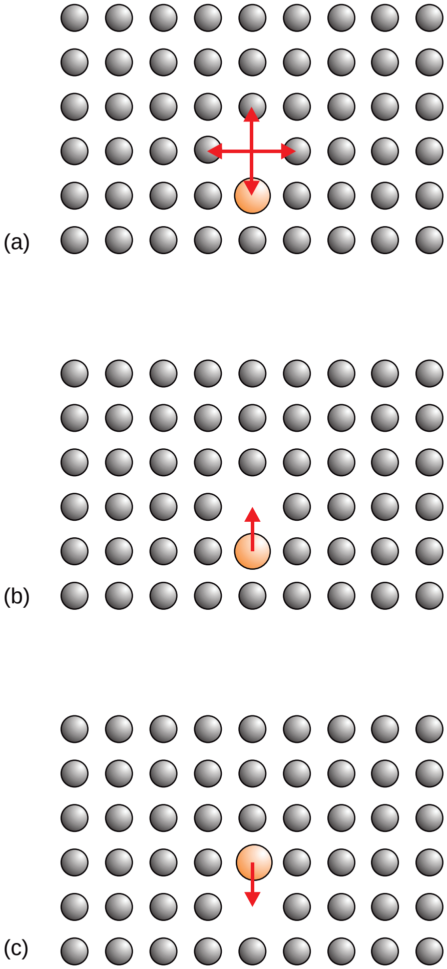


Figure 7.15 Correlated motion during vacancy diffusion: (a) a vacancy can jump to any surrounding position; the motion of a tracer is correlated so that a jump into a vacancy (b) is most likely to be followed by a reverse jump (c).

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:

$$f = \frac{\langle x^2 \rangle_\mathrm{c}}{\langle x^2 \rangle_\mathrm{r}} = \frac{D^*}{D_\mathrm{r}}$$

that is:

$$D^* = f\, D_\mathrm{r}$$

where $\langle x^2 \rangle_\mathrm{c}$ represents the mean square displacement of a correlated walk by the diffusing atom and $\langle x^2 \rangle_\mathrm{r}$ is 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 7.5).

Table 7.5 Correlation factors for self-diffusion

Crystal structure	Correlation factor (*f*)
Vacancy diffusion	
Diamond	0.50
Tungsten (A2, body-centred cubic)	0.73
Copper (A1, face-centred cubic)	0.78
Magnesium (A3, hexagonal close-packed), 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

All data from Compaan and Haven, *Trans. Faraday Soc.*, **52**: 786–801 (1956); **54**: 1498–1508 (1958).

7.12 Ionic conductivity

7.12.1 Ionic Conductivity in Solids

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:

$$\sigma = c_\mathrm{a}\, Z_\mathrm{a} e\, \mu_\mathrm{a} + c_\mathrm{c}\, Z_\mathrm{c} e\, \mu_\mathrm{c}$$

where σ is the conductivity, c_a, c_c represent the concentrations of mobile anions and cations, $Z_\mathrm{a}e$, $Z_\mathrm{c}e$ 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:

$$\sigma_\mathrm{c} = t_\mathrm{c}\sigma \quad \sigma_\mathrm{a} = t_\mathrm{a}\sigma$$

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:

$$\sigma = \sigma\,(t_\mathrm{c} + t_\mathrm{a})$$
$$t_\mathrm{c} + t_\mathrm{a} = 1$$

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 7.12) by reducing it by $^1\!/_2\, ZeaV$ in one direction and raising it by the same amount in the other (Figure 7.16).

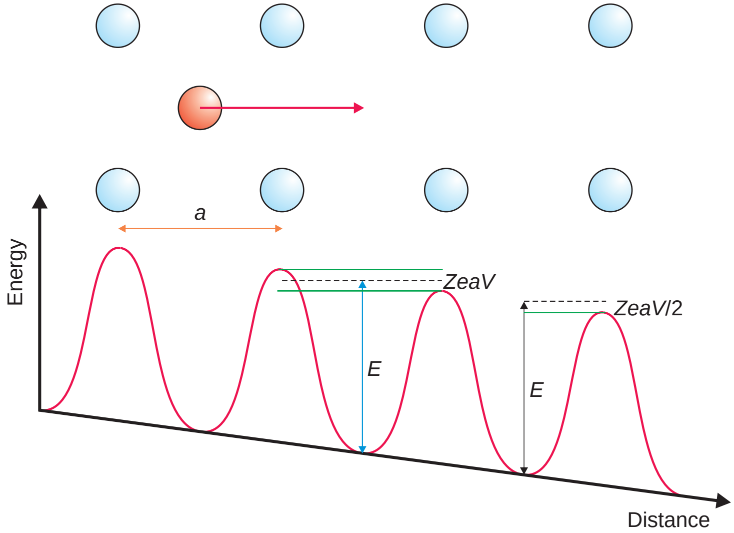


Figure 7.16 Energy barriers to be surmounted by an ion during ionic conductivity.

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:

$$\Gamma_+ = \nu \exp\left(\frac{-(E - \frac{1}{2}ZeaV)}{k_\mathrm{B}T}\right)$$

where the potential barrier is $(E - {}^1\!/_2\, ZeaV)$ and ν is the vibration frequency of the solid. In a direction against the field the number of successful jumps will be given by:

$$\Gamma_- = \nu \exp\left(\frac{-(E + \frac{1}{2}ZeaV)}{k_\mathrm{B}T}\right)$$

where the potential barrier is $(E + {}^1\!/_2\, ZeaV)$. 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:

$$\begin{aligned} \mathrm{v} &= \nu a \left[\exp\left(\frac{-(E - \frac{1}{2}ZeaV)}{k_\mathrm{B}T}\right) - \exp\left(\frac{-(E + \frac{1}{2}ZeaV)}{k_\mathrm{B}T}\right)\right] \\ &= \nu a \exp\left(\frac{-E}{k_\mathrm{B}T}\right)\left[\exp\left(\frac{ZeaV}{2k_\mathrm{B}T}\right) - \exp\left(\frac{-ZeaV}{2k_\mathrm{B}T}\right)\right] \end{aligned}$$

For low field strengths $ZeaV$ is much less than $k_\mathrm{B}T$, and:

$$\exp\left(\frac{ZeaV}{2k_\mathrm{B}T}\right) - \exp\left(\frac{-ZeaV}{2k_\mathrm{B}T}\right)$$

may be replaced by[1]:

$$\frac{ZeaV}{k_B T}$$

Using this approximation:

$$\mathrm{v} = \frac{v\,Z\,a^2 eV}{k_B T}\exp\left(\frac{-E}{k_B T}\right)$$

The mobility, μ, of the ion is defined as the velocity when the value of V is unity, so:

$$\mu = \frac{v\,Z\,a^2 e}{k_B T}\exp\left(\frac{-E}{k_B T}\right)$$

Often a geometrical factor g is included to take into account different crystallographic features to give:

$$\mu = \frac{gv\,Z\,a^2 e}{k_B T}\exp\left(\frac{-E}{k_B T}\right)$$

Substituting for μ in the equation: $\sigma = c\,Ze\,\mu$ gives:

$$\sigma = \frac{cgv\,a^2 Z^2 e^2}{k_B T}\exp\left(\frac{-E}{k_B T}\right)$$

This equation takes the form:

$$\sigma\,T = \sigma_0\exp\left(\frac{-E}{k_B T}\right) \qquad (7.10)$$

where

$$\sigma_0 = \frac{cgv\,a^2 Z^2 e^2}{k_B}$$

When ion migration takes place via a vacancy diffusion mechanism:

$$\sigma_v\,T = \sigma_0\exp\left(\frac{-E_v}{k_B T}\right)$$

and when it takes place by the migration of interstitials:

$$\sigma_i\,T = \sigma_0\exp\left(\frac{-E_i}{k_B T}\right)$$

The value of E_v or E_i can be obtained from the Arrhenius-like plots of ln σT versus 1/T (Figure 7.17). Note T is not constant, and a plot of ln σ versus 1/T will not give a true value for E.

In the case of very pure solids, it is necessary to take into account the number of intrinsic defects (Section 7.10). The value of c must then reflect the type of intrinsic defect present. For example, should Schottky defects predominate, substitution of equation (3.2) for c in equation (7.10) gives:

$$\sigma_S = \sigma_0\exp\left(\frac{-E_v}{k_B T}\right)\exp\left(\frac{-\Delta H_S}{2k_B T}\right)$$

Should Frenkel defects predominate, substituting equation (3.4):

$$\sigma_F = \sigma_0\exp\left(\frac{-E_i}{k_B T}\right)\exp\left(\frac{-\Delta H_F}{2k_B T}\right)$$

In these cases, Arrhenius plots of ln σT versus 1/T will show a knee between the low-temperature and high-temperature regions (Figure 7.16). The

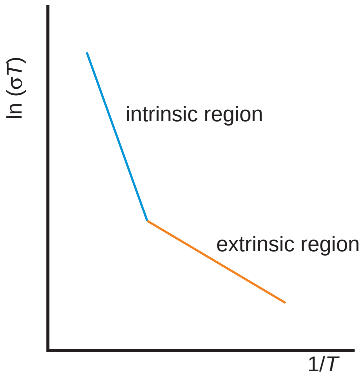


Figure 7.17 Arrhenius-type plot of ln (σT) versus 1/T for ionic conductivity in an almost pure crystal.

[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}.

high-temperature value for E will be composed of two terms:

$$E_{\mathrm{S}} = E_{\mathrm{v}} + \frac{\Delta H_{\mathrm{S}}}{2}$$

for Schottky defects, and:

$$E_{\mathrm{F}} = E_{\mathrm{i}} + \frac{\Delta H_{\mathrm{F}}}{2}$$

for Frenkel defects.

7.12.2 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:

$$D = g\,\nu\,a^2 \exp\left(\frac{-E}{k_{\mathrm{B}}T}\right)$$

The ionic conductivity of an ion is:

$$\sigma = \frac{cg\nu\,a^2Z^2e^2}{k_{\mathrm{B}}T} \exp\left(\frac{-E}{k_{\mathrm{B}}T}\right)$$

leading to the *Nernst-Einstein equation:*

$$D_\sigma = \frac{k_{\mathrm{B}}T\sigma}{c\,Z^2e^2}$$

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:

$$H_{\mathrm{R}} = \frac{D^*}{D_\sigma}$$

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 7.11):

$$f = \frac{D^*}{D_r}$$

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:

$$H_{\mathrm{R}} = \frac{D^*}{D_\sigma} = f$$

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:

$$D_\sigma = \frac{f\,k_{\mathrm{B}}T\sigma}{c\,Z^2e^2}$$

where f is the correlation factor appropriate to the diffusion mechanism. For vacancy diffusion in a cubic structure,

$$D_\sigma = \frac{f_{\mathrm{v}}\,k_{\mathrm{B}}T\sigma}{c\,Z^2e^2}$$

where f_{v} is the correlation factor for vacancy self-diffusion, and:

$$D_\sigma = \frac{f_{\mathrm{iy}}\,k_{\mathrm{B}}T\sigma}{c\,Z^2e^2}$$

for interstitialcy diffusion, where f_{iy} is the appropriate interstitialcy correlation factor.

Further reading

Kirkaldy, J.S. and Young, D.J. (1987) *Diffusion in the Condensed State*. Institute of Metals, London.

LeClaire, A.D. (1976) Chapter 1, in *Treatise on Solid State Chemistry* (ed. N.B. Hannay). Plenum, New York.

Metselaar, R. Diffusion in solids, Parts 1–3. *J. Materials Education*, **6**: 229 (1984); **7**: 653 (1985); **10**: 621 (1988).

Open University Course Team (2009) *Random Walks and Diffusion*. Open University Press, Milton Keynes.

Tilley, R.J.D. (2008) *Defects in Solids*, Chapters 5 and 6. John Wiley & Sons, Ltd., Hoboken.

Problems and exercises

Quick quiz

1 Diffusion through a crystalline structure is called:
(a) Tracer diffusion.
(b) Volume diffusion.
(c) Self-diffusion.

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.

3 Tracer diffusion refers to the diffusion of:
(a) Marked atoms.
(b) Traces of impurities.
(c) Single atoms.

4 A diffusion profile is a graph of:
(a) Concentration versus time.
(b) Concentration versus distance.
(c) Distance versus time.

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.

6 What graph would you plot to determine the tracer diffusion coefficient in a diffusion couple?
(a) c versus x.
(b) ln c versus x^2.
(c) ln c versus ln x.

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.

8 In steady-state diffusion the diffusion coefficient is obtained via:
(a) The diffusion equation.
(b) The Arrhenius equation.
(c) Fick's first law.

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.

10 The activation energy for diffusion can be determined by a plot of:
(a) ln D versus 1/T.
(b) ln c versus 1/T.
(c) ln D versus T.

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.

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.

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.

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.

15 In crystals containing Frenkel defects:

(a) One diffusion mechanism is possible.

(b) Two diffusion mechanisms are possible.

(c) Three diffusion mechanisms are possible.

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.

17 Correlation factors take into account:

(a) Non-random diffusion.

(b) Impurity content.

(c) Frenkel and Schottky defects.

Calculations and questions

7.1 Show that the units of the diffusion coefficient are m^2 s^{-1}.

7.2 Show that the two equations are equivalent

(a) $$\frac{c_x - c_0}{c_s - c_0} = 1 - erf\left(\frac{x}{2\sqrt{Dt}}\right)$$

(b) $$\frac{c_s - c_x}{c_s - c_0} = erf\left(\frac{x}{2\sqrt{Dt}}\right)$$

7.3 Radioactive nickel-63 was coated onto a crystal of CoO and made into a diffusion couple. The sample was heated for 30 min 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

7.4 Radioactive cobalt-60 was coated onto a crystal of CoO and made into a diffusion couple. The sample was heated for 30 min 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

7.5 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 s 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 × 10^{-4}
520	3
400	4
270	5
185	6
130	6.5
90	7
53	8

7.6 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 at 1 mm below the surface after a heating time of 2 hours.

7.7 Using the data in Question 7.6, how long would it take to make the carbon content 0.40 wt% at 1 mm below the surface?

7.8 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 1 hour?

7.9 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?

7.10 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 A1 (face-centred cubic) structure, with lattice parameter $a = 0.389$ nm.

7.11 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}

7.12 Using the data in question 7.11, at what temperature (°C) will the penetration depth be 2 μm after 10 hours of heating?

7.13 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}

7.14 Using the data in question 7.13, at what temperature (°C) will the penetration depth be 10^{-4} m after 20 hours of heating?

7.15 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}

7.16 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}

7.17 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}

7.18 Calculate the diffusivity of ^{51}Cr in titanium metal at 1000 °C. $D_0 = 1 \times 10^{-7}\,m^2\,s^{-1}$, $E = 158\,kJ\,mol^{-1}$.

7.19 Calculate the diffusivity of ^{51}Cr in a titanium–18 wt% Cr alloy at 1000 °C. $D_0 = 9 \times 10^{-2}\,m^2\,s^{-1}$, $E = 186\,kJ\,mol^{-1}$.

7.20 Calculate the diffusivity of ^{55}Fe in forsterite, Mg_2SiO_4, at 1150 °C. $D_0 = 4.17 \times 10^{-10}\,m^2\,s^{-1}$, $E = 162.2\,kJ\,mol^{-1}$.

7.21 Calculate the diffusivity of ^{18}O in Co_2SiO_4 at 1250 °C. $D_0 = 8.5 \times 10^{-3}\,m^2\,s^{-1}$, $E = 456\,kJ\,mol^{-1}$.

7.22 Calculate the diffusivity of Li in quartz, SiO_2, parallel to the c-axis, at 500 °C. $D_0 = 6.9 \times 10^{-7}\,m^2\,s^{-1}$, $E = 85.7\,kJ\,mol^{-1}$.

7.23 The diffusion coefficient of Ni^{2+} tracers in NiO is $1 \times 10^{-15}\,m^2\,s^{-1}$ at 1100 °C. Estimate the penetration depth of the radioactive Ni^{2+} ions into a crystal of NiO after heating for 1 hour at 1100 °C.

7.24 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 $1 \times 10^{-11}\,m^2\,s^{-1}$.

7.25 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 $127\,kJ\,mol^{-1}$.

7.26 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.

7.27 The ionic conductivity of F^- ions in the fast ionic conductor $Pb_{0.9}In_{0.1}F_{2.1}$ at 423 K is $1 \times 10^{-4}\,\Omega^{-1}\,m^{-1}$. 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.

7.28 The conductivity of SrO (halite (B1) 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.

8

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 or nanostructure of the solid. Phase diagrams (Chapter 4) can be used to determine which phases occur when the temperature, pressure or composition of the system is specified, and how these 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 changes that involve little or no change of composition (sections 8.1–8.5). This is followed by a discussion of a few of the consequences that compositional change necessitates (sections 8.6–8.8).

8.1 Sintering

8.1.1 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. The first stage that occurs during sintering 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 8.1). Usually this involves shrinkage, change of shape and the formation of pores. Finally, the solid becomes denser by the elimination of internal pores and voids, resulting in a compact grain structure. Sintering is widely used in the ceramics and powder metallurgy industries to form small components such as valves, and in the polymer industry to fabricate porous polymer blocks used in filtration.

In general, the procedure involves the careful preparation of powders with a suitable particle size

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

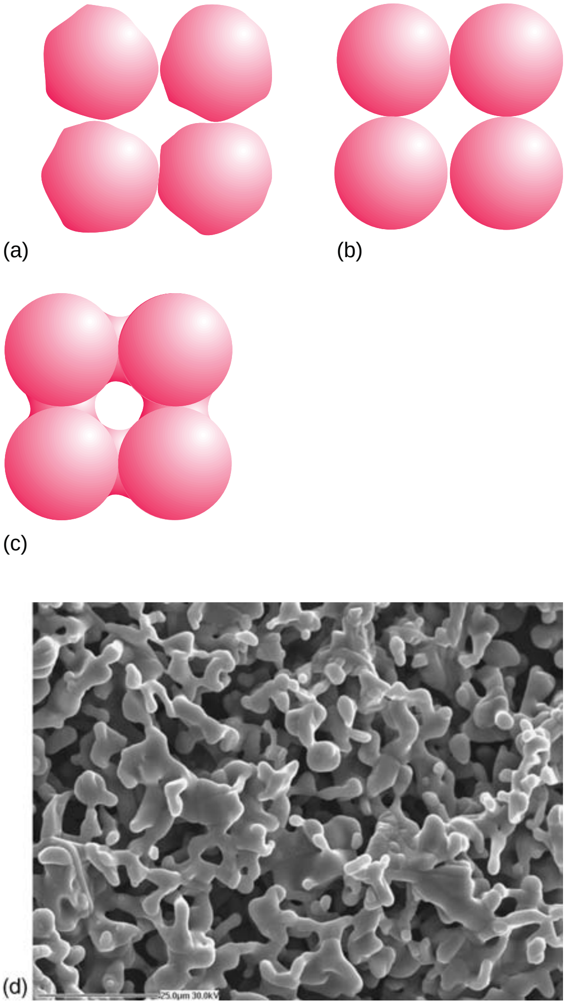


Figure 8.1 Stages in sintering: (a) an initial compact of slightly uneven particles; (b) decrease in surface roughness; (c) transformation into a porous solid; (d) scanning electron-micrograph of lightly sintered $Sr_2Ti_2O_7$ grains.

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.

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 whilst 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. A technique that is being widely used, especially for the production of biomedical components, is *selective laser sintering* (*SLS*). In this process, a bed of powder, either single-component, coated or mixed, is heated to a point close to but below its melting point for ceramics and metals, and to near the glass transition temperature for polymers. The powder bed is heated in a defined pattern via a high-power laser controlled by a computer routine. In this procedure, note that, unlike in classical sintering, the particles actually melt on the surface and fuse together, largely pulled together by surface tension. Once one layer is treated, a new bed of powder is laid down on top of the first and the whole process repeated. 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 a complicated three-dimensional shape can be fabricated.

Sintering can be brought about by a variety of reactions. Of these, material transport by viscous flow is important in glasses but less so in metals or ceramics. Evaporation and condensation is important in rather volatile compounds such as halides and some oxides. Diffusion, both bulk and

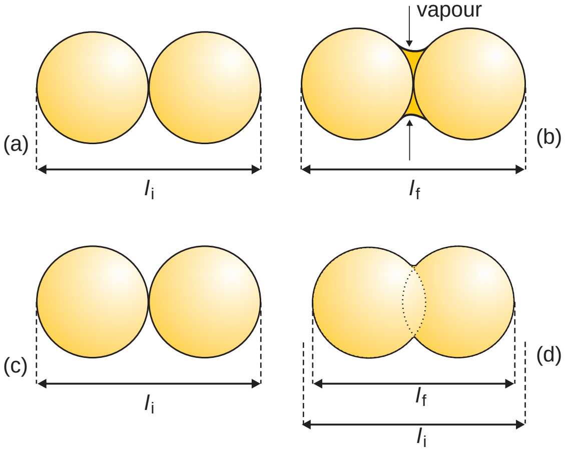


Figure 8.2 Sintering via vapour transport (a, b) does not cause shrinkage. Sintering via solid-state diffusion (c, d) causes shrinkage. l_i, l_f are initial and final lengths.

short-circuit, is important for refractory materials, and for these the presence of traces of liquid phase are also greatly beneficial in speeding up the reaction. In practice, the mechanism that operates has 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 8.2). It is also necessary to remember that many of the properties of a sintered object, such as mechanical strength, and the optical, electrical and magnetic properties, differ from those of a bulk material. Sintering practice must take these end properties into account when the overall production process is designed.

8.1.2 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 8.3a). The vapour pressure over a curved surface p is related to the vapour pressure over a flat surface p_0 by the *Kelvin equation*:

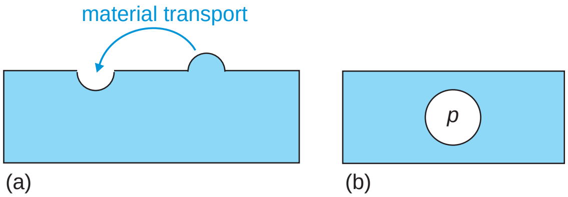


Figure 8.3 (a) Material transports from hills to valleys during vapour-phase sintering. (b) The excess pressure p inside a void balances the surface energy of the void.

$$RT \ln\left(\frac{p}{p_0}\right) = \frac{2V\gamma}{r} = \Delta G$$

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:

$$RT \ln\left(\frac{p}{p_0}\right) = \frac{2V\gamma}{-r} = \Delta G$$

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.

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 8.3b), the excess pressure in the pore, p, needed to balance the surface energy is given by:

$$p = \frac{2\gamma}{r}$$

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:

$$\frac{r_\mathrm{i}}{r_\mathrm{f}} = \sqrt{\frac{2\gamma}{p_\mathrm{i} r_\mathrm{i}}}$$

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 $r_\mathrm{i} = r_\mathrm{f}$, that is:

$$p_\mathrm{i} r_\mathrm{i} = 2\gamma$$

Material transport also occurs (for a similar reason) when two spheres touch to form a neck (Figure 8.2a). 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.

The transport of material to achieve sintering is not restricted to transport via the vapour phase. Bulk or surface diffusion can also be called into play to achieve the same result. In fact, many refractory oxides have very low vapour pressures so that solid-state diffusion becomes paramount in importance.

8.1.3 The kinetics of neck 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 50 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:

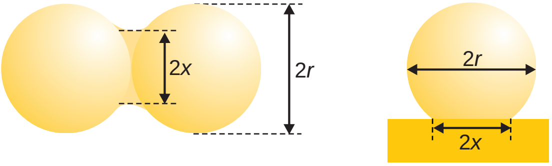


Figure 8.4 Definition of neck radius, x, and sphere radius, r, used in the kinetics of sintering.

$$\begin{aligned}
&\text{Viscous flow :} && \frac{x^2}{r} = k_1 t\\
&\text{Surface diffusion :} && \frac{x^5}{r^2} = k_2 t\\
&\text{Bulk diffusion :} && \frac{x^3}{r} = k_3 t\\
&\text{Vapour transport :} && \frac{x^7}{r^3} = k_4 t
\end{aligned}$$

In these equations x and r are the neck and sphere radii (defined in Figure 8.4), k_1, k_2, k_3 and k_4 are constants which vary with temperature, and t is the time of reaction.

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.

8.2 First-order and second-order 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*.

8.2.1 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 8.5a, b). 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:

$$\left(\frac{\partial g}{\partial T}\right)_P^{\text{high}} - \left(\frac{\partial g}{\partial T}\right)_P^{\text{low}} = s_\mathrm{m}^{\text{high}} - s_\mathrm{m}^{\text{low}} = \Delta s_\mathrm{m}$$

The slope $(\partial g/\partial P)_T$ of the g versus P curve is the molar volume of the phase, $\mathrm{v_m}$, and the difference in the slopes at the transition point is simply the change in molar volume occurring at the transition, $\Delta \mathrm{v_m}$:

$$\left(\frac{\partial g}{\partial P}\right)_T^{\text{high}} - \left(\frac{\partial g}{\partial P}\right)_T^{\text{low}} = \mathrm{v}_\mathrm{m}^{\text{high}} - \mathrm{v}_\mathrm{m}^{\text{low}} = \Delta \mathrm{v_m}$$

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. Rocksalt 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 atmosphere pressure. Similar transitions occur in many MX halides and oxides.

8.2.2 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 fact that although the first derivatives 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 8.5c,d). 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 $-c_P/\mathrm{T}$, where c_P is the molar specific heat. The derivative $(\partial^2 g/\partial P^2)_T$ is equal to the volume term $-\mathrm{v}\beta$, where β is the compressibility. Second-order transitions include ferroelectric to paraelectric transformations (Chapter 11), ferromagnetic to paramagnetic transformations (Chapter 12) or superconducting to metallic or semiconducting transformations (Chapter 13).

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

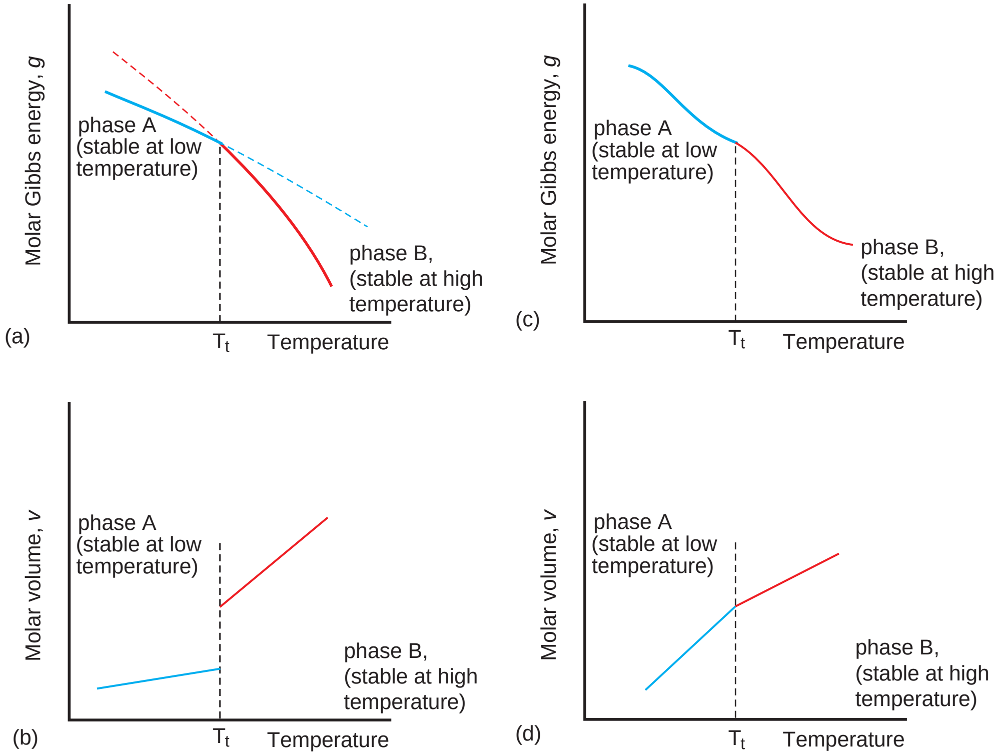


Figure 8.5 First-order phase transitions: (a) the variation of molar Gibbs energy with temperature; (b) variation of molar volume with temperature. Second-order phase transitions: (c) the variation of molar Gibbs energy with temperature; (d) variation of molar volume with temperature.

always be categorically made (see the notes on C_{60}, Section 8.4.2).

8.3 Displacive and reconstructive transitions

8.3.1 Displacive transitions

In crystallographic terms it is convenient to consider a phase transformation as a structural change. 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 breakage (Figure 8.6a–c). 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).

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 8.7a). 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 8.7b,c), resulting in an orthorhombic unit cell. At a temperature of 523 K thermal vibrations are sufficient to disrupt the distortion so

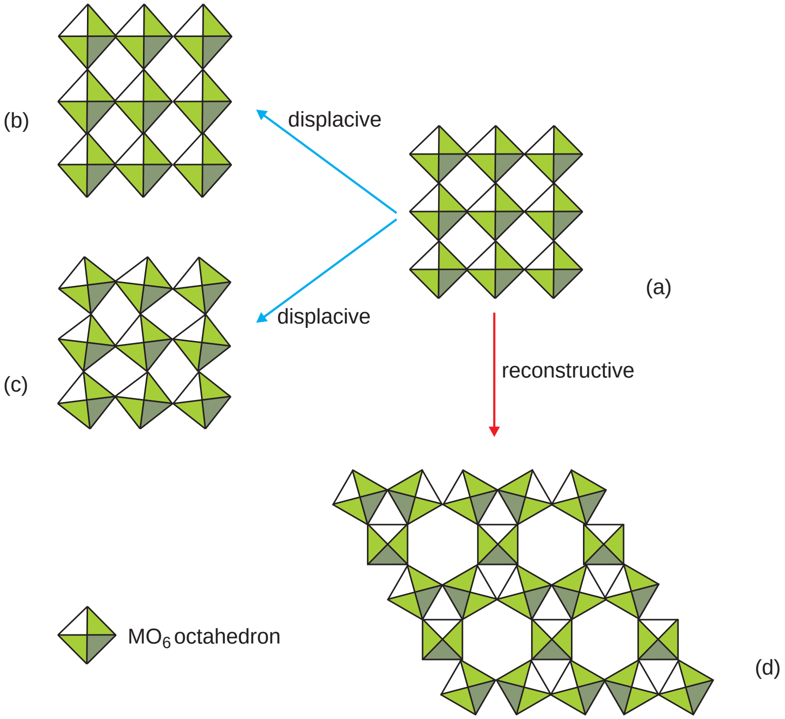


Figure 8.6 Displacive and reconstructive transitions, schematic: (a) undistorted octahedra; (b) octahedra lengthened axially; (c) octahedra rotated; (d) octahedra rearranged.

that the octahedra become regular and the unit cell becomes tetragonal.

Displacive transformations involving octahedral rotation have been studied extensively in the perovskite oxides. An example is provided by $CaTiO_3$. At room temperature the structure is orthorhombic with "rotated" TiO_6 octahedra (Figure 8.8a). The phase transforms by a discontinuous (first-order) transition to tetragonal at 1498 K and then to cubic (Figure 8.8b) 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.

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 (Section 11.2.2, Figures 11.7 and 11.8). Other examples include important transitions in ferroelectric ceramics that include hydrogen-bonded materials in which the hydrogen in the bond is offset to one side of centre at low temperatures, but at higher temperatures the hydrogen is found to occupy, at least statistically, a central position (Section 11.3.5, Figure 11.20), and the perovskite oxide $BaTiO_3$, in which distortions of TiO_6 octahedra give rise to the ferroelectric effect (Section 11.3.7, Figure 11.22).

8.3.2 Reconstructive transitions

Reconstructive transitions describe the situation where a major reorganisation of a crystal structure occurs during which primary bonds are broken and reformed (Figure 8.6d). 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

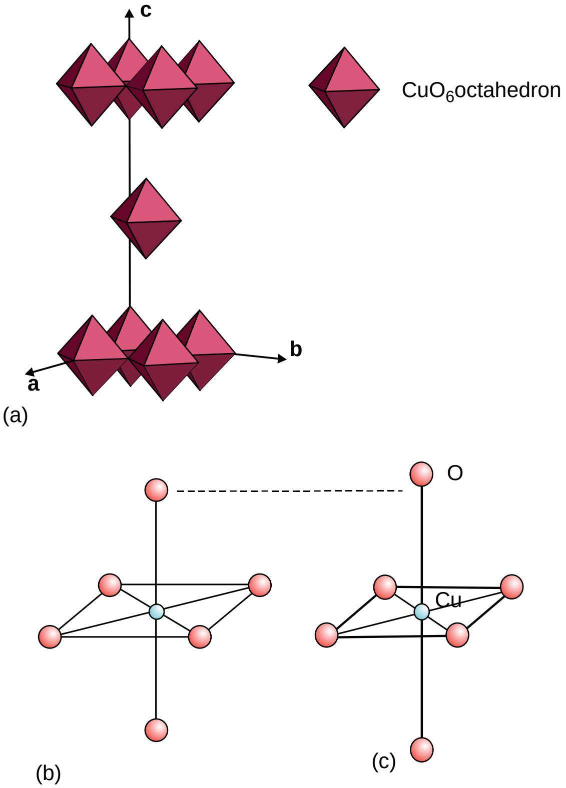


Figure 8.7 Displacive transition at 523 K in La_2CuO_4 due to distortion of CuO_6 octahedra: (a) idealised sheets of CuO_6 octahedra forming the skeleton of the phase; (b) regular octahedron present above 523 K; (c) elongated octahedron present below 523 K (exaggerated).

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.

An example of importance to studies of the Earth is the transformation of minerals with the olivine structure to the spinel structure under the high pressures that occur deep within the mantle. 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 an eighth 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 (sections 5.3.9 and 5.4.3, Table 5.2). 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 an eighth 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 (e.g. 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.

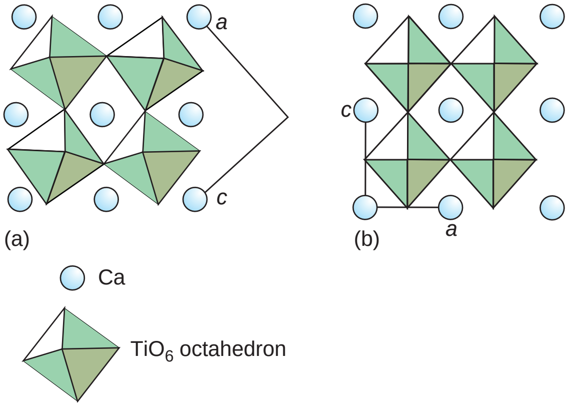


Figure 8.8 Displacive transformation in $CaTiO_3$: (a) the room-temperature orthorhombic phase with rotated and slightly deformed octahedra; (b) the high-temperature cubic form with regular octahedra.

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

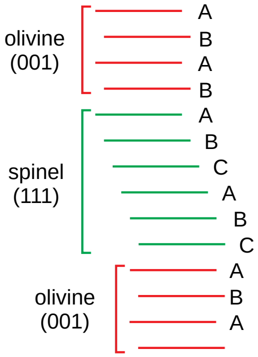


Figure 8.9 A reconstructive transformation in which the ABAB packing of (001) oxygen atom planes in olivine shear in a topotactic way to become ABCABC packing of (111) oxygen planes in spinel.

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 no correspondence between the crystal structures of the parent and product material. An alternative mechanism involves converting the ABAB packing in olivine into ABCABC packing by a shearing of the structure followed by subsequent diffusion of the cations into new sites (Figure 8.9). 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.

8.4 Order-disorder transitions

An *order–disorder transition*, as the name states, is a transition in which the degree of order in a phase changes. Such transitions occur frequently in both metal alloy and ceramic phases. 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, Section 8.4.1.
- Orientational disordering, Section 8.4.2.
- Ordering of electronic or nuclear spin states, giving rise to magnetism (Chapter 12).

8.4.1 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 time-scales 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 (Section 6.1.3.1, Figure 6.3a). Similar order–disorder occurs in many non-metallic compounds. For example, the oxide $LiFeO_2$, prepared at high temperatures, adopts the halite structure in which the cations are distributed at random over the available octahedral sites (Chapter 5, Sections 5.3.8, 5.4.2). On annealing this phase at lower temperatures, the cations will order, leading to a tetragonal unit cell distortion. A similar transformation has been recorded with the phase Li_2NiTiO_4. This phase adopts a disordered halite 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 the oxide spinels (Section 5.3.9), which have a general formula of AB_2O_4 with cations able to 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 sulphides related to TiS_2 can be considered to contain Ti interstitials. The structure of these phases consists of sheets of composition TiS_2 (Figure 8.10). 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 of 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.

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.

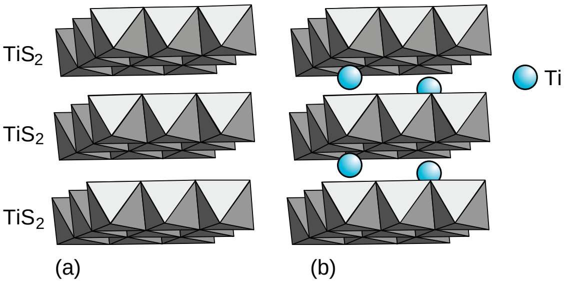


Figure 8.10 (a) The idealised structure of TiS_2, consisting of sheets of TiS_6 octahedra. (b) The idealised structure of phases with compositions between TiS_2 and TiS. The additional Ti interstitials are distributed over octahedral sites between the layers.

Other examples of order–disorder transitions are given in Section 8.5.3 and Chapter 11).

8.4.2 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 lowest temperature 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 (Sections 11.2 and 11.3). 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 $+\mathbf{b}$-axis at low temperatures but orient randomly in either the $+\mathbf{b}$ or $-\mathbf{b}$ direction at high temperatures (Section 11.3.6). 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 8.11a,b). Linear groups, such as polar HX molecules in the halogen halides, also adopt a single direction at lower temperatures, and random directions constrained only by the crystal structure at higher temperatures (Figure 8.11c,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 halite.

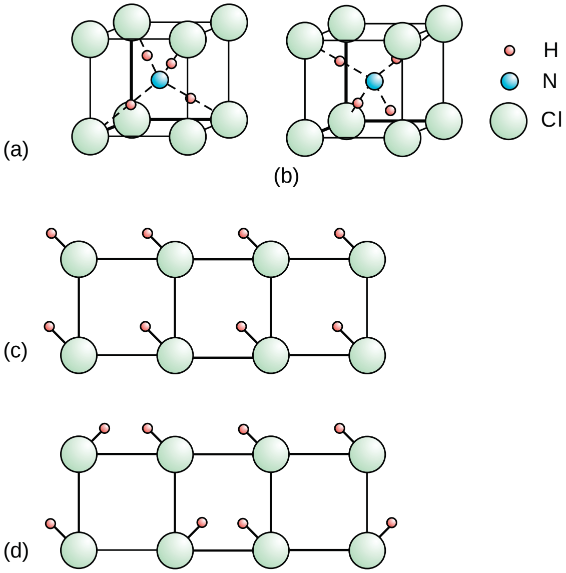


Figure 8.11 Orientational transformations: (a, b) alternative orientations of NH^{4+} tetrahedra as in ammonium halides; (c, d) alternative orientations of linear HCl molecules as in hydrogen halides.

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. The C atoms are situated at the vertices of a truncated icosahedron, to produce a shape that resembles a soccer ball (Figure 3.5a). 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'.

8.5 Martensitic transformations

Martensitic transformations are essentially solid-state displacive transitions that take place at a constant temperature and composition due to synchronized 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.

8.5.1 The austenite–martensite transition

Rapidly cooling the homogeneous alloy austenite (the γ phase, face-centred cubic, A1 structure) to room temperature, (a process called *quenching*), leads to the formation of *metastable* martensite with a tetragonal structure that is a slightly distorted body-centred cubic (A2) structure (the α' phase). The lattice parameters of the α' phase vary with carbon content, a taking a value of approximately 0.285 nm and c taking values between 0.292 and 0.300 nm. The ratio of c/a 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 interstitial sites in austenite (Figure 6.4), and this does not change on transformation, so that they remain in octahedral sites in body-centred martensite (Figure 6.5). Once a nucleus of a martensite crystal forms it will grow to its final size in a time of the order of 10^{-7} sec. 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. This 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*. Cubic austenite can be represented in terms of a body-centred tetragonal unit cell with $a(\text{tet}) = a/\sqrt{2}$, $c(\text{tet}) = a$ (Figure 8.12). To convert this prototype martensite unit cell into the true body-centred cubic unit cell of martensite, a(tet) must increase by about 22% at the same time as c (tet) 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:

$(111)_\gamma \rightarrow (011)_{\alpha'}$
$[110]_\gamma \rightarrow [100]_{\alpha'}$
$[101]_\gamma \rightarrow [111]_{\alpha'}$
$[112]_\gamma \rightarrow [011]_{\alpha'}$

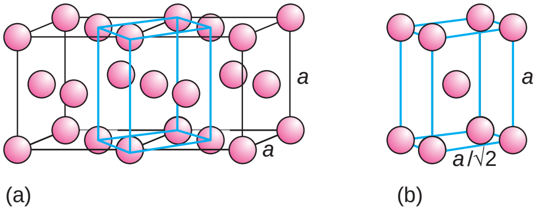


Figure 8.12 (a) Two unit cells of the cubic A1 structure of austenite with an alternative body-centred tetragonal unit cell outlined. (b) The dimensions of the body-centred tetragonal unit cell prototype of martensite.

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, and is called *plate martensite*. At compositions between 0.6 and 1.0 wt% carbon, a mixture of the two forms is found. 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 8.13), and usually

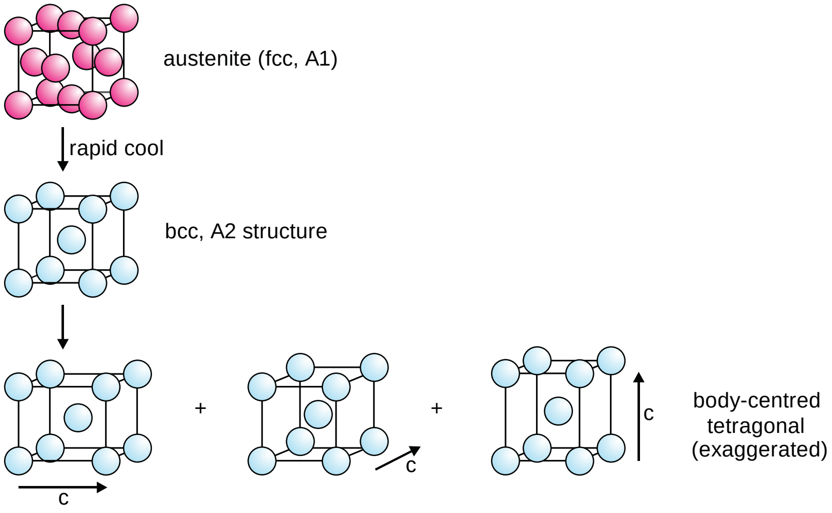


Figure 8.13 Schematic crystallographic changes occurring during the formation of martensite from austenite.

several different *variants* or *domains* (i.e. orientations) occurs in an austenite grain.

The production of each martensite platelet involves a shear parallel to the plate and usually an expansion normal to it. These cause considerable strain in the two-phase matrix, which 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 8.14a). 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 8.14b). In addition the platelets show extensive slip lines and large numbers of closely-spaced twins, called *polysynthetic twinning*. These features arise because the deformation of the platelets (Figure 8.15a,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 either by repeated slip, which may be brought about by the passage of dislocations through the crystal (Figure 8.15c, also see Figure 3.15), or by polysynthetic twinning (Figure 8.15d). These alternatives and the underlying transition mechanism are still under active study.

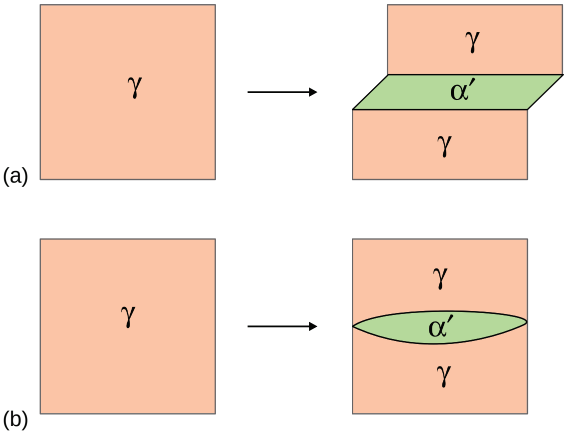


Figure 8.14 (a) The unconstrained transformation of austenite (γ phase) to martensite (α′ phase). (b) The constrained transformation produces lens-shaped volumes.

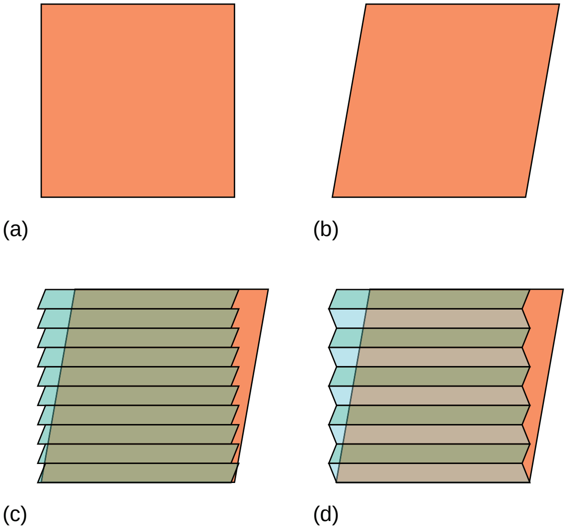


Figure 8.15 The shape change caused by shear deformation (a, b) can be minimised by repeated slip (c) or polysynthetic twinning (d).

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 8.16). The martensite start temperature for a steel depends upon a large numbers 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

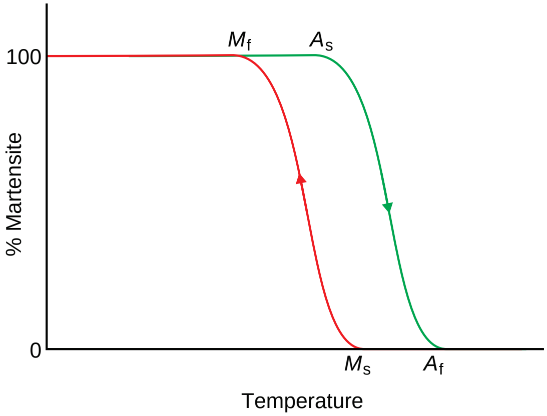


Figure 8.16 Hysteresis (schematic) between martensite start and finish temperatures, M_s and M_f, and austenite start and finish temperatures, A_s and A_f, in a typical steel.

the construction of computer models to predict M_s for a wide range of steel-related alloys.

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 (Section 8.6 4). 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.

8.5.2 Martensitic transformations in zirconia

Zirconia, ZrO_2, also known by the mineral name baddeleyite, and isostructural HfO_2 show a progressive transformation from low-symmetry monoclinic to high-symmetry cubic as the temperature rises:

$$ZrO_2 : (\text{monoclinic}) \xrightarrow{1170\,^\circ\text{C}} (\text{tetragonal}) \xrightarrow{2370\,^\circ\text{C}} (\text{cubic})$$

$$HfO_2 : (\text{monoclinic}) \xrightarrow{1720\,^\circ\text{C}} (\text{tetragonal}) \xrightarrow{2600\,^\circ\text{C}} (\text{cubic})$$

The high-temperature phases cannot be quenched to room temperature but can be stabilised by doping with cations such as Ca^{2+} and Y^{3+} (Section 3.4.5, Figure 3.12, Figure 6.13). 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 crystal structures of the tetragonal and monoclinic zirconia phases are slightly distorted versions of the high-temperature fluorite structure cubic phase with lattice parameters:

$$\begin{array}{lll}
\text{Monoclinic}: & a = 0.5193\text{ nm}, & b = 0.5204\text{ nm}, \\
 & c = 0.5362\text{ nm}, & \beta = 99.2^\circ \\
\text{Tetragonal}: & a = 0.5132\text{ nm}, & c = 0.5228\text{ nm} \\
\text{Cubic}: & a = 0.5144\text{ nm} &
\end{array}$$

The structural correspondence (Figure 8.17) 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 the Zr ions appear to be displaced by a shear, rapidly followed by the oxygen ions shuffling into new positions, leading to the growth of heavily twinned martensitic plates in the untransformed matrix.

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 (Figure 8.18). In this context it is of interest to note that the tetragonal form of zirconia can be stabilised to room

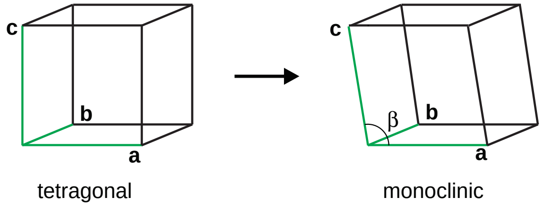


Figure 8.17 The structural relationship between the tetragonal and monoclinic structures of zirconia.

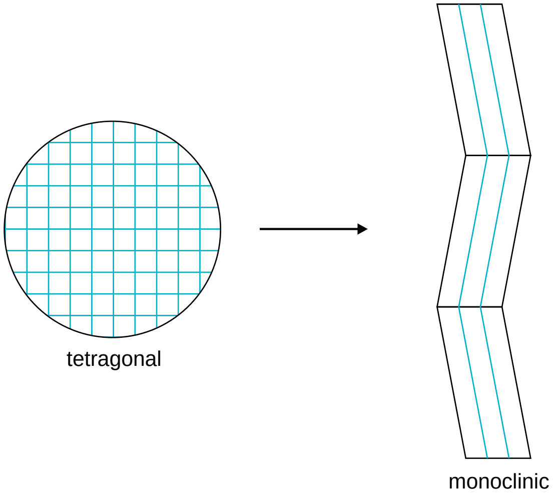


Figure 8.18 The martensitic transformation of spherical tetragonal zirconia nanoparticles produces twinned rod-shaped monoclinic zirconia nanoparticles.

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 10–30 nm diameter.

8.5.3 Martensitic transitions in Ni–Ti alloys

An important martensitic transformation occurs in the titanium–nickel alloy TiNi (Figure 8.19), called Nitinol, used in shape-memory alloys (Section 8.5.4). 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 B2 (CsCl) structure (Figure 8.20a). 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 8.20b,c). If these shear so as to stack in the hexagonal close-packing arrangement ABAB, the resulting structure is the AuCd type (Strukturbericht symbol B19, 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 (Strukturbericht symbol B19′, typical monoclinic lattice parameters of $a = 0.4120$ nm, $b = 0.315$ nm, $c = 0.4623$ nm, $\beta = 96.8°$).

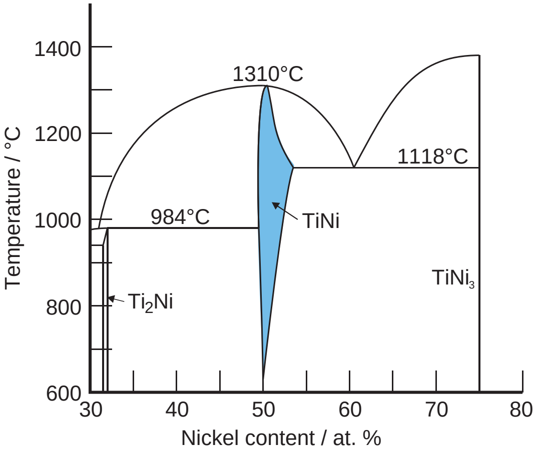


Figure 8.19 Part of the phase diagram of the Ti–Ni system.

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

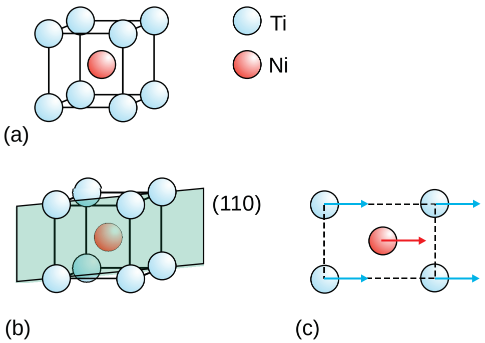


Figure 8.20 The B2 (CsCl) structure of the high-temperature form of TiNi: (a) a unit cell; (b) atoms in the (110) plane; (c) a shear displacement that occurs upon transformation to martensite, arrowed.

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 8.21).

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.

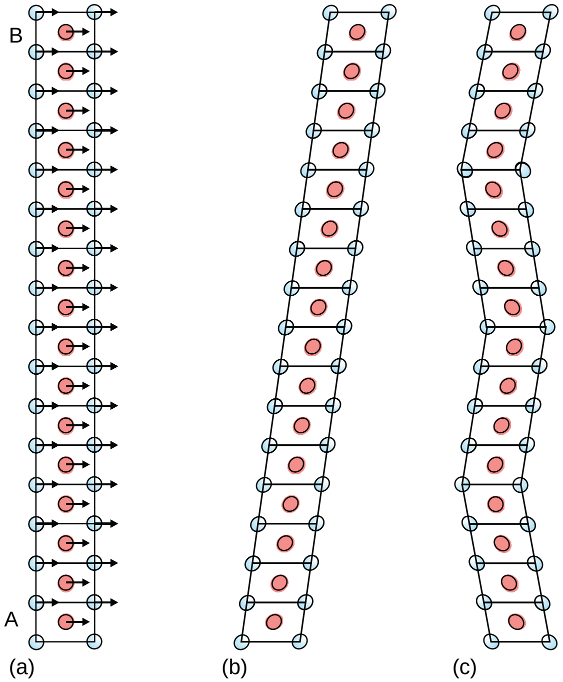


Figure 8.21 (a) A stack of unit cells of NiTi through (110). Shear displacement is applied to each cell in sequence, starting at A and ending at B, to produce a martensitic structure. (b) The deformed structure resulting from shear. (c) The overall shape of the original stack is more nearly maintained by repeated twinning.

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 (Figure 8.19). 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 upon 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.

8.5.4 *Shape-memory alloys*

Shape-memory alloys are 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 spacecraft, 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 tens 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 10.1.5).

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 8.22). 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 8.23).

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 is '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. There is considerable research interest in developing and exploiting two-way shape-memory effect alloys at present.

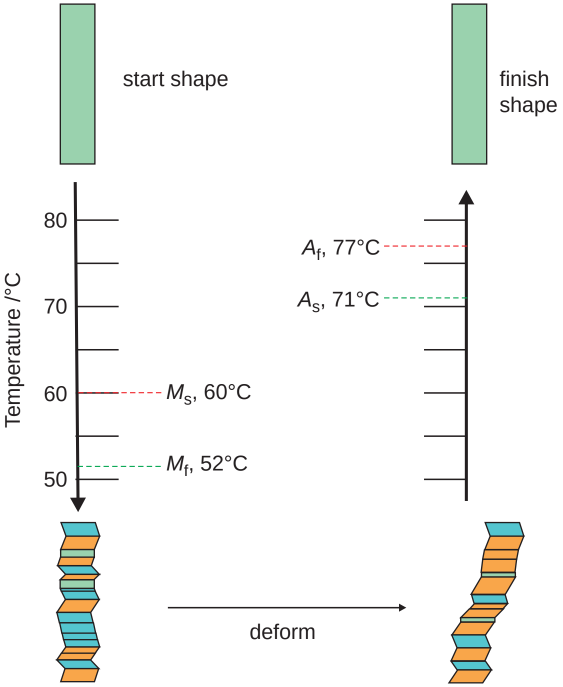


Figure 8.22 The sequence of events taking place during the deformation and recovery of a shape using a shape-memory alloy. Cooling the high-temperature shape below M_f transforms it into a multiply twinned form with the same overall shape. Deformation alters the distribution of the twin boundaries. Reheating the sample above A_f removes the twins and causes the material to revert to the high-temperature form. The temperatures are appropriate to TiNi.

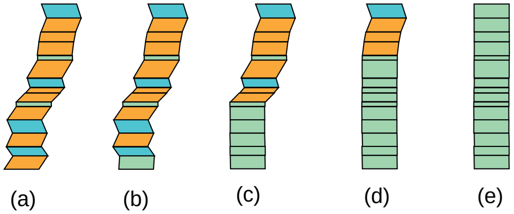


Figure 8.23 (a–e) The progressive recovery of the original shape of a deformed rod by removal of twins.

Nitinol only has a useful temperature working range, when changes due to composition are included, from approximately −20–80 °C. This is rather limiting, as space applications need an application range of approximately −150–120 °C and automotive applications require a range of approximately −50–150 °C. For this reason many metal alloy systems are being investigated experimentally and theoretically using CALPHAD techniques (Section 4.5). Alloys in the Fe-Ni-Co-Al system can be useful from approximately −80–40 °C, and alloys in the Fe-Mn-Al-Ni system have shown shape-memory effects over the increased temperature range of −200–150 °C. The shape-memory effect is also apparent in zirconia ceramics. Practical applications of the transformation are rather limited because of the brittleness of the materials. However, the transformation takes place at a much higher temperature (several hundred degrees) than in Nitinol, making it of potential use in extreme conditions.

8.6 Phase diagrams and microstructures

8.6.1 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 8.24a). The initial composition of the liquid is l_1 (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.

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 8.24b), 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 8.24c). 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 8.24d). 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 and so on, and the composition of the liquid in equilibrium with the solid runs down the liquidus from l_1 to l_2 and so on, 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 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 8.25a). Further cooling in the solid will not initiate change of composition.

8.6.2 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 8.24a). However, there will be insufficient time on further cooling for the original solid to equilibrate, and new material of composition s_2 (Figure 8.24b) 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 8.25b). 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.

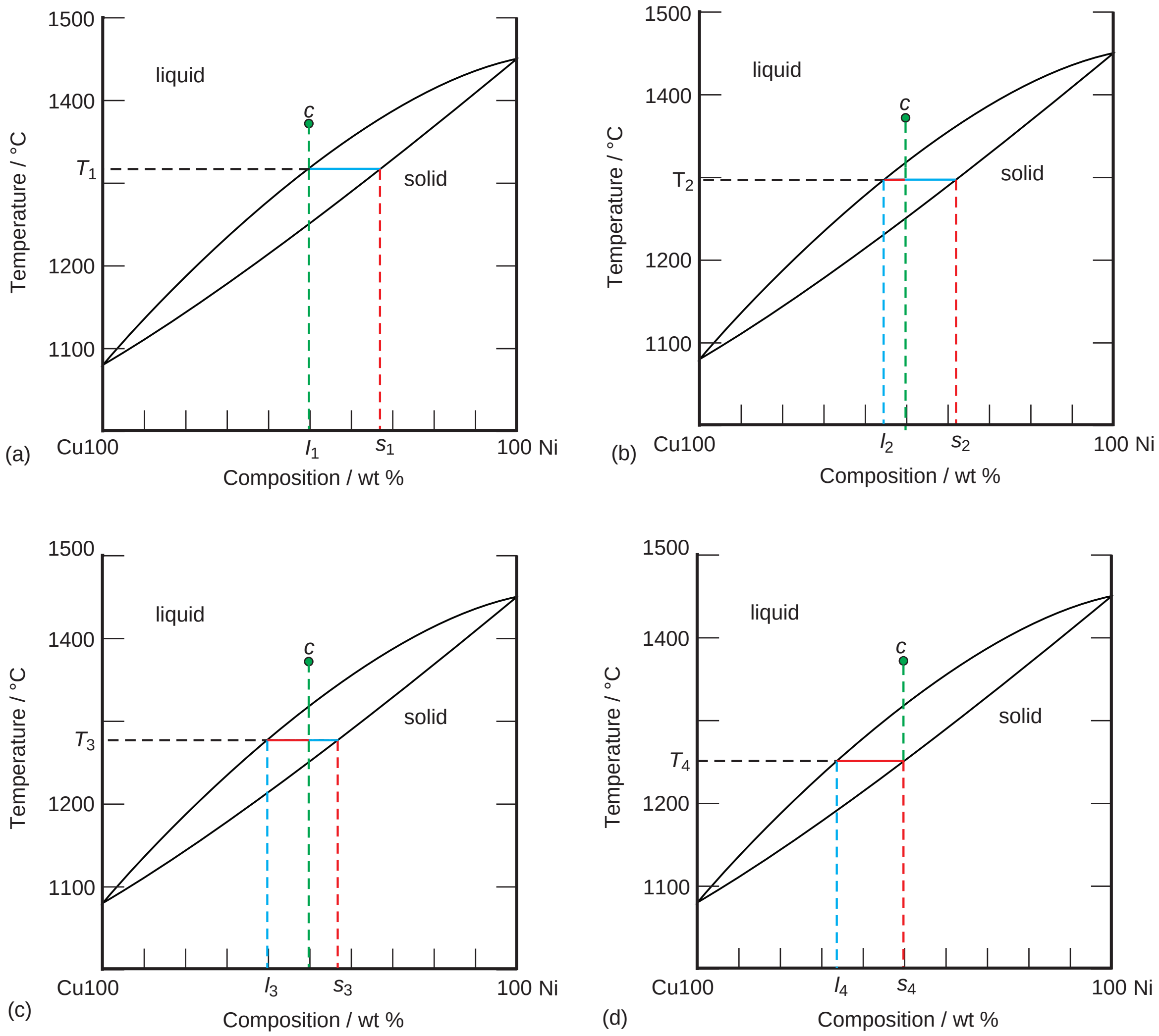


Figure 8.24 A sample of composition *c* in the Cu–Ni system will consist of liquid, liquid plus solid, or solid, depending upon the temperature. (a) At T_1, the liquidus, the liquid has a composition l_1, equal to *c*, and the infinitesimal amount of solid has a composition s_1. (b) At T_2, the liquid has a composition l_2 and the solid, s_2. (c) At T_3, the liquid has a composition l_3 and the solid a composition s_3. (d) At T_4, the solidus, the infinitesimal amount of liquid has a composition l_4 and the solid a composition s_4, equal to *c*.

8.6.3 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 (Section 4.2.4).

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

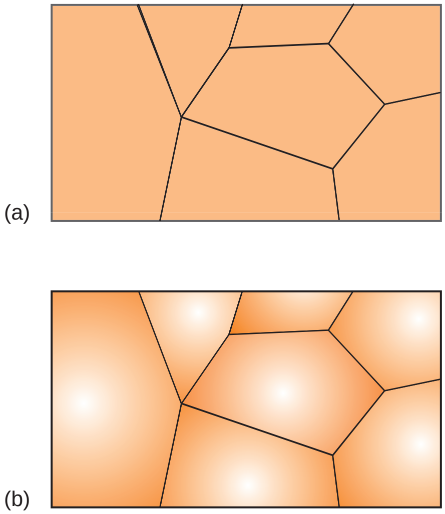


Figure 8.25 Microstructures of a solidified Cu–Ni alloy: (a) after slow cooling the grains are homogeneous; (b) after faster cooling, the grains are richer in one component (Ni) at the grain centres and richer in the other (Cu) at the grain surfaces.

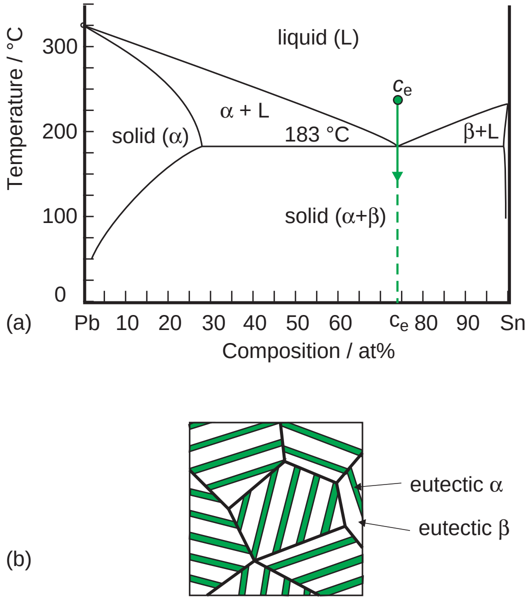


Figure 8.26 Microstructure of solidified Pb–Sn alloys. (a) The eutectic composition c_e. (b) Each resulting grain consists of alternating lamellae of eutectic α and eutectic β.

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 8.26a). 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 8.26b). 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.

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 8.27a), is slowly cooled. At temperature T_1, below the liquidus, some solid α-phase will have nucleated (Figure 8.27b). 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 above 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 8.27a). 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 8.27c). 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.

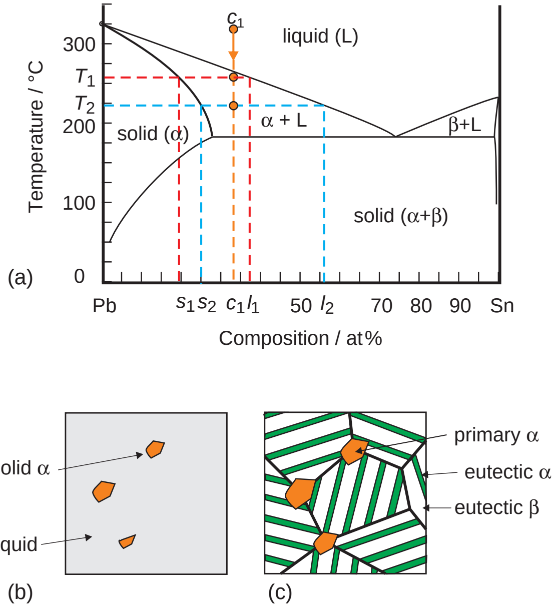


Figure 8.27 Microstructure of solidified Pb–Sn alloys: (a) at a lead-rich composition c_1; (b) cooling into the $(\alpha + \mathrm{L})$ two-phase region results in the formation of crystallites of solid α in a liquid matrix; (c) cooling below the solidus causes the remaining liquid to solidify into alternating lamellae of eutectic α and eutectic β, containing grains of primary α located in the grain boundaries.

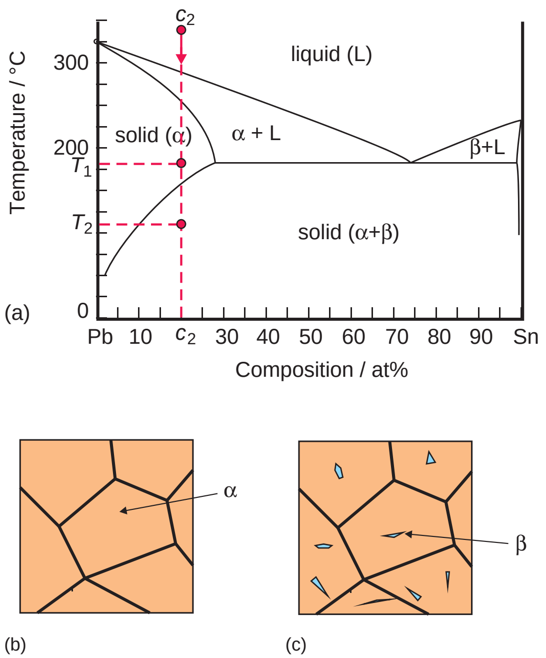


Figure 8.28 Microstructure of solidified Pb–Sn alloys: (a) at a very lead-rich composition c_2; (b) cooling to T_1 produces homogeneous grains of composition c_2; (c) cooling to T_2 results in the formation of small precipitates of β in each grain of α.

In the case of a very lead-rich composition such as c_2 (Figure 8.28a), 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 8.28b). On cooling further, at temperature T_2, for example (Figure 8.5a), 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 8.28c).

The compositions and amounts of the phases present at all times can be calculated by use of tie lines and the lever rule (Section 4.2).

8.6.4 Equilibrium heat treatment of steel in the Fe–C phase diagram

The changes brought about by quenching of austenite (Section 8.5) are quite different from those produced by near-equilibrium cooling of austenite. The principle microstructures found can be explained by reference to the partial iron (Fe)–carbon (C) phase diagram (Figure 8.29).

The most important transformation point in the Fe–C phase diagram from the point of view of these 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 8.29a), an alloy with the eutectoid composition, c_e, is single phase austenite (Figure 8.29b). The face-centred cubic structure of austenite is strained by the interstitial carbon atoms, and this strain 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 ferrite (α-ferrite) and cementite (Fe_3C) at a temperature of 727 °C (T_2, Figure 8.29a), with internal strain likely to provide the driving force.

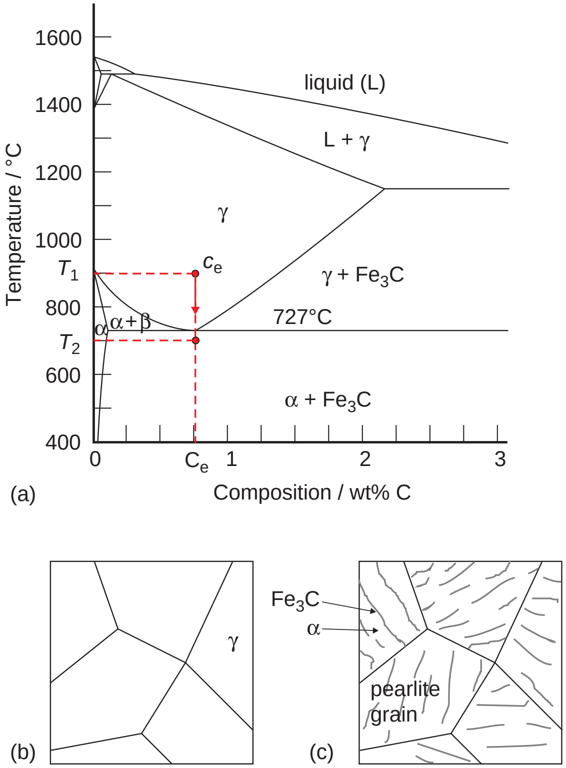


Figure 8.29 (a) The existence diagram of the Fe–C system. (b) The microstructure of the solid with composition c_e, the eutectoid composition, at temperature T_1, consists of γ (austenite). (c) On cooling to T_2 the microstructure consists of grains of pearlite, composed of alternating lamellae of Fe_3C and α.

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. The simultaneous 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 8.5). 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 8.29c), and in this context the phases are called *eutectoid ferrite* and *eutectoid cementite*.

The microstructures 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 8.30a), consists of homogeneous grains of austenite (Figure 8.30b). As these cool slowly, the austenite region is exchanged for a two-phase region (temperature T_2, Figure 8.30a). This material is transformed to ferrite plus austenite. The ferrite often forms at grain boundaries, as the reaction is kinetically favoured in these disordered regions (Figure 8.30c). 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. Ultimately the temperature reaches the eutectoid temperature, 727 °C. Further cooling causes the austenite to transform to pearlite (temperature T_3, Figure 8.30a). The microstructure now consists of pearlite (*eutectoid ferrite* and *eutectoid cementite*), together with the precipitates of the ferrite formed earlier, called *proeutectoid ferrite* (Figure 8.30d).

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 8.31a). The microstructure will

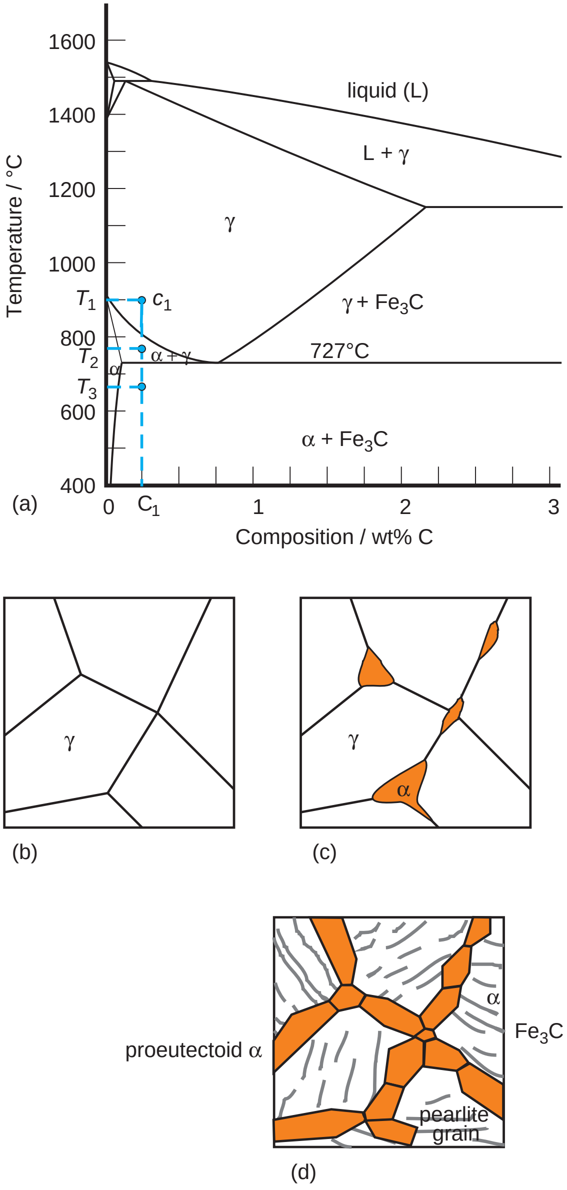


Figure 8.30 (a) The existence diagram of the Fe–C system. (b) The microstructure of the solid with composition c_1, a hypoeutectoid composition, at a temperature T_1. (c) On cooling to T_2 the microstructure consists of grains of γ with precipitates of proeutectoid α at the grain boundaries. (d) At T_3 the microstructure consists of grains of pearlite with proeutectoid α at the grain boundaries.

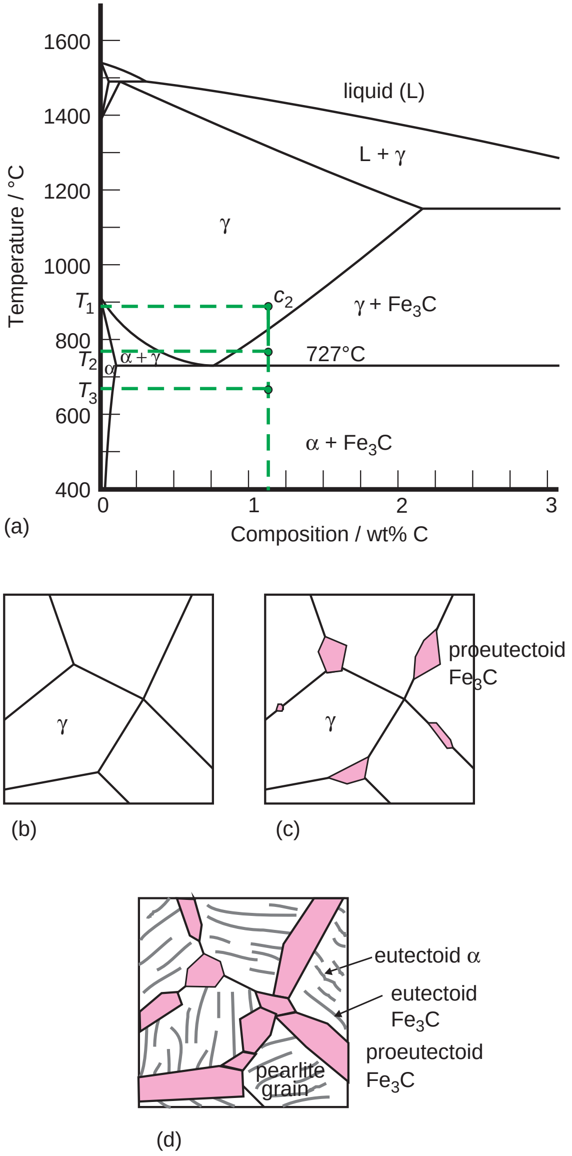


Figure 8.31 (a) The existence diagram of the Fe–C system. (b) The microstructure of the solid with composition c_2, a hypereutectoid composition, at a temperature T_1. (c) On cooling to T_2 the microstructure consists of grains of γ with precipitates of proeutectoid Fe_3C at the grain boundaries. (d) At T_3 the microstructure consists of grains of pearlite with regions of proeutectoid Fe_3C at the grain boundaries.

consist of grains of austenite, (Figure 8.31b). On cooling to temperature T_2 (Figure 8.31a), cementite separates from the austenite, forming preferentially at the grain boundaries (Figure 8.31c). 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 8.31a). The microstructure of the final solid will consist of proeutectoid cementite and pearlite, which itself consists of eutectoid cementite and eutectoid ferrite (Figure 8.31d).

8.7 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 9.4).

8.7.1 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} atmospheres to 10^{-40} atmospheres. As the oxygen partial pressure in air is about a fifth 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 °C 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.

8.7.2 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:

$$x = \sqrt[3]{k_c t} \qquad \text{cubic law}$$

$$x = k_1 - k_2 \ln t \qquad \text{logarithmic law}$$

$$\frac{1}{x} = k_1 - k_2 \ln t \qquad \text{reciprocal logarithmic law}$$

$$x = \sqrt{k_p t} \qquad \text{parabolic law}$$

where, in each equation, x is the thickness of the film, t is the time of reaction and k_c, k_p and so on 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*.

8.7.3 Oxide film microstructure

The initial step of an oxidation reaction is usually the adsorption of oxygen onto the metal surface. Initially the adsorbate will consist of oxygen molecules that are weakly bound to the surface (Figure 8.32a). 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 8.32b). This stage is called *chemical adsorption* or *chemisorption*.

The oxidation begins by the diffusion of oxygen atoms into surface layers of the metal to form a dilute solid solution (Figure 8.32c). Nucleation of chemically recognisable oxides then occurs. At high temperatures and low partial pressures of oxygen, this takes place at random (Figure 8.32d). 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 proceeds, the islands grow together to give a surface layer consisting of randomly oriented grains of oxide (Figure 8.32e). Thereafter, further oxidation requires transport of material across the film.

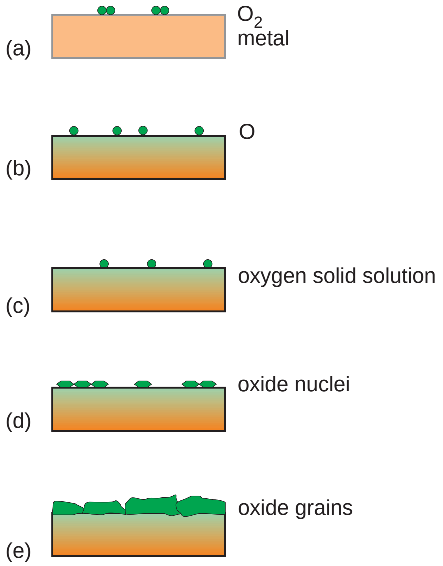


Figure 8.32 Oxide formation on a metal surface, schematic: (a) physical absorption of oxygen molecules; (b) dissociation into separated oxygen atoms (O) strongly bound to the surface; (c) penetration of some oxygen atoms into the metal to form a solid solution as more oxygen arrives at the surface; (d) saturation of the surface and subsurface with oxygen, leading to formation of oxide nuclei on the surface; (e) surface film made up of oxide grains.

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 general oxidation:

$$m\text{M} + \frac{n}{2}\text{O}_2 \rightarrow \text{M}_m\text{O}_n$$

$$X_{\text{PB}} = \frac{m_{\text{o}}\rho_m}{m\, m_m \rho_o}$$

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 1, 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.

8.7.4 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 8.33a). 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 remains balanced – a process called *ambipolar*

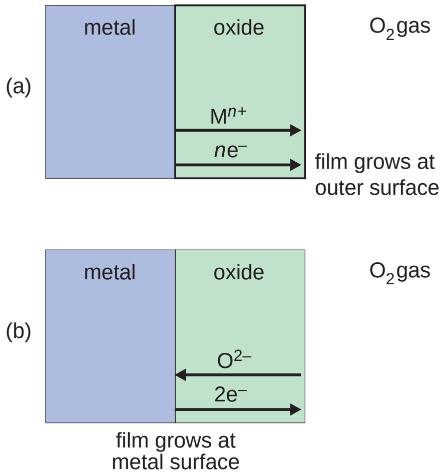


Figure 8.33 Growth of an oxide film on a metal surface: (a) diffusion of metal ions and electrons leads to growth at the outer (oxide/gas) side of the oxide film; (b) counter-diffusion of electrons and oxide ions leads to growth at the inner (metal/oxide) side of the oxide film.

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.

A similar mechanism envisages that oxygen ions diffuse across the film from the outer surface towards the metal (Figure 8.33b). 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 the following section.

The oxidation of copper metal in a low partial pressure of oxygen produces Cu_2O, by a mechanism involving diffusion of Cu^+ cations and electrons. The reaction is described by the chemical equation:

$$2Cu + {}^1\!/\!{}_2O_2 \rightarrow Cu_2O$$

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. The initial reaction of aluminium is similar and gives a thin, continuous and well-attached film of oxide. However, unlike copper, 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.

In these reactions the rate of growth of the film is controlled by the relatively slow diffusion of the ions, compared with electron transport. The reaction rate can be derived from Fick's first law of diffusion (Section 7.3):

$$J = -D\frac{dc}{dx}$$

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, $d(\Delta x)/d(\Delta t)$, will be proportional to the net flow of ions J:

$$\frac{d(\Delta x)}{d(\Delta t)} \propto J$$

The concentration gradient (dc/dx) 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:

$$\frac{d(\Delta x)}{d(\Delta t)} = \frac{k\,\Delta c}{x}$$

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:

$$x^2 = k_p\, t$$

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:

$$k \propto D_A$$

where D_A is the diffusion coefficient of the ionic species.

8.7.5 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, 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.) 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.

8.8 Solid-state reactions

Reactions between two solids are analogous to the oxidation of a metal, because the product of the reaction separates the two reactants. Further reaction is dependent upon the transport of material across this barrier. As with oxidation, cracking, porosity and volume mismatch can all help in this. In this section, the case where 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. To illustrate the problems encountered a typical solid-state reaction, the formation of oxide spinels is described.

8.8.1 Spinel formation

Spinel is a mineral with a composition $MgAl_2O_4$. The spinel formation reaction can be represented by the chemical equation:

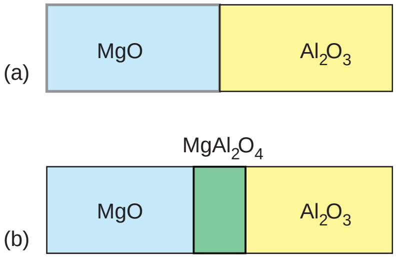


Figure 8.34 Growth of a layer of spinel ($MgAl_2O_4$) between crystals of magnesium oxide (MgO) and alumina (Al_2O_3).

$$MgO + Al_2O_3 \rightarrow MgAl_2O_4$$

Suppose that a crystal of Al_2O_3 is placed in close contact with a crystal of MgO to form a *reaction couple* (Figure 8.34a). Initial reaction will result in the separation of the MgO and Al_2O_3 by a layer of $MgAl_2O_4$ (Figure 8.34b). 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 of O^{2-} anions and Mg^{2+} cations (Figure 8.35a). 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 antiparallel diffusion of Al^{3+} cations might occur (Figure 8.35b). Because of the difference in the ionic charges, the diffusion of two Al^{3+} cations needs 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 8.35c). 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.

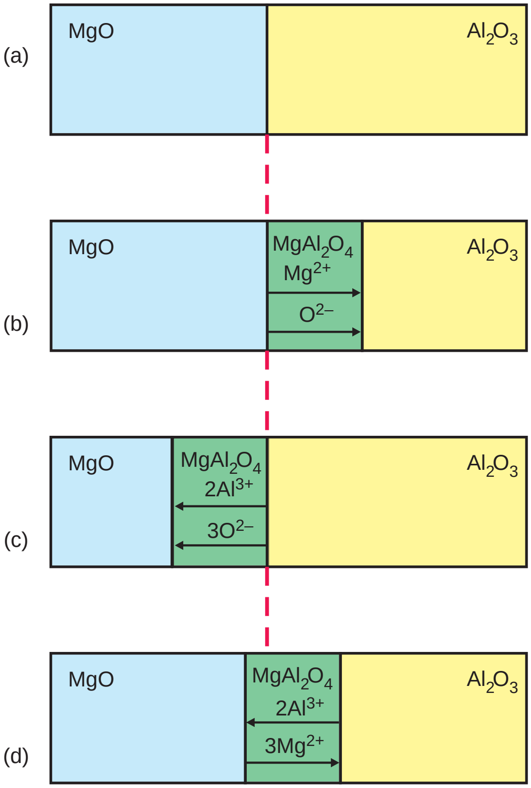


Figure 8.35 Mechanisms of spinel formation: (a) initial state; (b) diffusion of equal numbers of Mg^{2+} and O^{2-} ions; (c) diffusion of $2n$ Al^{3+} ions and $3n$ O^{2-} ions; (d) counter-diffusion of $2n$ Al^{3+} ions and $3n$ Mg^{2+} ions.

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 8.36). 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.

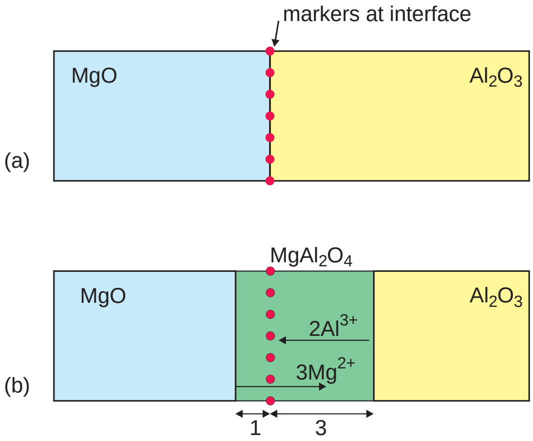


Figure 8.36 Markers used to determine the mechanism of spinel formation from magnesium oxide and alumina: (a) inert markers at the interface between MgO and Al_2O_3 crystals; (b) after the reaction the markers will appear to be within the $MgAl_2O_4$ layer when a cation counter-diffusion is in operation.

8.8.2 The kinetics of spinel formation

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 (Section 8.7.4):

$$x^2 = 2k_p t$$

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.

Further reading

Sintering:

Remmey, G. Bickley (1994) *Firing Ceramics*. World Scientific, Singapore.

Germain, R.M. (1996) *Sintering Theory and Practice*. Wiley-Interscience, New York.

Kang, S.-J.L. (2005) *Sintering, Densification, Grain Growth and Microstructure*. Elsevier Butterworth-Heinemann, Oxford.

For information on selective laser sintering, see:

Duan, B. and Wang, M. (2011) *Mater. Res. Bull.*, **36**: 998–1005; and references therein.

Computer simulations of sphere to sphere and sphere to plate sintering are to be found at www.roentzsch.org.

Transformations and microstructures:

Anderson, J.C., Leaver, K.D., Rawlings, R.D. and Alexander, J.M. (1998) *Materials Science*, 4th edn. Stanley Thornes, Cheltenham.

Callister, W.D. (2000) *Materials Science and Engineering: An Introduction*, 5th edn. John Wiley & Sons, Ltd., New York.

Dove, M.T. (1997) Theory of displacive transitions in minerals. *American Mineralogist*, **82**: 213–44.

Putnis, A. (1992) *Introduction to Mineral Sciences*. Cambridge University Press, Cambridge.

Shape-memory alloys:

Schetky, L McD. (1979) *Scientific American*, **241** (November): 74–82.

Materials Research Society Bulletin, **27** (February 2002): 91–127 (several authors).

Osuka, K. and Wayman, C.M. (1998) *Shape Memory Materials*. Cambridge University Press, Cambridge.

High-temperature oxidation and solid-state reactions:

Khanna, K.S. (2002) *High Temperature Oxidation and Corrosion*. ASM International, OH, Blackwell, Oxford.

Young, D. (2008) *High Temperature Oxidation and Corrosion of Metals*. Elsevier, Oxford.

Problems and exercises

Quick quiz

1 The transformation of a compacted powder into a solid by heating is called:
(a) Annealing.
(b) Tempering.
(c) Sintering.

2 Liquid phase sintering involves using:
(a) Only liquids for the reaction.
(b) Solids mixed with liquids for the reaction.
(c) An additive which melts during the reaction.

3 The driving force for sintering is:
(a) Reduction in surface area.
(b) Reduction in total volume.
(c) Reduction in Gibbs energy of reaction.

4 During slow cooling of a Ni–Cu alloy, the composition of the crystallites:
(a) Is constant.
(b) Varies continuously.
(c) Has the composition of the liquid.

5 When a 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 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.

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.

8 When austenite is cooled slowly through the eutectoid point, the material that forms is called:

(a) Cementite.

(b) Ferrite.

(c) Pearlite.

9 The microstructure of pearlite is composed of:

(a) Eutectoid ferrite plus austenite.

(b) Eutectoid ferrite plus eutectoid cementite.

(c) Eutectoid cementite plus austenite.

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.

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.

12 Martensite is produced from austenite by:

(a) Quenching the austenite.

(b) Annealing the austenite.

(c) Tempering the austenite.

13 A martensitic transformation is described as a:

(a) Reconstructive transformation.

(b) Diffusionless transformation.

(c) Equilibrium transformation.

14 Shape memory alloys utilise:

(a) Martensitic transformations.

(b) Eutectoid transformations.

(c) Reconstructive transformations.

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.

16 The shape-memory alloy Nitinol has an approximate formula:

(a) $NiTi_3$.

(b) NiTi.

(c) Ni_3Ti.

17 Direct corrosion of a metal requires the presence of:

(a) A gas plus water.

(b) A dry gas.

(c) Air.

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.

19 The Pilling-Bedworth ratio is used to predict:

(a) The likelihood of corrosion of a metal.

(b) The reactivity of a metal with a gas.

(c) The rate of corrosion of a metal.

20 The formation of spinel, $MgAl_2O_4$, involves:

(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.

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.

Calculations and questions

8.1 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.

8.2 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.

8.3 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.

8.4 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.

8.5 Using the data in question 8.4, what is the equilibrium size of a pore that will neither shrink nor expand?

8.6 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.

8.7 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.

8.8 The ruthenium–rhenium system forms a complete solid solution (see Figure 4.22 for the phase diagram). 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.

8.9 The Al_2O_3–Cr_2O_3 system forms a complete solid solution (see Figure 4.21 for the phase diagram). 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.

8.10 With respect to the phase diagram of the copper–silver system, Figure 8.37: (a) Sketch and

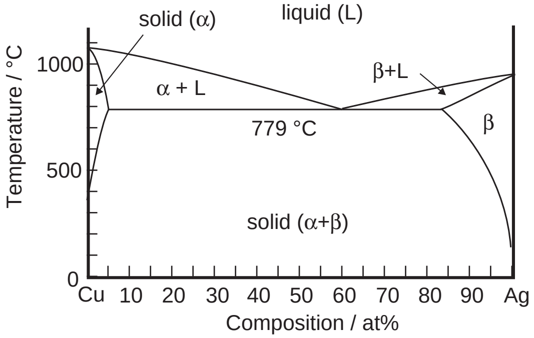


Figure 8.37 Phase diagram of the Cu–Ag system.

label the microstructure of a solid containing 2.5 at.% Ag (about halfway 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.

8.11 With respect to the iron–carbon phase diagram, Figure 4.15: (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?

8.12 With respect to the iron–carbon phase diagram, Figure 4.15: (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.

8.13 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
Na/Na_2O	966	2270

8.14 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}

8.15 When copper is oxidised under low partial pressures of oxygen gas, it forms Cu_2O via a parabolic rate law. The rate constant is $5.38 \times 10^{-10}\,m^2\,s^{-1}$ 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?

8.16 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/h
1.0	20
1.4	40
1.8	60
2.0	80
2.3	100

9

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.)

9.1 Galvanic cells

9.1.1 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 9.1). 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 as 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*. *Anions* (negative ions) travel towards the anode through the electrolyte. *Cations* (positive ions) travel towards the cathode through the electrolyte. The electrons travel via an external circuit from one electrode to the other and are used to do useful work. The driving force for this is the energy of the cell reaction, which stops when the external circuit is broken. Batteries

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

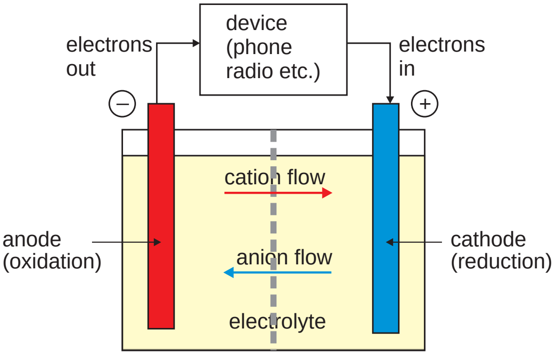


Figure 9.1 The components of a galvanic cell.

are spent, or need recharging, when one or both of the components have been used up.

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 sulphate, and a copper rod in a solution of copper sulphate. To complete the circuit, 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 9.2). A current of electrons flows from the zinc anode to the copper cathode when the external connection is completed with a metallic conductor.

The reactions taking place at each electrode, defined by *half-reactions*, are:

$$\textit{anode half-reaction (oxidation)}:\quad \mathrm{Zn(s)} \rightarrow \mathrm{Zn^{2+}(aq)} + 2\mathrm{e^-}$$

$$\textit{cathode half-reaction (reduction)}:\quad \mathrm{Cu^{2+}(aq)} + 2\mathrm{e^-} \rightarrow \mathrm{Cu(s)}$$

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:

$$\textit{cell reaction (redox reaction)}:\quad \mathrm{Zn(s)} + \mathrm{Cu^{2+}(aq)} \rightarrow \mathrm{Zn^{2+}(aq)} + \mathrm{Cu(s)}$$

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

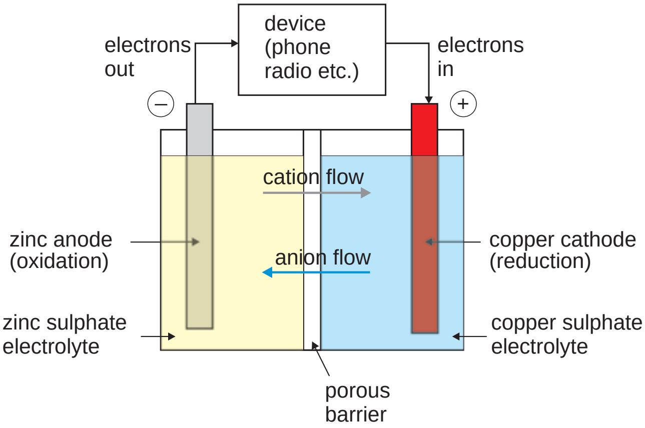

electrons out
device (phone radio etc.)
electrons in
−
+
cation flow
anion flow
zinc anode (oxidation)
zinc sulphate electrolyte
copper cathode (reduction)
copper sulphate electrolyte
porous barrier

Figure 9.2 Schematic diagram of a Daniell cell.

rod gains weight and a layer of new copper metal forms on the surface. Electroplating is occurring.

9.1.2 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:

$$\begin{aligned} E_{\text{cell}} &= E_{\text{c}}(\text{cathode; reduction half-reaction}) \\ &\quad + E_{\text{a}}(\text{anode; oxidation half-reaction}) \end{aligned}$$

In general, only the reduction half-reaction potentials are listed in tables (Table 9.1). 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^0, are recorded:

$$E^0_{\text{cell}} = E^0_{\text{c}} + E^0_{\text{a}}$$

For cells involving ionic solutions, the standard state is a solution of *1 molar concentration*, and for cells involving gases these are at *1 atmosphere pressure* at 298.15 K (25°C).[1]

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 atmosphere. The standard electrode potential of this half-reaction is *defined* as zero:

$$2H^+(aq) + 2e^- \rightarrow H_2(g) \qquad E^0 = 0\,\text{V}$$

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

Table 9.1 Standard reduction potentials at 25°C

Half-reaction	E^0/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
$Al^{3+} + 3e^- \rightarrow Al$	−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.

[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. (1.01325×10^5 Pa).

found to be the anode and the hydrogen electrode is the cathode.

$$\begin{aligned}&\textit{anode half-reaction}:\\ &\quad \mathrm{Zn(s)} \rightarrow \mathrm{Zn^{2+}(aq)} + 2\mathrm{e^-} \quad E^0 = +0.76\ \mathrm{V}\end{aligned}$$

(Because zinc is the anode, the value of E^0 is the negative of that given in Table 9.1.)

$$\begin{aligned}&\textit{cathode half-reaction}:\\ &\quad 2\mathrm{H^+(aq)} + 2\mathrm{e^-} \rightarrow \mathrm{H_2(g)} \quad E^0 = 0\ \mathrm{V}\\ &\textit{cell reaction}:\\ &\quad \mathrm{Zn(s)} + 2\mathrm{H^+(aq)} \rightarrow \mathrm{Zn^{2+}(aq)} + \mathrm{H_2(g)}\end{aligned}$$

$$E^0_{\mathrm{cell}} = +0\ \mathrm{V} + (+0.76\ \mathrm{V}) = 0.76\ \mathrm{V}$$

The Zn is oxidised to $\mathrm{Zn^{2+}}$ and the $\mathrm{H^+(aq)}$ is reduced to $\mathrm{H_2(g)}$.

In a cell using copper, the hydrogen electrode is found to be the anode and the copper electrode the cathode:

$$\begin{aligned}&\textit{anode half-reaction}:\\ &\quad \mathrm{H_2(g)} \rightarrow 2\mathrm{H^+(aq)} + 2\mathrm{e^-} \quad E^0 = 0\ \mathrm{V}\\ &\textit{cathode half-reaction}:\\ &\quad \mathrm{Cu^{2+}(aq)} + 2\mathrm{e^-} \rightarrow \mathrm{Cu(s)} \quad E^0 = +0.34\ \mathrm{V}\\ &\textit{cell reaction}:\\ &\quad \mathrm{H_2(g)} + \mathrm{Cu^{2+}(aq)} \rightarrow 2\mathrm{H^+(aq)} + \mathrm{Cu(s)}\\ &\quad E^0_{\mathrm{cell}} = +0.34\ \mathrm{V} + 0\ \mathrm{V} = +0.34\ \mathrm{V}\end{aligned}$$

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 above, it is seen that the order of the electrodes with respect to this tendency is $\mathrm{Zn^{2+}/Zn} < \mathrm{H^+/H_2(g)} < \mathrm{Cu^{2+}/Cu}$. Thus, in a Daniell cell, Zn forms the anode and Cu the cathode:

$$\begin{aligned}&\textit{anode half-reaction}:\\ &\quad \mathrm{Zn(s)} \rightarrow \mathrm{Zn^{2+}(aq)} + 2\mathrm{e^-} \quad E^0 = +0.76\ \mathrm{V}\\ &\textit{cathode half-reaction}:\\ &\quad \mathrm{Cu^{2+}(aq)} + 2\mathrm{e^-} \rightarrow \mathrm{Cu(s)} \quad E^0 = +0.34\ \mathrm{V}\\ &\textit{cell reaction}:\\ &\quad \mathrm{Zn(s)} + \mathrm{Cu^{2+}(aq)} \rightarrow \mathrm{Cu(s)} + \mathrm{Zn^{2+}(aq)}\\ &\quad E^0_{\mathrm{cell}} = +0.34\ \mathrm{V} + (+0.76\ \mathrm{V}) = 1.10\ \mathrm{V}\end{aligned}$$

The comparison of each element to a hydrogen electrode in a standard galvanic cell allows the reduction tendency to be ranked. A table of these values 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* (Table 9.1). 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, $\mathrm{F_2}$, 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 $\mathrm{Li^+}$ and $\mathrm{F^-}$ 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:

$$\begin{aligned}&\mathrm{Zn(s)|Zn^{2+}(aq)\ ||\ H^+(aq)|H_2(g)|Pt} \quad && E^0 = 0.76\ \mathrm{V}\\ &\mathrm{Pt|H_2(g)|H^+(aq)\ ||\ Cu^{2+}(aq)|Cu(s)} \quad && E^0 = 0.34\ \mathrm{V}\\ &\mathrm{Zn(s)|Zn^{2+}(aq)\ ||\ Cu^{2+}(aq)|Cu(s)} \quad && E^0 = 1.10\ \mathrm{V}\end{aligned}$$

9.1.3 Cell potential and Gibbs energy

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:

$$\Delta G_{\mathrm{r}} = -nE_{\mathrm{cell}}F$$

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:

$$\mathrm{Zn(s)} + \mathrm{Cu^{2+}(aq)} \rightarrow \mathrm{Cu(s)} + \mathrm{Zn^{2+}(aq)}$$

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:

$$\begin{aligned}\Delta G_r &= -2E_{\text{cell}}F \\ &= -1.93 \times 10^5\, E_{\text{cell}}\end{aligned}$$

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:

$$\Delta G_r^0 = -nE^0F$$

9.1.4 Concentration dependence

The potential generated by a cell is dependent upon the concentration of the components present. The relationship is given by the *Nernst equation*:

$$E_{\text{cell}} = E^0 - \left(\frac{RT}{nF}\right)\ln Q \qquad (9.1)$$

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 below. Inserting values for the constants, the Nernst equation can be written:

$$E_{\text{cell}} = E^0 - \left(\frac{0.02569}{n}\right)\ln Q$$

$$\text{or} \quad E_{\text{cell}} = E^0 - \left(\frac{0.05916}{n}\right)\log Q$$

The reaction quotient, Q, of a reaction:

$$a\mathrm{A} + b\mathrm{B} \leftrightarrow x\mathrm{X} + y\mathrm{Y}$$

is given by:

$$Q = Q_c = \frac{[\mathrm{X}]^x[\mathrm{Y}]^y}{[\mathrm{A}]^a[\mathrm{B}]^b}$$

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:

$$Q = Q_p = \frac{p_{\mathrm{X}}^x p_{\mathrm{Y}}^y}{p_{\mathrm{A}}^a p_{\mathrm{B}}^b}$$

Pure liquids or *solids*, or *water in solutions*, that appear in the cell reaction equation are *not* entered into the equations for Q. (Note that, for most accurate work, the quantity of importance is the *activity* rather than concentration. Activity and concentration are equal in dilute solutions.)

9.2 Chemical analysis using galvanic cells

9.2.1 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 $Hg_2Cl_2/Hg, Cl^-$.

[2] The pH of a liquid is defined as pH = $-\log\,[H_3O^+] = -\log\,[H^+]$, where $[H_3O^+]$ or $[H^+]$ are equivalent to the concentration of hydrogen ions present.

The cell is :

$$\mathrm{Pt}|\mathrm{H_2(g)}|\mathrm{H^+(aq)}||\mathrm{Cl^-(aq)}|\mathrm{Hg_2Cl_2}|\mathrm{Hg(l)}$$

anode half-reaction :

$$\mathrm{H_2(g)} \rightarrow \mathrm{2H^+(aq)} + 2\mathrm{e^-} \qquad E^0 = 0$$

cathode half-reaction :

$$\mathrm{Hg_2Cl_2(s)} + 2\mathrm{e^-} \rightarrow \mathrm{2Hg(l)} + \mathrm{2Cl^-(aq)}$$
$$E^0 = +0.27\ \mathrm{V}$$

cell reaction :

$$\mathrm{Hg_2Cl_2(s)} + \mathrm{H_2(g)} \rightarrow \mathrm{2H^+(aq)} + \mathrm{2Cl^-(aq)} + \mathrm{2Hg(l)}$$

The reaction quotient for the cell reaction is:

$$Q = \frac{[\mathrm{H^+}]^2[\mathrm{Cl^-}]^2}{p_{\mathrm{H2}}}$$

The Nernst equation is then:

$$E_{\mathrm{cell}} = E^0 - \left(\frac{0.05916}{2}\right)\log\left(\frac{[\mathrm{H^+}]^2[\mathrm{Cl^-}]^2}{p_{\mathrm{H2}}}\right)$$

To simplify matters, take the hydrogen pressure and $[\mathrm{Cl^-}]$ as standard, in which case:

$$E_{\mathrm{cell}} = E^0 - 0.05916 \log [\mathrm{H^+}]$$

i.e. $$E_{\mathrm{cell}} = E^0 + 0.05916\,\mathrm{pH}$$

In cases where the hydrogen pressure and $[\mathrm{Cl^-}]$ are not standard, they are incorporated into a *cell constant* E':

$$E_{\mathrm{cell}} = E' + 0.05916\,\mathrm{pH}$$

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 $\mathrm{H^+}$ ions (Figure 9.3). 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.

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 molar solution of potassium chloride (KCl), saturated with silver chloride (AgCl). The half-reaction is:

$$\mathrm{AgCl(s)} + \mathrm{e^-} \rightarrow \mathrm{Ag(s)} + \mathrm{Cl^-}$$
$$E^0 = 0.2046\ \mathrm{V}\ \text{at } 25^\circ\mathrm{C}$$

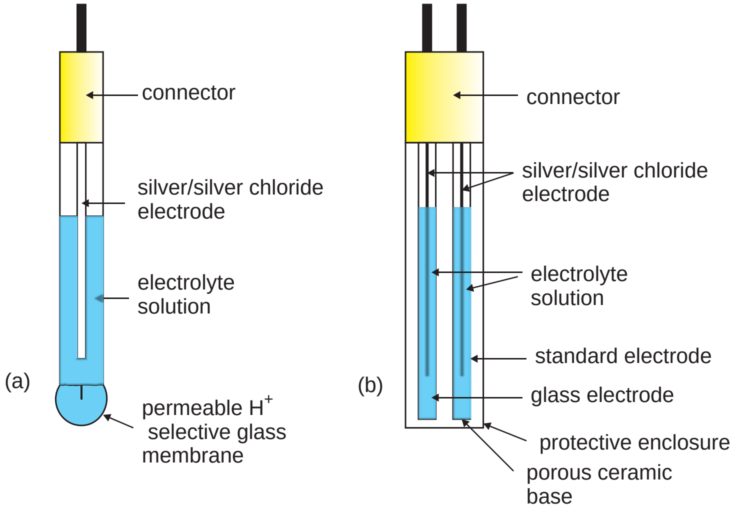


Figure 9.3 (a) A glass electrode for the measurement of pH. (b) Experimental arrangement of glass electrode and standard electrode in a single cell.

In practice the hydrogen ion selective electrode and the standard electrode are packaged together in a small plastic cylinder, which is easily transported (Figure 9.3). Measurement is made by dipping the electrode into the solution to be checked, and the voltage is transformed into pH electronically.

9.2.2 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 9.4). The potential developed by such a combination is of the general form:

$$E = E' + \left(\frac{2.303\,RT}{nF}\right)\log[C]$$

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 50–60 mV for monovalent ions and 25–30 mV for divalent ions.

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^- substructure, 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

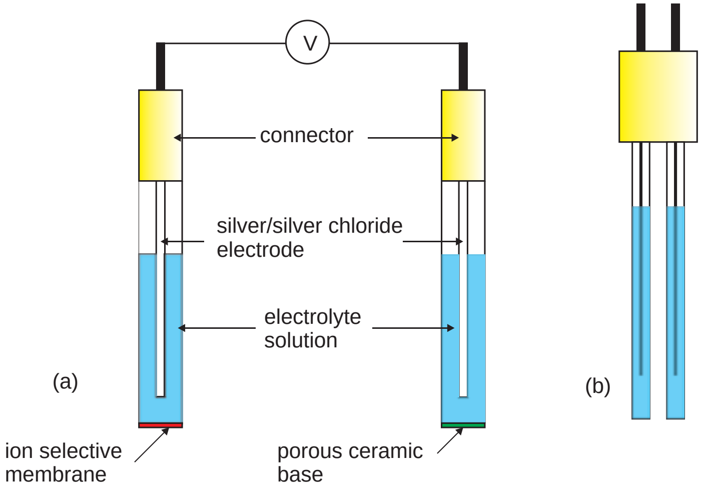


Figure 9.4 (a) Schematic arrangement of an ion selective electrode for the measurement of ion concentrations in solution. (b) Experimental arrangement, where the electrodes are attached to a connecting unit.

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.

9.2.3 Oxygen sensors

The Nernst equation indicates that a cell potential will develop even when the electrode materials are the same, provided that there 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 (Section 3.4.5, Figure 3.12). Doping leads to a high concentration of oxygen vacancies, which means that oxygen ions can diffuse very rapidly through the ceramic. 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:

$$\text{Pt, } O_2(p'_{O2}) \,\|\, \text{stab. zirconia} \,\|\, O_2(p''_{O2}),\ \text{Pt}$$

The reactions taking place are:

$$\begin{aligned}
&\textit{anode reaction}: && 2O^{2-} \rightarrow O_2(g)(p'_{O2}) + 4e^- \\
&\textit{cathode reaction}: && 4e^- + O_2(g)(p''_{O2}) \rightarrow 2O^{2-} \\
&\textit{overall cell reaction}: && O_2(g)(p''_{O2}) \rightarrow O_2(g)(p'_{O2})
\end{aligned}$$

The cell voltage is related to the oxygen pressures by the Nernst equation, equation (9.1):

$$E_{\text{cell}} = E^0 - \left(\frac{RT}{nF}\right)\ln Q$$

In this case, the number of electrons transferred, n, is 4 and the appropriate reaction quotient is:

$$Q = \frac{(p'_{O2})}{(p''_{O2})}$$

Noting that E^0 for this cell reaction is zero:

$$E = -\left(\frac{RT}{4F}\right)\ln\left[\frac{(p'_{O2})}{(p''_{O2})}\right]$$

$$E = -\left(\frac{RT}{4F}\right)\ln\left[\frac{\text{anode pressure (low)}}{\text{cathode pressure (high)}}\right]$$

Solving this equation for the oxygen partial pressure gives:

$$(p'_{O2}) = (p''_{O2})\exp\left(\frac{-4E}{RT}\right)$$

The high-pressure p''_{O2} is taken as a reference pressure so that the unknown pressure p'_{O2} is readily determined. These equations are often seen in the form:

$$E = +\left(\frac{RT}{4F}\right)\ln\left[\frac{(p''_{O2})}{(p'_{O2})}\right]$$

$$E = +\left(\frac{RT}{4F}\right)\ln\left[\frac{\text{cathode pressure (high)}}{\text{anode pressure (low)}}\right]$$

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:

$$E = -\left(\frac{RT}{4F}\right)\ln\left[\frac{[O_2(\text{solution})]}{p_{O2}(\text{reference})}\right]$$

where [O_2 (solution)] is the concentration of the oxygen molecules, which is assumed to be lower than the reference pressure. If p_{O2} is taken as 1 atmosphere the equation becomes:

$$E = -\left(\frac{RT}{4F}\right)\ln\,[O_2(\text{solution})]$$

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:

$$E = -\left(\frac{RT}{2F}\right)\ln\left[\mathrm{O}\,(\text{solution})\right]$$

The design of the sensor depends upon its ultimate use. For high-temperature applications a stabilised zirconia tube coated inside and out with porous platinum forms a suitable design. The tube can be used directly as an oxygen meter if p''_{O2} is a standard pressure, such as 1 atmosphere of oxygen or else the pressure of oxygen in air, which is approximately 0.21 atmosphere. Such a system is utilised to monitor the exhaust oxygen content for a car engine 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 air–fuel mix automatically to optimise engine efficiency.

9.3 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 20th century was inherently due to inadequate batteries. Even at present, hybrid electric vehicles are hampered by battery difficulties.

Batteries are simple in concept. It is merely necessary to combine two couples such as those in Table 9.1 and link them via a suitable electrolyte. 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 the table. The battery should ideally have a high specific energy (energy/weight) and a high energy density (energy/volume).

The 19th and early 20th centuries 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-19th century, are *still* widely available – the principal recent cell in widespread use is the lithium-ion (Li-ion) cell, employed in power tools and hybrid electric vehicles.

9.3.1 '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 $\mathrm{Zn/MnO_2}$. The current collector is a graphite rod buried in the positive cathode, a mixture of $\mathrm{MnO_2}$ and carbon. This is kept moist by the electrolyte, aqueous ammonium chloride. The negative anode is the container itself, made from zinc (Figure 9.5).

As with all commercial batteries, the real cell reaction is complex, and the reactions are approximated by:

$$\textit{anode reaction}: \mathrm{Zn(s)} \rightarrow \mathrm{Zn^{2+}(aq)} + 2\mathrm{e^-}$$

$$\textit{cathode reaction}: \mathrm{MnO_2(s)} + \mathrm{H_2O(l)} + \mathrm{e^-} \rightarrow \mathrm{MnO(OH)(s)} + \mathrm{OH^-(aq)}$$

$$\textit{cell reaction}: \quad \mathrm{Zn(s)} + 2\mathrm{MnO_2(s)} + 2\mathrm{H_2O(l)} \rightarrow 2\mathrm{MnO(OH)(s)} + 2\mathrm{OH^-(aq)}$$

The main problem encountered with this cell is the build-up of $\mathrm{Zn^{2+}}$ and $\mathrm{OH^-}$ 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.

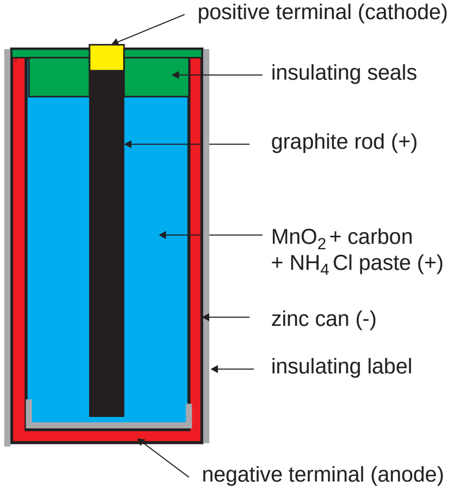


Figure 9.5 Section through a carbon–zinc battery, schematic.

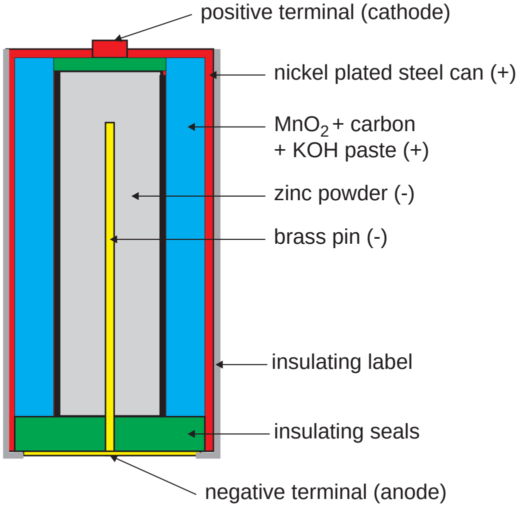


Figure 9.6 Section through an alkaline battery, schematic.

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 9.6). 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.

9.3.2 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 9.1). 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. 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 9.7).

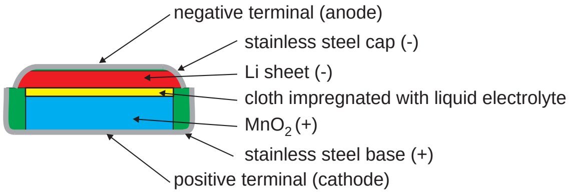

negative terminal (anode)
stainless steel cap (-)
Li sheet (-)
cloth impregnated with liquid electrolyte
MnO_2 (+)
stainless steel base (+)
positive terminal (cathode)

Figure 9.7 Section through a lithium coin cell primary battery, schematic.

The cell reactions are poorly understood, but can be written schematically as:

$$\textit{anode reaction}: \quad Li(s) \rightarrow Li^+(aq) + e^-$$

$$\textit{cathode reaction}: \quad MnO_2(s) + e^- \rightarrow MnO_2^-(s)$$

$$\textit{cell reaction}: \quad Li(s) + MnO_2(s) \rightarrow LiMnO_2(s)$$

$$E^0 = 3.2\ \mathrm{V}$$

9.3.3 The lead–acid secondary battery

Plante 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 is dilute sulphuric acid (Figure 9.8). 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 :

$$PbO_2(s) + Pb(s) + 2H^+(aq) + HSO_4^-(aq) \rightarrow 2PbSO_4(s) + 2H_2O$$

The cell potential is about 2 V. A car battery consists of (usually) 6 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.

9.3.4 Lithium-ion secondary batteries

The advantages of lithium primary cells extend to secondary cells. In particular, the high power available and the lightness make them ideal for portable

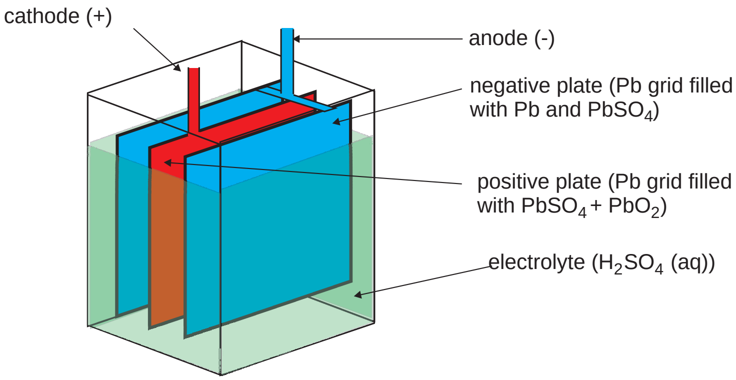

cathode (+)
anode (-)
negative plate (Pb grid filled with Pb and $PbSO_4$)
positive plate (Pb grid filled with $PbSO_4$ + PbO_2)
electrolyte (H_2SO_4 (aq))

Figure 9.8 A single cell in a lead–acid battery, schematic.

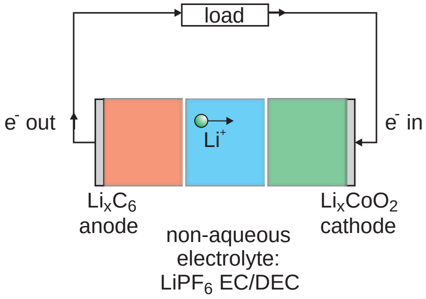


Figure 9.9 A lithium-ion cell in discharge operation, schematic.

electronic devices. The first successful lithium-ion rechargeable battery was introduced by Sony in 1991, and is often called the *Sony cell* (Figure 9.9). The difficulties of working with lithium metal are overcome by using non-stoichiometric intercalation compounds (Chapter 3). The electrolyte is, as with the lithium primary cells, a non-aqueous solution of lithium salts in a polar organic liquid.

The active component of the anode is lithium metal contained in graphite (Section 5.3.7, Figure 5.23). It has long been known that graphite can take in alkali metal atoms between the sheets of carbon hexagons to form intercalation compounds. The nominal composition of the lithium–graphite intercalation material 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 9.10a,b). The stacking, therefore, alters with the degree of lithium incorporation, limiting the performance of the electrode. The cathode material is the non-stoichiometric oxide Li_xCoO_2. This material is also an intercalation compound, in which the Li^+ ions lie between layers of composition CoO_2 (Figure 9.10c,d). In theory, the composition range of the cathode material is from $LiCoO_2$ to CoO_2, that is, 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 (ABAB) layers of oxygen atoms, while $LiCoO_2$ has cubic (ABCABC) close-packing. This change degrades the

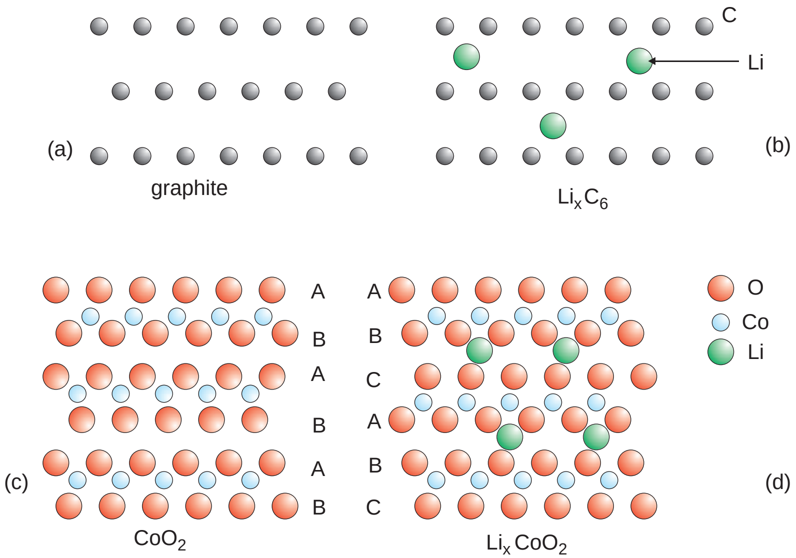


Figure 9.10 Transformations in electrode materials: (a, b) graphite changes the layer stacking when Li atoms are intercalated; (c, d) CoO_2 changes from hexagonal close-packing of the oxygen anions to cubic close-packing when Li atoms are intercalated.

oxide during use, and is the major reason why the useful composition range is limited. At present there is much research work on improving both the cathode and anode materials in these cells, and a number of alternative layer compounds are being actively tested. 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).

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.

9.3.5 Lithium–air batteries

There are a number of new battery types being explored, especially for vehicle use, that are still in the developmental stage. Among these are lithium–air batteries using an organic electrolyte, lithium–air batteries using an aqueous electrolyte, and lithium–sodium batteries.

The cell reaction of a typical lithium–air battery is:

$$2\,Li^+ + 2e^- + O_2 \rightarrow Li_2O_2$$

with lithium metal as the anode and air as the cathode (Figure 9.11). The electrolyte is similar to that in the Li-ion battery, consisting of $LiPF_6$ in polypropylene carbonate (PPC). There are a number of problems that remain before this type of cell can be commercially viable. One of these is that the cathode cannot be exposed to ordinary air, because water vapour will react with the Li_2O_2 to produce lithium hydroxide and carbon dioxide will 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. Nevertheless, these batteries are eagerly anticipated and expected to be commercially available in the future.

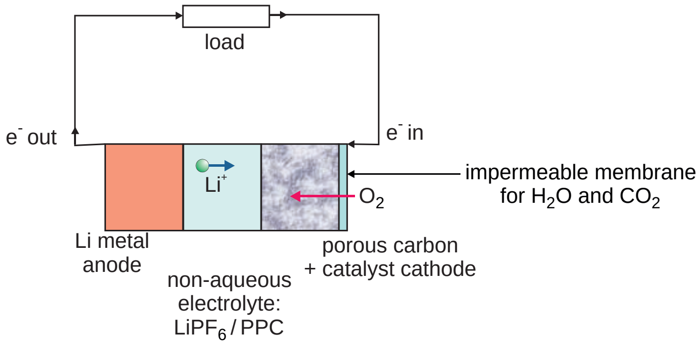

load
e- out
e- in
Li+
O2
impermeable membrane for H2O and CO2
Li metal anode
porous carbon + catalyst cathode
non-aqueous electrolyte: LiPF6 / PPC

Figure 9.11 A lithium–air battery, schematic.

9.3.6 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 reaction favoured 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 9.12), and early fuel cells, containing an alkaline solution, typically potassium hydroxide solution, as electrolyte, demonstrate the principle involved.

$$\textit{anode reaction}: \\ H_2(g) + 2OH^-(aq) \rightarrow 2H_2O(l) + 2e^- \; E^0 = -0.83\ \text{V}$$

$$\textit{cathode reaction}: \\ O_2(g) + 2H_2O(l) + 4e^- \rightarrow 4OH^-(aq) \; E^0 = 0.40\ \text{V}$$

$$\textit{cell discharge reaction}: \\ 2H_2(g) + O_2(g) \rightarrow 2H_2O(l)$$

Current research is centred upon making compact cells of high efficiency. They are described in terms of the electrolyte that is used. The principle types are *alkali* (AFC), described above, with aqueous KOH as electrolyte transferring OH^- ions; *proton exchange membrane* (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

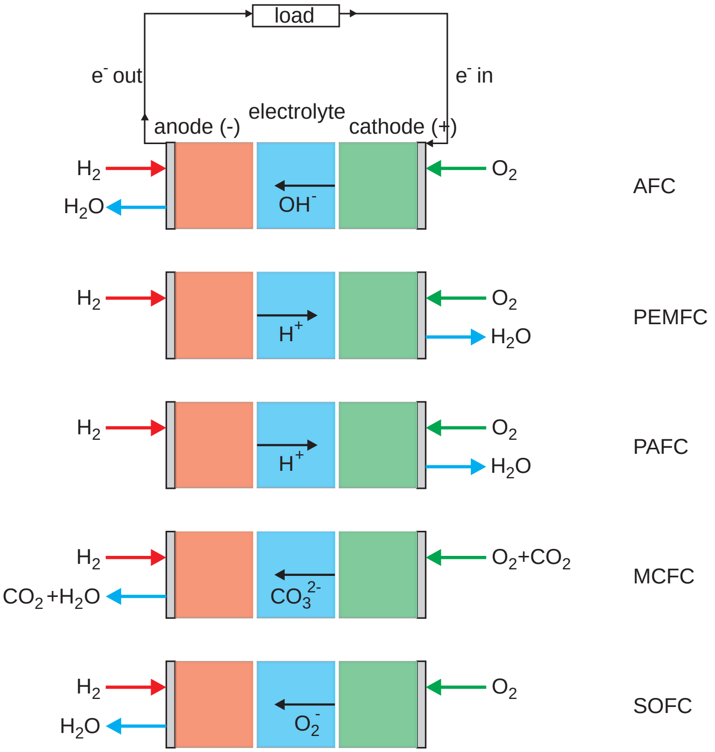


Figure 9.12 Schematic diagrams fuel cells: AFC, alkali fuel cell; PEMFC, proton exchange membrane fuel cell; PAFC, phosphoric acid fuel cell; MCFC, molten carbonate fuel cell; SOFC, solid oxide fuel cell.

H^+ ions; *molten carbonate fuel cell* (MCFC), with molten Li_2CO_3 as electrolyte in a $LiAlO_2$ matrix, transporting CO_3^{2-} ions; and solid oxide fuel cell (SOFC), with solid yttria-stabilised zirconia electrolytes that allow 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, but 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 the cell, although *direct methanol conversion fuel cells* (DMCFC) are also under investigation.

Of the cells available, solid oxide fuel cells (SOFCs) appear to be nearest to commercial use for large-scale electricity generation. A variety of designs are being explored, including planar and tubular geometries (Figure 9.13). In all cases, single cells are linked to give a fuel-cell stack by an interconnecting 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 3.4.5). 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.

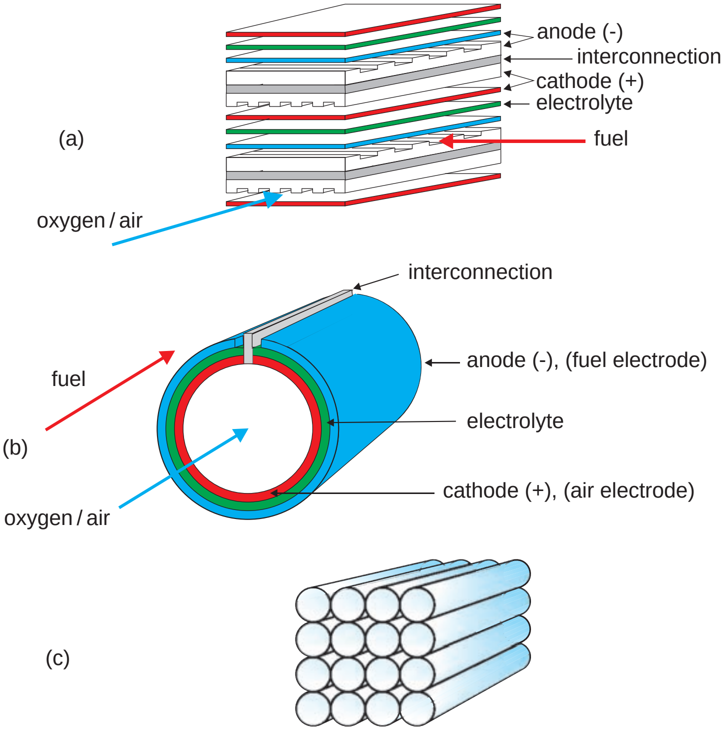


Figure 9.13 (a) An expanded view of a stack of planar-design solid oxide fuel cells. (b) Tubular design of a solid oxide fuel cell. (c) An array of tubular fuel cells.

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 :

$$H_2(g) + \tfrac{1}{2}O_2(g) \rightarrow H_2O(l)$$

A problem with solid oxide fuel cells is 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 temperatures in excess of 650°C. Intensive research is lowering this temperature continually.

9.4 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 8.7.) Aqueous corrosion involves a set of complex electrochemical reactions in which the metal reverts to a more stable condition, usually an oxide or mixture of oxides and hydroxides (Figure 9.14). In many cases the products are not crystalline and are frequently mixtures of compounds. Aside from the loss of metal, the corrosion products may be voluminous. In this case, they force overlying protective layers away from the metal, and so allow corrosion to proceed unchecked, which exacerbates the damage.

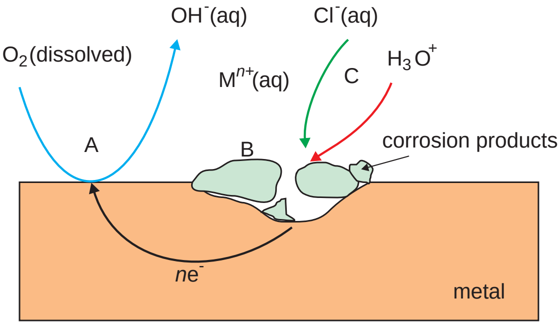


Figure 9.14 Aqueous corrosion reactions: A, electrode reactions; B, deposition reactions; C, impurity reactions.

The extent of aqueous corrosion often depends upon the presence of impurities and trace contaminants in the water present. 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.

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 is a major economic factor across the world.

9.4.1 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 in Table 9.1. Metals below H_2 in the electrochemical series, those with a negative standard reduction potential, will react in 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:

$$Zn + 2HCl \rightarrow ZnCl_2 + H_2(g)$$

However, more insight into the process is given by using the two half-reactions:

$$2H^+(aq) + 2e^- \rightarrow H_2(g)$$
$$Zn(s) \rightarrow Zn^{2+}(aq) + 2e^-$$
$$\text{i.e. } Zn(s) + 2H^+(aq) \rightarrow Zn^{2+}(aq) + H_2(g)$$

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:

$$2H_2O(l) + 2e^- \rightarrow H_2(g) + 2OH^-(aq)$$
$$E^0 = -0.83\text{ V at pH} = 14$$

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 molar 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.

$$2H_2O(l) + 2e^- \rightarrow H_2(g) + 2OH^-(aq)$$
$$E = -0.42\text{ V at pH} = 7$$

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 little tendency to be corroded by pure water. Corrosion of iron only takes place in water containing dissolved oxygen, discussed below.

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:

$$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$$
$$E^0 = +1.23\text{ V at pH} = 0$$

The presence of the $H^+(aq)$ ions indicates that the water is acidic. The concentration of $H^+(aq)$ in the standard state is one molar, and the pH will be 0, equivalent to very acidic conditions. In these conditions, 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:

$$Co^{3+}(aq) + e^- \rightarrow Co^{2+}(aq)$$
$$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$$
$$\text{i.e. } 4Co^{3+}(aq) + 2H_2O(l) \rightarrow 4Co^{2+}(aq) + O_2(g) + 4H^+(aq)$$

The reduction potential for neutral water can be calculated via the Nernst equation.

$$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l)$$
$$E = +0.81\text{ V at pH} = 7$$

There are a number of half-reactions involving nonmetals that lie above this value in the electrochemical series, and all will be reduced. In these cases, reactions 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:

$$2H_2O(l) \rightarrow 4H^+(aq) + O_2(g) + 4e^-$$
$$2Cl_2(g) + 4e^- \rightarrow 4Cl^-(aq)$$
$$2Cl_2 + 2H_2O \rightarrow 4HCl + O_2(g)$$

9.4.2 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.

9.4.2.1 Copper and iron/steel

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 :

$$2Fe(s) + O_2 + 2H_2O(l) \rightarrow 4(OH^-) + 2Fe^{2+}(aq)$$

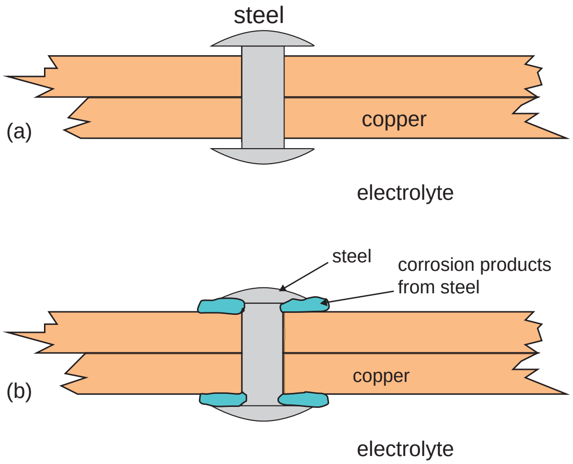


Figure 9.15 Dissimilar metal corrosion of a steel rivet in contact with copper and an electrolyte.

The Fe^{2+} is soluble and the iron will gradually dissolve (Figure 9.15). 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.

These cells have had considerable influence historically. Wooden sailing ships were attacked below the waterline 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. 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 lost.

9.4.2.2 Galvanizing

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

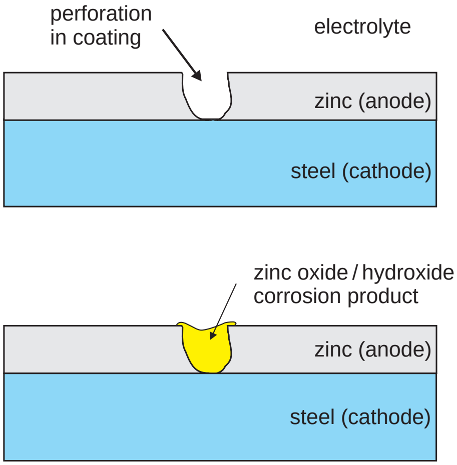


Figure 9.16 The protective coating of a hydroxy-oxide formed by zinc on steel following galvanising.

reaction (Section 8.7). 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 9.16). 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 :

$$2Zn(s) + O_2 + 2H_2O(l) \rightarrow 4(OH^-) + 2Zn^{2+}(aq)$$

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.

9.4.2.3 Tin-plate

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 to uncoated steel. This accounts for the fact that old tin cans on rubbish tips are always badly corroded.

9.4.3 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 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.

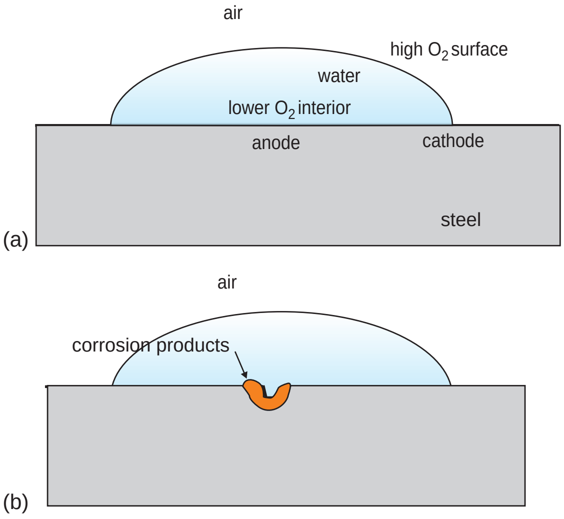


Figure 9.17 Differential aeration leading to pitting in steel: (a) initial situation, a cathode is formed at the outer oxygen-rich circumference of the water drop, and an anode at the oxygen-poor centre; (b) after corrosion.

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 9.17). The less aerated centre forms an anode, and corrosion produces a pit at the centre of the drop:

$$\textit{anode reaction}:$$
$$\mathrm{Fe(s)} \rightarrow \mathrm{Fe^{2+}} + 2\mathrm{e^-}$$
$$\textit{cathode reaction}:$$
$$\mathrm{O_2(g)} + \mathrm{H_2O(e)} + 4\mathrm{e^-} \rightarrow 4(\mathrm{OH^-})$$
$$\textit{cell reaction}:$$
$$2\mathrm{Fe(s)} + \mathrm{O_2} + 2\mathrm{H_2O(l)} \rightarrow 4(\mathrm{OH^-}) + 2\mathrm{Fe^{2+}(aq)}$$

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 9.18). This problem is often enhanced when the corrosion product has a high volume. The resulting stress may lever the rivet head off.

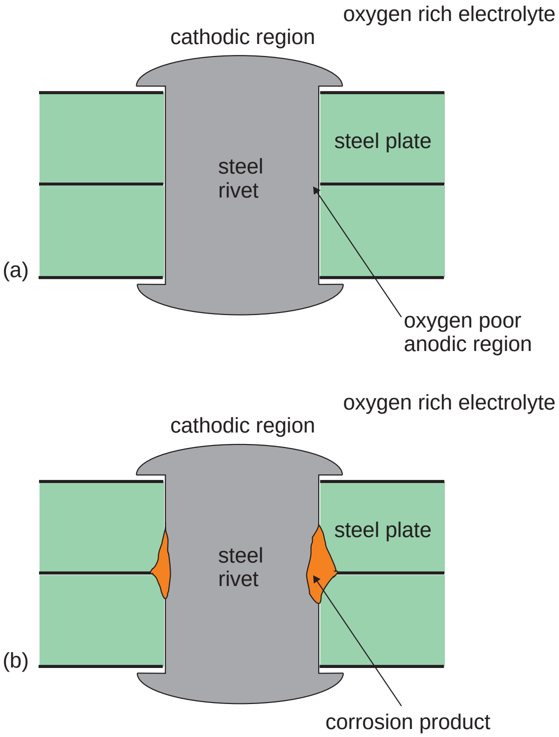


Figure 9.18 Corrosion in a crevice due to differential aeration: (a) initial situation, a cathode is formed at the outer oxygen-rich region of the electrolyte and an anode in the oxygen-poor crevice; (b) after corrosion.

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 to 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.

9.5 Electrolysis

While a galvanic cell uses a spontaneous chemical reaction to produce an electric current, an

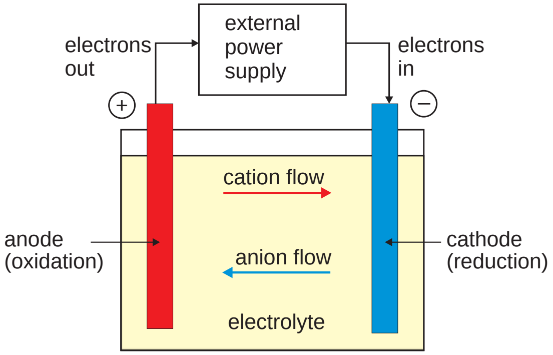


Figure 9.19 The components of an electrolytic cell.

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.

9.5.1 Electrolytic cells

Electrolytic cells do not need the electrodes in separate compartments and so are simpler in construction than galvanic cells (Figure 9.19). 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 +.

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.

9.5.2 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 above, in practice the production of a high-quality film is critically dependent upon large numbers of factors, especially the cleanliness of the surface to be plated. In addition, commercial plating solutions contain organic additives to enhance film adherence.

The process can be illustrated by the *schematic* description of nickel plating (Figure 9.20). 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:

$$Ni^{2+}(aq) + 2e^- \rightarrow Ni(s) \qquad E^0 = +0.23\ V$$

Two moles of electrons must be supplied to deposit 1 mole of nickel. At the anode, the chloride ions are discharged, to form chlorine gas:

$$2Cl^-(aq) \rightarrow Cl_2(g) + 2e^- \qquad E^0 = +1.36\ V$$

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:

$$Ni(s)\ \text{anode} \rightarrow Ni(s)\ \text{cathode}$$

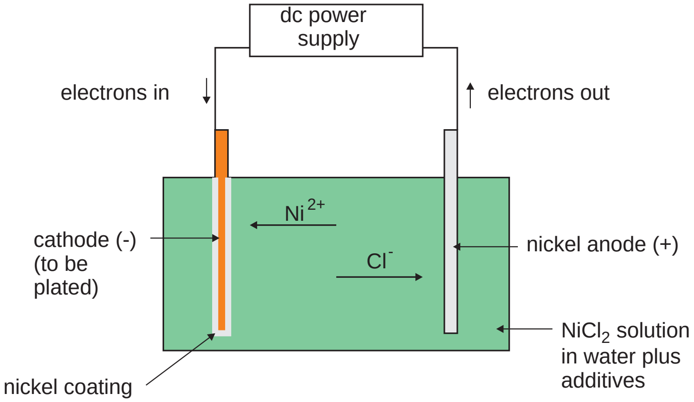


Figure 9.20 Nickel electroplating: the nickel anode is dissolved and transported to the object to be plated, which is the cathode, under the driving force of the external power supply.

It is this same process that causes the corrosion of anodes in all batteries. As a first approximation the balanced reactions are:

anode reaction :
$$\mathrm{Ni(s) + 2Cl^-(aq) \rightarrow Ni^{2+}(aq) + Cl_2(g) + 4e^-}$$
cathode reaction :
$$\mathrm{2Ni^{2+}(aq) + 4e^- \rightarrow 2Ni(s)}$$
plating reaction :
$$\mathrm{2Cl^-(aq) + Ni^{2+}(aq) \rightarrow Ni(s) + Cl_2(g)}$$

9.5.3 The amount of product produced during electrolysis

The *chemical nature* of the products 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 *Faraday's 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:

$$Q = It$$

where I is the current and t the time. One mole of electrons has a charge:

$$(1.6022 \times 10^{-19}\ \mathrm{C}) \times (6.0222 \times 10^{23}\ \mathrm{mol^{-1}})$$
$$= 9.6485 \times 10^4\ \mathrm{C\ mol^{-1}}$$

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:

$$Q_\mathrm{m} = \frac{It}{F}$$

To obtain the mass of an element that is formed, it is necessary to multiply by the molar mass produced by 1 mole of electrons:

$$m = \left(\frac{It}{F}\right)\frac{M}{Z}$$

where m is the mass produced, M is the molar mass of the element and Z is the charge on the ion involved.

9.5.4 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. To date, none of these has worked well. 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 recent process, called the *FFC Cambridge Process*, uses a slightly different electrochemical approach. The method is named after its discoverers, Fray, Farthing and Chen, working in the University of Cambridge, England. The key to the method, and what distinguishes it from earlier electrolysis attempts, lies in the use of slightly non-stoichiometric titanium dioxide, $TiO_{2\text{-}x}$, 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, which can be written $TiO_{2-\delta}$ or $Ti_{1+\delta}O_2$, could be imagined to contain either oxygen vacancies or titanium interstitials (Section 3.4.5). However, neither of these alternatives is correct. Instead the structure collapses along planes, initially (132), then at greater degrees of reduction (121), called *crystallographic shear* (CS) planes (Section 3.5.4). The collapse is due to the removal of a complete sheet of oxygen atoms and it results in lamellae of a structure analogous to that of the lower oxide Ti_2O_3. This latter phase is a metallic conductor, and so the slightly reduced TiO_2 has planes of conducting material throughout the bulk. The reduced materials are then good electronic conductors, and even a composition as close to TiO_2 as $TiO_{1.995}$ conducts electricity well.

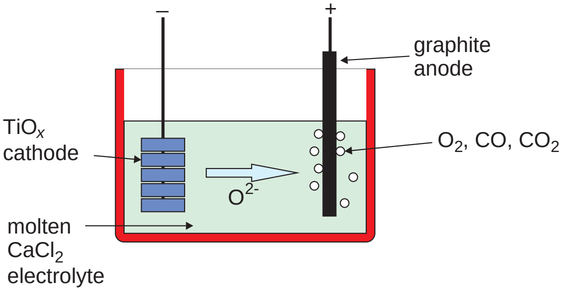


Figure 9.21 Schematic design of the FFC Cambridge cell for the production of titanium metal by the electrolysis of slightly reduced titanium dioxide, TiO_x.

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 9.21). 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.

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:

$$\begin{aligned} &\textit{anode reaction}: && xC + xO^{2-} \rightarrow xCO(g) + 2xe^- \\ &\textit{cathode reaction}: && TiO_x + 2xe^- \rightarrow xO^{2-} + Ti \\ &\textit{electrolysis reaction}: && TiO_x + xC \rightarrow Ti + xCO(g) \end{aligned}$$

In reality, carbon monoxide, carbon dioxide and oxygen are produced at the anode. These anode reactions can be approximated by:

$$\begin{aligned} aC + aO^{2-} &\rightarrow aCO(g) + 2ae^- \\ bC + 2bO^{2-} &\rightarrow bCO_2(g) + 2be^- \\ cO^{2-} &\rightarrow (c/2)O_2(g) + 2ce^- \end{aligned}$$

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 new 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 can be produced in an electronically conducting form by small degrees of reduction, although only Nb_2O_5 and WO_3 use crystallographic shear to accommodate the oxygen loss, the other oxides relying upon point defects for this. 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.

9.6 Pourbaix diagrams

9.6.1 Passivation, corrosion and leaching

Many reactions that occur in water are sensitive to acidity, electrolyte concentration, and to the relative oxidising or reducing conditions in the neighbourhood. 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 is 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 certain species are 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 (Section 4.5).

9.6.2 The stability field of water

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 9.22). Above the upper boundary, water is oxidised to O_2 gas, and below the lower line it is reduced to H_2 gas.

The starting point in plotting the stability field of water (at 25°C) is the Nernst equation in the form:

$$E = E^0 - \left(\frac{0.05916}{n}\right)\log Q$$

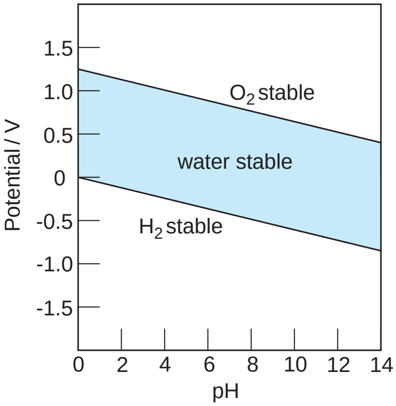


Figure 9.22 Pourbaix diagram showing the stability field of water.

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:

$$\begin{aligned} E &= 1.23 - \left(\frac{0.05916}{4}\right)\log\left(\frac{1}{p_{O_2}[H^+]^4}\right) \\ &= 1.23 + (0.01479)\log\left(p_{O_2}[H^+]^4\right) \\ &= 1.23 + (0.01479)\log p_{O_2} \\ &\quad + (0.05916)\log[H^+] \end{aligned}$$

For the normal oxygen partial pressure, p_{O_2}, of 0.21 atm, we find:[3]

$$E = 1.22 - 0.05916\,pH\ (volts) \qquad (9.2)$$

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 $Q = p_{H_2}[OH^-]^2$:

$$E = E^0 - \left(\frac{0.05916}{2}\right)\log(p_{H_2}[OH^-]^2)$$

Taking the standard pressure of H_2 (1 atm):

$$E = -0.83 - (0.05916)\log[OH^-]$$

Converting to pH:[4]

$$E = -0.05916\,pH \qquad (9.3)$$

Equations (9.2) and (9.3) are plotted on the Pourbaix diagram. The stability field of water is the area between the two lines.

9.6.3 Pourbaix diagram for a metal showing two valence states, M^{2+} and M^{3+}

Transition metals display several valence states. Generally, the valence state that is stable depends upon the acidity and the oxidation potential of the environment. Iron and its compounds illustrate these possibilities. 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 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 9.23).

[3] $pH = -\log_{10}[H^+]$.

[4] $\log[OH^-] = pH - 14$.

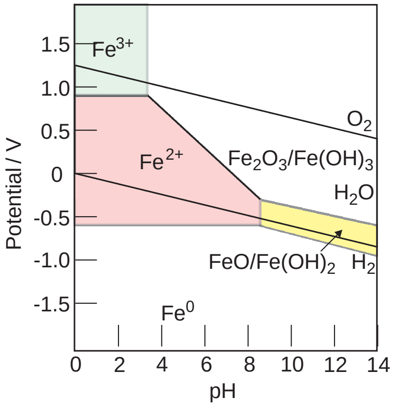


Figure 9.23 Pourbaix diagram showing the stable species in the iron–water–oxygen system.

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* since it separates 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 above:

$$K = \frac{[H^+]^3}{[Fe^{3+}]}$$

so
$$\begin{aligned}\log K &= 3\log[H^+] - \log[Fe^{3+}]\\ &= -3\text{pH} - \log[Fe^{3+}]\end{aligned}$$

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 the 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 it separates 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^0:

$$Fe^{2+} + 2e^- \rightarrow Fe(s) \qquad E^o = -0.44\text{ V}$$

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:

$$K = \frac{[H^+]^2}{[Fe^{2+}]}$$

$$\text{so} \qquad \log K = 2\log[H^+] - \log[Fe^{2+}]$$
$$= -2\text{pH} - \log[Fe^{2+}]$$

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.

$$2Fe(OH)_3(s) + 6H^+(aq) + 2e^-$$
$$\rightarrow 2Fe^{2+}(aq) + 6H_2O(l) \quad E^0 = 0.94\ \text{V}$$

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:

$$M^{3+}(OH)_3(s) + H^+(aq) + e^-$$
$$\rightarrow M^{2+}(OH)_2(s) + H_2O(l)$$

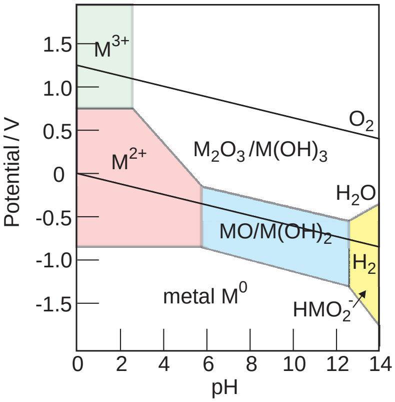


Figure 9.24 Generalised Pourbaix diagram showing the stable species in a system containing a metal capable of two valence states, M^{2+} and M^{3+}.

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 9.24).

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.

9.6.4 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 9.23). The regions in which a solid occurs are less likely to corrode extensively as the initial

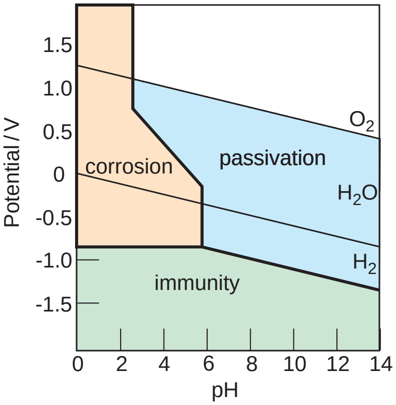


Figure 9.25 Simplified version of a Pourbaix diagram showing the range of conditions over which corrosion, passivation and immunity are likely to occur for a metal capable of two valence states, M^{2+} and M^{3+}.

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 subject to corrosion, can be used to label the stability fields on the Pourbaix diagram, if that is of prime importance (Figure 9.25).

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. These are of high quality for well-known systems such as iron–water–air, but for some systems involving radioactive materials the data are 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 8.7) is an attempt to determine whether a film is likely to fulfil the requirements of passivation.

Further reading

For a general introduction to electrochemistry, see:

Shriver, D.F., Atkins, P.W. and Langford, C.H. (1994) Chapter 7, *Inorganic Chemistry*, 2nd edn. Oxford University Press, Oxford.

Structure–property relations and defects in electrode and electrolyte solids is described in:

Tilley, R.J.D. (2008) *Defects in Solids*. John Wiley & Sons, Ltd., Hoboken; especially Chapters 6 and 8.

Batteries:

Armand, M. and Tarascon, J.-M. (2008) Building better batteries. *Nature*, **451**: 652–7.

Bruce, P.G., Freunberger, S.A., Hardwick, L.J. and Tarascon, J.-M. (2012) Li-O_2 and Li-S batteries with high energy storage. *Nature Materials*, **11**: 19–29.

Dell, R.M. and Rand, D.A.J. (2001) *Understanding Batteries*. Royal Society of Chemistry, London.

Sadoway, D.R. and Mayes, A.M. (2002) *Materials Research Society Bulletin*, **27**: 590.

Vincent, C.A. and Scrosati, B. (1997) *Modern Batteries*, 2nd edn. Elsevier, Amsterdam.

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**: 4245–69.

Yoshino, A. (2012) The birth of the lithium-ion battery. *Angew. Chem. Int. Ed.*, **51**: 5798–5800.

Corrosion:

Roberge, P.R. (2008) *Corrosion Engineering: Principles and Practice*. McGraw-Hill.

Trethewey, K.R. and Chamberlain, J. (1995) *Corrosion*, 2nd edn. Longman, New York.

Electroplating:

Kanami, N. (2004) *Electroplating: Basic Principles, Processes and Practice*. Elsevier, Oxford.

Pourbaix diagrams:

Brookins, D.A. (1988) *Eh–pH Diagrams for Geochemistry*. Springer-Verlag, Berlin.

Geological Survey of Japan (2005) *Atlas of Eh–pH Diagrams*, at www.gsj.jp.

Problems and exercises

Quick quiz

1 Oxidation is equivalent to:
(a) Electron gain.
(b) Electron loss.
(c) Electron transfer.

2 A redox reaction is one in which:
(a) Oxidation *and* reduction occurs.
(b) Oxidation *or* reduction occurs.
(c) Oxygen takes part.

3 In a galvanic cell the anode:
(a) Is the negative terminal.
(b) Is the positive terminal.
(c) Links the cell compartments.

4 Batteries are examples of:
(a) Electrolytic cells.
(b) Corrosion cells.
(c) Galvanic cells.

5 In a battery, oxidation takes place at the:
(a) Cathode.
(b) Anode.
(c) Neither.

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 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.

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.

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.

10 A rechargable battery is called:
(a) A primary cell.
(b) A secondary cell.
(c) A fuel cell.

11 In an alkaline cell the zinc is:
(a) Oxidised and forms the anode.
(b) Reduced and forms the cathode.
(c) Neutral and forms the container.

12 In a lithium primary cell the lithium is:
(a) Oxidised and forms the anode.
(b) Reduced and forms the cathode.
(c) Neutral and forms the container.

13 When a metal reacts with an acid it is:
(a) Reduced.
(b) Oxidised.
(c) Neither oxidised or reduced, simply dissolved.

14 When a metal reacts with water it is:
(a) Oxidised.
(b) Reduced.
(c) Neither oxidised or reduced, simply corroded.

15 During dissimilar metal corrosion, the metal that corrodes is:
(a) The anode.
(b) The cathode.
(c) Neither.

16 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.

17 A Pourbaix diagram plots:
(a) Oxidation potential against free energy.
(b) Oxidation potential against temperature.
(c) Oxidation potential against pH.

18 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

9.1 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.

9.2 What is the standard reaction free energy for the cell reaction of a Daniel cell?

9.3 Estimate the value of *RT/nF* for monovalent, divalent and trivalent ions at 27°C.

9.4 (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}.

9.5 Determine the voltage of the cell in the previous question if it is operated at 50°C.

9.6 A cell is constructed with a hydrogen electrode:

$$\mathrm{Pt}|\mathrm{H_2(g)}|\mathrm{H^+(aq)} \,\|\, \mathrm{Cu^{2+}(aq)}|\mathrm{Cu} \quad E^0 = 0.34\ \mathrm{V}$$

(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.

9.7 For the cell with an overall cell reaction:

$$\mathrm{Zn(s)} + \mathrm{Fe^{2+}(aq)} \rightarrow \mathrm{Zn^{2+}(aq)} + \mathrm{Fe(s)}$$

(a) Write the anode and cathode reactions.
(b) Determine the standard cell potential.
(c) Determine the reaction free energy of the cell reaction.

9.8 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}, and (b) if the temperature of the cell is subsequently raised to 35°C?

9.9 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.

9.10 A pH meter of the type in question 9.9 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.

9.11 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. The results found after several days are: (a) rust present, especially on the head and point; (b) little rust present; (c) dense ring of rust at the waterline and pitting just below the surface of the water. Explain these findings. (See

Guichelaar and Williams (1990) A simple demonstration of corrosion cells. *J. Mater. Education*, 12: 331.)

9.12 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?

9.13 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 and the cell reaction.

(c) Calculate the standard cell voltage.

9.14 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.

9.15 (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 3 hours through the cell?

9.16 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 moles of aluminium?

(b) What quantity of electricity must be passed to produce 2 kg of aluminium?

(c) If a current of 25 A is used, how long will it take to produce these amounts?

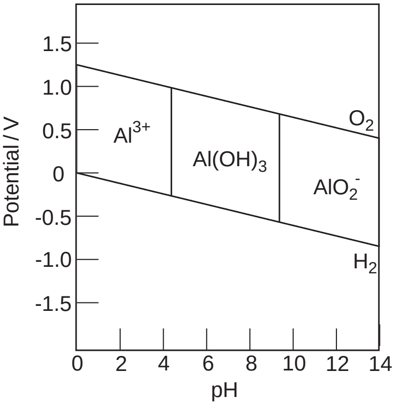


Figure 9.26 Pourbaix diagram for the Al–H_2O system.

9.17 A solution of a metal sulphate, MSO_4, is used to electroplate an object. The plating bath is operated for 2 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.

9.18 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?

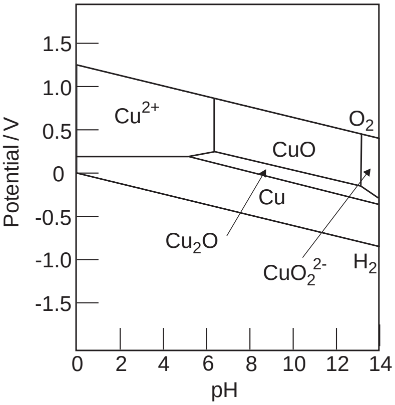


Figure 9.27 Pourbaix diagram for the Cu–H_2O system.

9.19 Figure 9.26 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?

9.20 Figure 9.27 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?

9.21 The water types met with in nature are:

(a) Fresh surface waters, pH ~ 7, E ~0.6 V.

(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 V.

Sketch these onto the stability field of water on the Pourbaix diagram (Figure 9.27). What copper species will be present in each of these water types?

PART 4

Physical properties

10

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 below 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 role of defects is described or when nanoscale properties are investigated.

10.1 Strength and hardness

10.1.1 Strength

Everyone has a subjective idea of strength. Some materials such as steel are universally regarded as strong, while others, like plastics, are considered weak. However, the strength of a material will depend upon exactly how it is evaluated. A reliable measure of the strength of a solid is the amount of force that can be applied to it before it breaks (Figure 10.1). 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

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

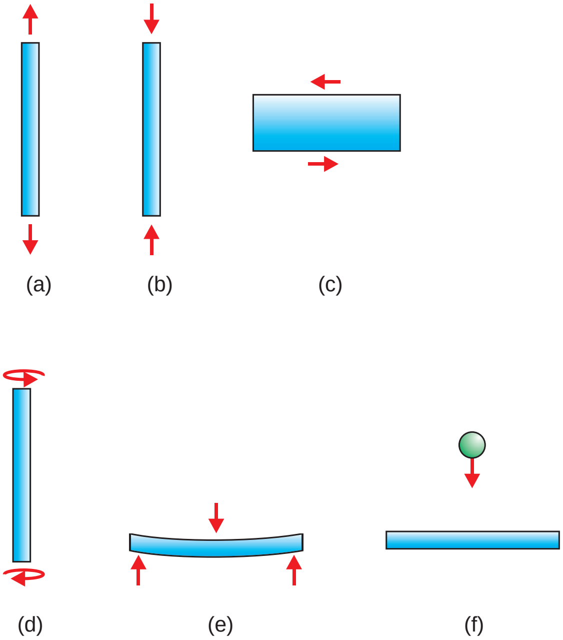


Figure 10.1 Mechanical loading of solids: (a) tension; (b) compression; (c) shear; (d) torsion; (e) flexing; (f) impact.

impact strength of a solid measures the resistance to a sudden blow. Glass has low impact strength, while wooden bats have high impact strength. A component may experience all of these forces and resultant deformations at the same time, but here each will be treated separately.

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 (see Section 10.2).

10.1.2 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 10.2), used for the evaluation of metal or polymer samples:

$$\sigma = \frac{F}{A}$$

where σ is the 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 ($\mathrm{N\,m^{-2}}$), commonly cited as $\mathrm{MPa} = 10^6\,\mathrm{N\,m^{-2}}$ or $\mathrm{GPa} = 10^9\,\mathrm{N\,m^{-2}}$. 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.

$$\sigma(\text{engineering stress}) = \frac{F}{A_0}$$

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 tensile 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:

$$\Delta\varepsilon = \frac{\Delta l}{l}$$

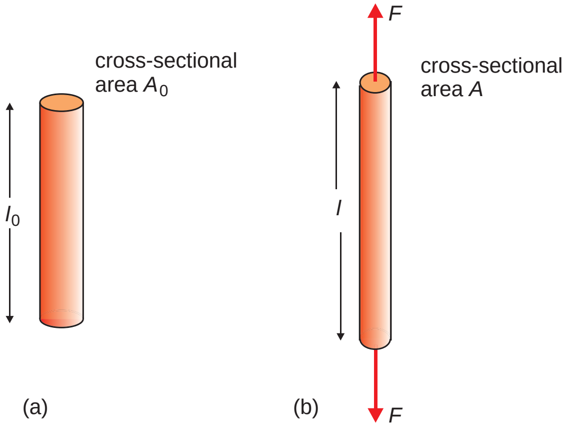


Figure 10.2 A rod in tension: (a) initial state; (b) final state.

The total strain is then given by:

$$\varepsilon = \int_{l_0}^{l} \frac{dl}{l} = \ln\left(\frac{l}{l_0}\right)$$

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:

$$\varepsilon\,(\text{engineering strain}) = \frac{(l - l_0)}{l_0}$$

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 (Figure 10.3). In this 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:

$$\sigma_m = \frac{3F\,l}{2a\,b^2}$$

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 10.3). The strain, ε_m, is related to the maximum deflection, δ, at the centre of the bar, given by:

$$\varepsilon_m = \frac{6\delta\,b}{l}$$

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.

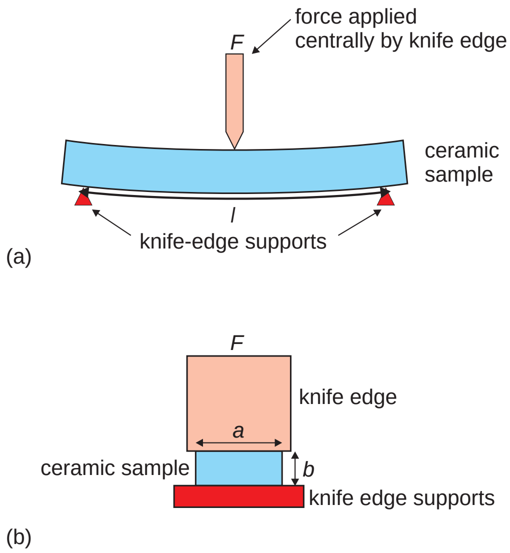


Figure 10.3 The three-point bend test for ceramic samples: (a) sideways; (b) end-on view of the test.

10.1.3 Stress–strain curves

A great deal can be learned about the mechanical properties of materials by stressing them until they fracture. 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 in form with a central cylindrical section, of known *gauge length*, usually 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 (Figure 10.4). 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 (Figure 10.4a). Metals initially show a similar linear relationship, but the plot ultimately curves and extensive deformation occurs before fracture (Figure 10.4b). This type of curve is typical of a *ductile* solid. Ductile solids are those that can be drawn out into wires or permanently deformed under an applied load. 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 10.4c). Elastomers (Figure 10.4d) deform at far lower stress levels than other materials.

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. In the case of elastomers, this reversibility is maintained over the whole of the stress–strain curve. 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*:

$$\sigma = E\,\varepsilon$$

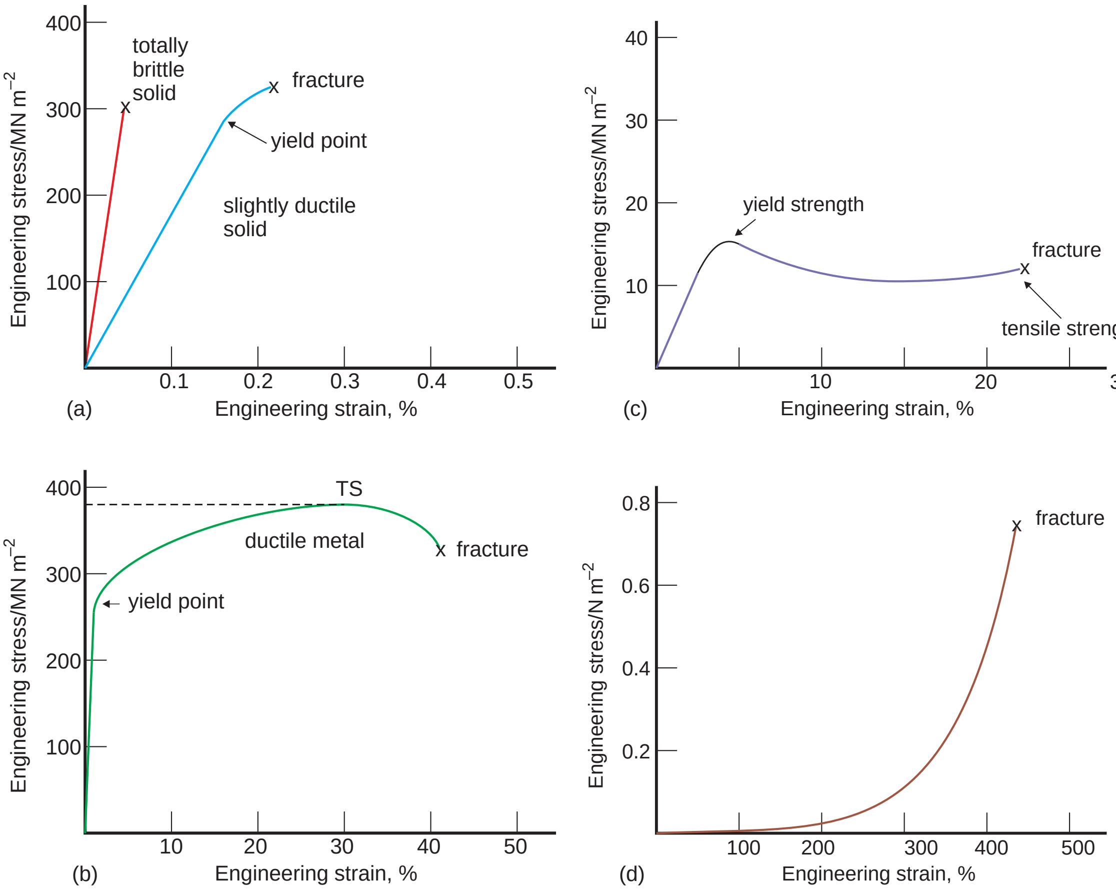


Figure 10.4 Schematic engineering stress–strain curves: (a) brittle and slightly ductile solids; (b) ductile metals; (c) a typical polymer; (d) a typical elastomer. Note the different stress scale in (d).

The constant of proportionality, E, given by the slope of the stress–strain plot, is the Young's modulus of the material (Figure 10.5a). As weight is an important consideration in many applications, the *specific modulus* is often quoted as a parameter:

$$\text{specific modulus} = \frac{\text{Young's modulus}}{\text{specific gravity}}$$

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 *non-linear* 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 10.5b,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 (Figure 10.4a), 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 10.4b) 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*:

$$\text{specific strength} = \frac{\text{tensile strength}}{\text{specific gravity}}$$

For a polymer, once the elastic region is passed, almost no increase in stress will bring about a large amount of plastic deformation (Figure 10.4c). Anyone who has carried an overloaded plastic bag for

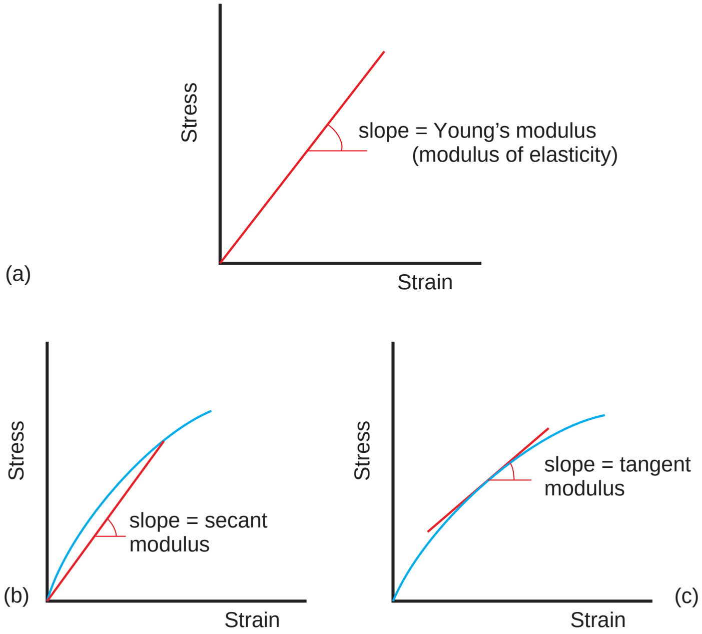


Figure 10.5 (a) Young's modulus (the modulus of elasticity) is defined as the slope of the stress–strain curve in the linear (elastic) region. The secant (b) and tangent (c) modulus is used in non-linear materials.

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–strain tensile curve are Young's modulus, yield strength, ultimate tensile strength, and amount of elongation at fracture.

10.1.4 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 can be estimated by the area under the stress–strain curve (Figure 10.6).

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 the 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 10.6) 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.

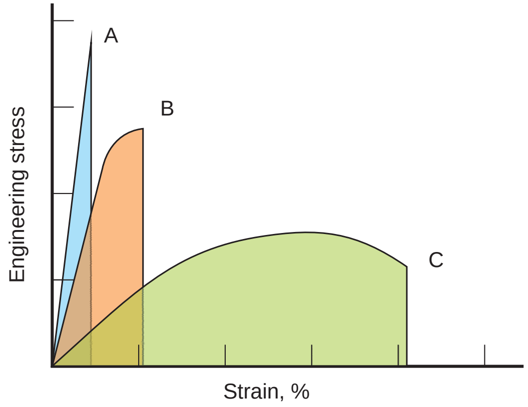


Figure 10.6 The toughness of a solid can be represented by the area under an engineering stress–strain curve up to the point of fracture.

10.1.5 Superelasticity

The stress–strain behaviour is different in many shape-memory alloys (Section 8.5.4), which show a region of extreme deformation called *superelasticity*. On release of the stress the curve usually shows a hysteresis effect (Figure 10.7). 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 8.5.3, 8.5.4). 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

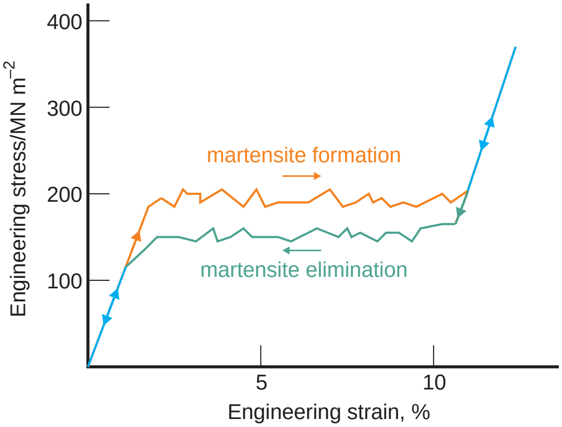


Figure 10.7 Schematic engineering stress–strain curve for a Nitinol shape-memory alloy showing superelasticity.

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.

10.1.6 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 ten minerals that, as far as possible, were equally spaced on a hardness scale of 1 to 10. Even today, the commonest description of hardness is still in terms of this scale, called *Mohs scale* of hardness. The minerals chosen, and their hardnesses are: 1, talc ($Mg_3Si_4O_{10}(OH)_2$); 2, gypsum ($CaSO_4.2H_2O$); 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 material 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 four major indentation hardness tests, which differ from each other in the shape of the indenter (Figure 10.8a–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 B, E, F and G scales, in which the indenter is steel, and the A, C and D scales, using a conical diamond indenter with a spherical tip. 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. 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. Both the Vickers and Knoop tests can use small loads, of the order of grams, and measure what is generally called *micro-hardness*.

Vickers and Knoop hardness numbers are obtained by dividing the load applied to the indenter by the area over which the force acts (not the simple cross-sectional area of the indentation). Vickers hardness number is given by:

$$\mathrm{VHN}\,(\mathrm{kg\ mm}^2) = \frac{2P\sin\left({}^{136}\!/\!_2\right)}{d^2} = \frac{1.854368\,P}{d^2}$$

where P is the load, and d is the average of the indent diagonals d_1 and d_2 (Figure 10.8d). The hardness number is then quoted as, say, 560 HV 3, where the first number specifies the hardness number ($\mathrm{kg\,mm}^{-2}$), 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: Mohs hardness $\sim 0.675 \times (\mathrm{VHN\ kg\,m}^{-2})^{1/3}$. Knoop hardness numbers are obtained in a similar way to VHN:

$$\mathrm{KHN}\,(\mathrm{kg\ mm}^2) = \frac{P}{c\,l^2}$$

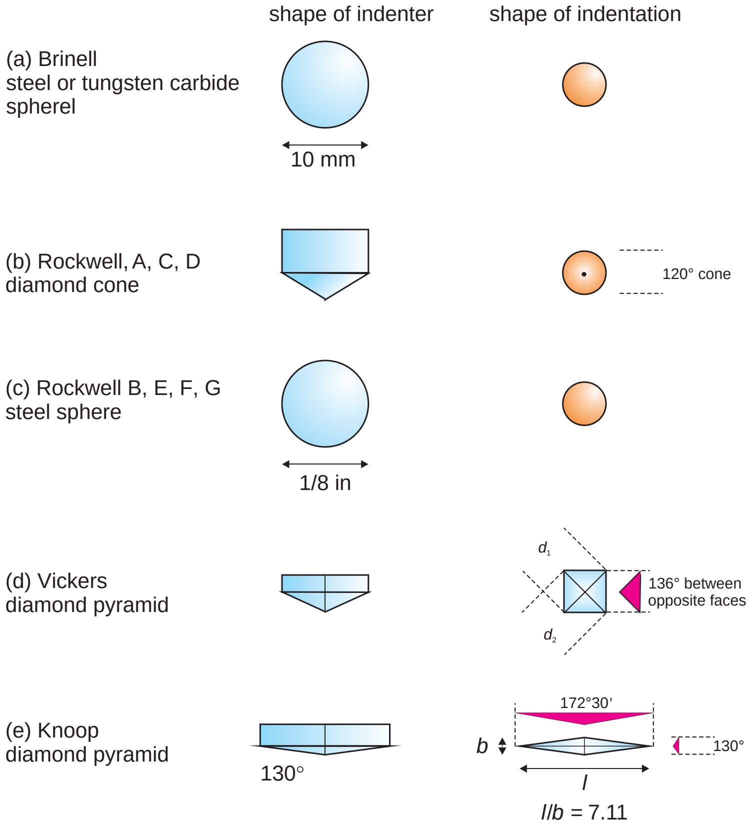


Figure 10.8 Hardness indenters.

where P is the load, l is the long diagonal of the indentation (mm) (Figure 10.8e) and c is a correction factor, ideally 0.070279. 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, 9.80665 m s^{-2}, and the dimensions (mm) are converted to metres by dividing by 1000. Multiplying the hardness number (kg mm^{-2}) by 9.80665 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 10.2.4) and hardness (Figure 10.9). This allows the hardness of new materials to be predicted from knowledge of the shear modulus.

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 focused on *iso-electronic crystals*. These are

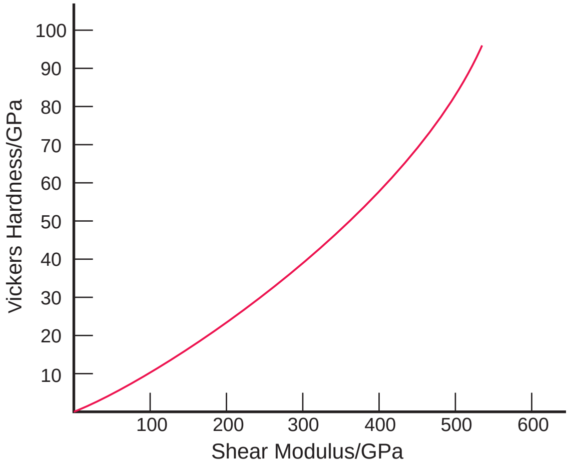


Figure 10.9 Schematic relationship between Vickers hardness and shear modulus (adapted from D.M. Teter, see Further reading).

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 known, 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 recently been synthesised, as well as other isoelectronic compounds such as B_2CO and B_5NO_2, but none, so far, has the hardness of diamond.

10.2 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. Glass 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 grains. Composites including wood are usually anisotropic, with properties in one direction different to those in 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.

10.2.1 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 10.10a,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. 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 properly defined in terms of the elastic properties of anisotropic solids, in which the stress and strain are vector quantities.)

The elastic deformation experienced is a result of pulling atoms apart or pushing atoms together, and so 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

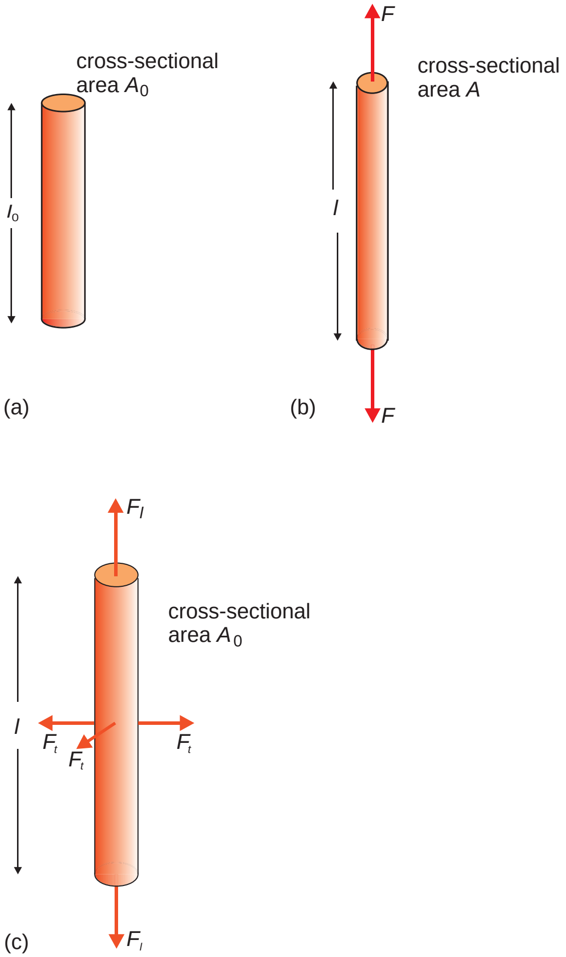


Figure 10.10 Elastic moduli: (a) sample before application of forces; (b) sample after application of tensile force (F) to obtain the Young's (elastic) modulus E; (c) sample after application of longitudinal (F_l) and transverse (F_t) forces to obtain the longitudinal modulus L.

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:

$$U = \frac{-C_1}{r} + \frac{C_2}{r^n}$$

where C_1, C_2 and n are empirical constants (Section 2.1.3). The first term in the equation represents the attractive energy between the atoms and the second term the repulsive energy (Figure 10.11a). The lattice energy, U_L, corresponds to the minimum in the total energy curve, reached at an interatomic separation of r_0. The force between the atoms is defined as F, where:

$$F = \frac{-dU}{dr} = \frac{-C_1}{r^2} + \frac{nC_2}{r^{n+1}}$$

This is also the sum of an attractive and repulsive term (Figure 10.11b). At the equilibrium separation, r_0, the force between the atoms will be zero, hence:

$$\begin{aligned} 0 &= \frac{-C_1}{r_0^2} + \frac{nC_2}{r_0^{n+1}} \\ C_2 &= \frac{C_1 r_0^{n-1}}{n} \end{aligned}$$

For small values of r lying close to r_0, this curve can be approximated as a straight line (Figure 10.11c). 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:

$$\sigma \approx \frac{\Delta F}{{r_0}^2}, \qquad \varepsilon \approx \frac{\Delta r}{r_0}$$

Hooke's law gives:

$$E = \frac{\sigma}{\varepsilon} = \frac{\Delta F}{\Delta r\, r_0}$$

$\Delta F/\Delta r$ is simply the slope of the curve (dF/dr) at r_0, hence:

$$\left(\frac{dF}{dr}\right)_{r=r_0} = \frac{2C_1}{{r_0}^3} - \frac{n(n+1)\,C_2}{{r_0}^{n+2}} = \frac{C_1(1-n)}{{r_0}^3}$$

Thus:

$$E = \frac{C_1(1-n)}{{r_0}^4}$$

In terms of this simple model, Young's modulus depends sensitively upon the interatomic distance in

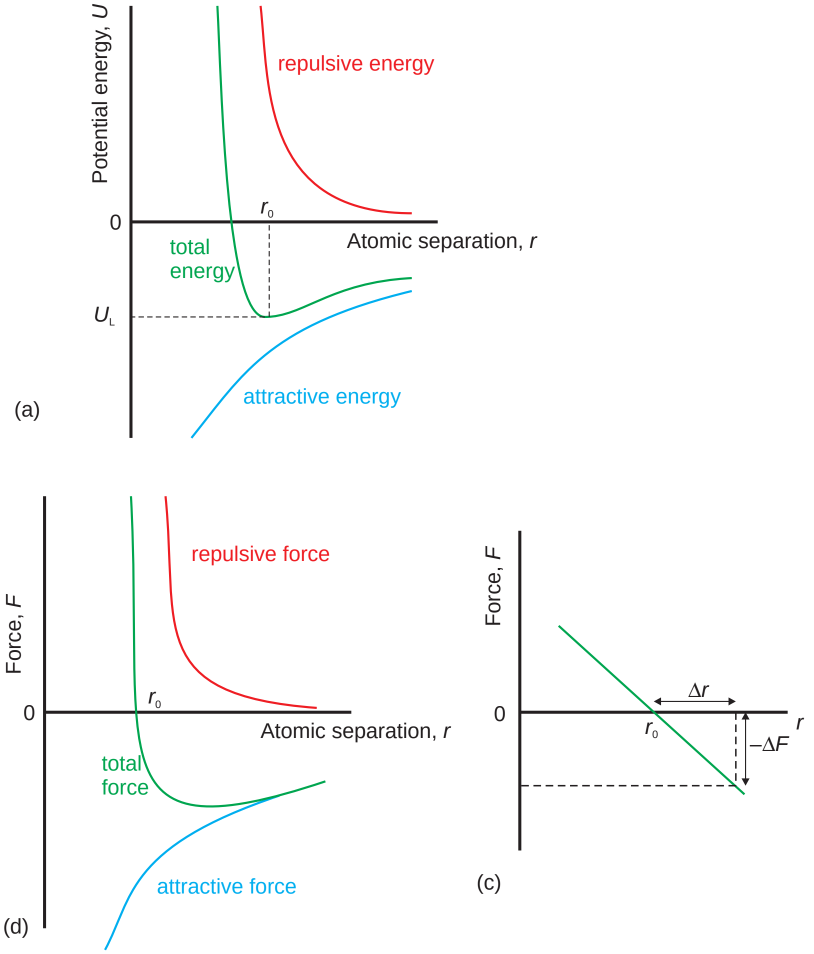


Figure 10.11 (a) The lattice potential energy U as a function of the atomic separation, r. (b) The forces between atoms, F as a function of the atomic separation, r. (c) Enlarged portion of the F–r curve close to r_0.

the solid. Atomistic simulations (Section 2.1.4) using accurate potential energy functions between different atoms perform these calculations routinely to give very accurate values of elastic constants.

Representative values of Young's modulus are given in Table 10.1.

10.2.2 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 $-\varepsilon_x = -\varepsilon_y$):

$$\nu = \frac{-\varepsilon_x}{\varepsilon_z}$$

The negative sign is to ensure that the numerical value of Poisson's ratio is positive (Table 10.1). For isotropic materials, the theoretical value of ν is $\frac{1}{2}$. Most metals show values in the range 0.25–0.35. Some

Table 10.1 Representative values of Young's modulus and Poisson's ratio

Material	*Young's modulus* E/GPa	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

materials, counter-intuitively, have a negative Poisson's ratio, and get thicker when under tension (Section 10.5.2).

10.2.3 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 10.10c). 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:

$$\varepsilon = \frac{(l - l_0)}{l_0}$$

This modulus determines the velocity of ultrasonic stress pulses through a solid.

10.2.4 The shear modulus or modulus of rigidity (G or μ)

The shear stress and resulting shear strain can be most easily defined for a rectangular block of material (Figure 10.12a,b). The shear stress, τ, is given by the ratio of the force (or load) F applied to one face of the block to the area of these faces:

$$\tau = \frac{F}{A_0}$$

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:

$$\gamma(\text{shear})\text{strain} = \frac{a}{b} = \tan\phi$$

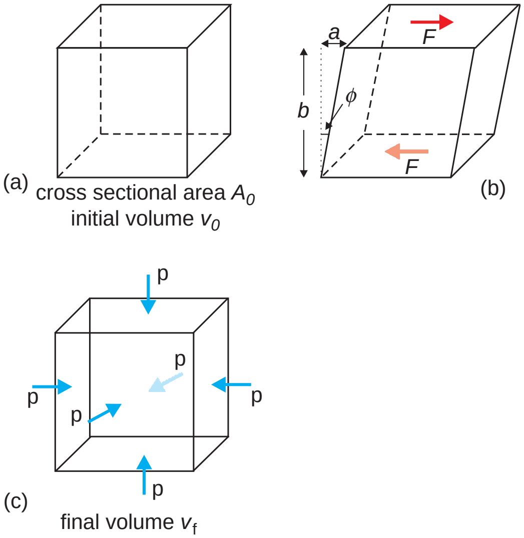


Figure 10.12 Elastic moduli: (a) sample before application of forces or pressures; (b) sample after application of shear forces (F) to give shear deformation and obtain the shear modulus G; (c) sample after application of a uniform pressure (p) to obtain the bulk modulus B or K.

The shear modulus defines the response of a body, the shear strain, to a shear stress tending to distort it. The relationship between these quantities 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* (also see Section 10.2.6).

10.2.5 The bulk modulus, K or B

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 10.12c):

$$p = \frac{B(\mathrm{v}_\mathrm{f} - \mathrm{v}_0)}{\mathrm{v}_0} = B\,\Delta V$$

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 2.1.4) 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.

10.2.6 The Lamé modulus (λ)

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 10.2.8).

10.2.7 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:

$$\begin{aligned}
E &= 2G\,(1+\nu) = 3B\,(1-2\nu)\\
G &= \frac{E}{2\,(1+\nu)}\\
B &= \frac{E}{3\,(1-2\nu)} = \frac{EG}{3(3G-E)}\\
L &= B + \left(\frac{4G}{3}\right)\\
\lambda &= B - \left(\frac{2G}{3}\right) = \frac{E\mathrm{v}}{(1+\mathrm{v})(1-2\mathrm{v})}
\end{aligned}$$

10.2.8 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 atoms making up the solid move in a direction parallel to the direction of propagation of the wave (Figure 10.13a). 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 atoms making up the solid move in a direction perpendicular to the direction of propagation of the wave (Figure 10.13b). These waves move the atoms to and fro 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.

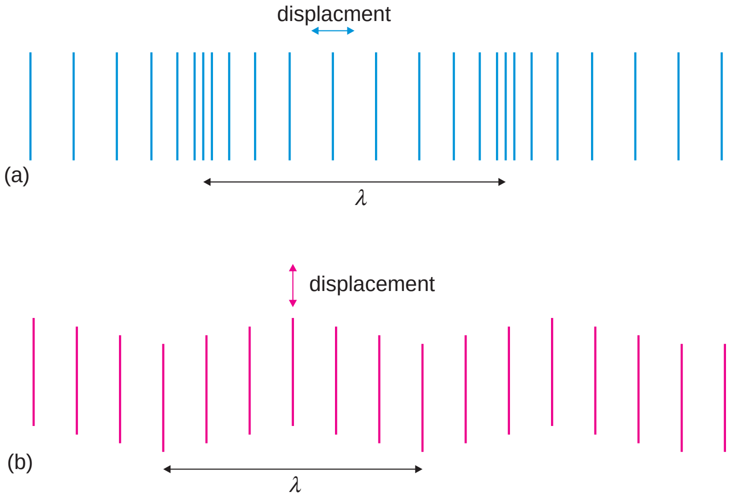


Figure 10.13 Waves in solids schematic: (a) longitudinal waves, atom planes move parallel to the wave propagation; (b) transverse waves, atom planes move perpendicular to the wave propagation.

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 11.2), 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 10.14a). The crystal can respond in one of two ways, depending upon

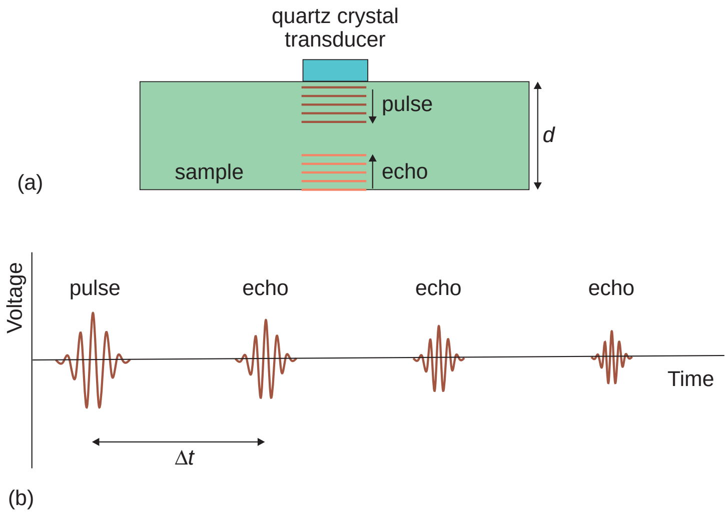


Figure 10.14 The pulse–echo technique for elastic moduli: (a) schematic of pulse and echo generation; (b) voltage (pulse and echoes) as a function of time.

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 10.14b). The round time taken by the pulse gives the velocity of the ultrasound v in the material:

$$\mathrm{v} = \frac{2d}{\Delta t}$$

where d is the sample thickness and Δt the delay time between the pulse echoes.

The pulse velocity of longitudinal waves $\mathrm{v_L}$ is given by:

$$\mathrm{v_L} = \sqrt{\frac{E(1-\nu)}{\rho\,(1+\nu)(1-2\nu)}}$$

where the symbols have the same meanings as in the previous sections and ρ is the sample density. The formula can also be written in terms of other elastic constants, for example:

$$\mathrm{v_L} = \sqrt{\frac{\lambda + 2G}{\rho}} = \sqrt{\frac{B + {}^{4G}\!/_{3}}{\rho}}$$

The pulse velocity of transverse waves $\mathrm{v_T}$ is given by:

$$\mathrm{v_T} = \sqrt{\frac{E}{2\rho\,(1+\nu)}} = \sqrt{\frac{G}{\rho}}$$

and

$$\frac{\mathrm{v_L}}{\mathrm{v_T}} = \sqrt{\frac{2\,(1-\nu)}{(1-2\nu)}}$$

Taking a value of ν to be 0.3 it is found that:

$$\frac{\mathrm{v_L}}{\mathrm{v_T}} = \sqrt{3.5}$$

The elastic constants can be written in terms of the $\mathrm{v_L}$ and $\mathrm{v_T}$ velocities thus:

$$\begin{aligned}
E &= 2\rho \mathrm{v_T}(1+\nu)\\
G &= \rho \mathrm{v_T}\\
B &= \rho\left(\mathrm{v_L^2} - \frac{4\mathrm{v_T^2}}{3}\right)\\
\nu &= {}^{1}\!/_{2}\left[1 - \left(\frac{\mathrm{v_T^2}}{\mathrm{v_L^2} - \mathrm{v_T^2}}\right)\right] = \frac{1 - 2\left({}^{\mathrm{v_T^2}}\!/_{\mathrm{v_L^2}}\right)}{2 - 2\left({}^{\mathrm{v_T^2}}\!/_{\mathrm{v_L^2}}\right)}
\end{aligned}$$

A measurement of the velocities of both the longitudinal and transverse waves then allows all of the elastic moduli to be determined.

10.3 Deformation and fracture

10.3.1 Brittle fracture

Many materials are elastic and brittle, especially at lower temperatures, and have a linear stress–strain curve to fracture (Figure 10.4a). Single crystals frequently fracture by cleavage, along crystal planes in which the bonding is relatively weak (Figure 10.15a). 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 10.15b). Amorphous materials such as glass or polymers that become brittle at low temperatures will fracture to produce a smooth surface resembling the inside of a seashell. This is called *conchiodal* fracture (Figure 10.15c). Materials containing voids, such as porcelain, frequently fracture in the neighbourhood of these defects (Figure 10.15d).

Rapid failure occurs when a crack propagates through the solid by the breaking of successive

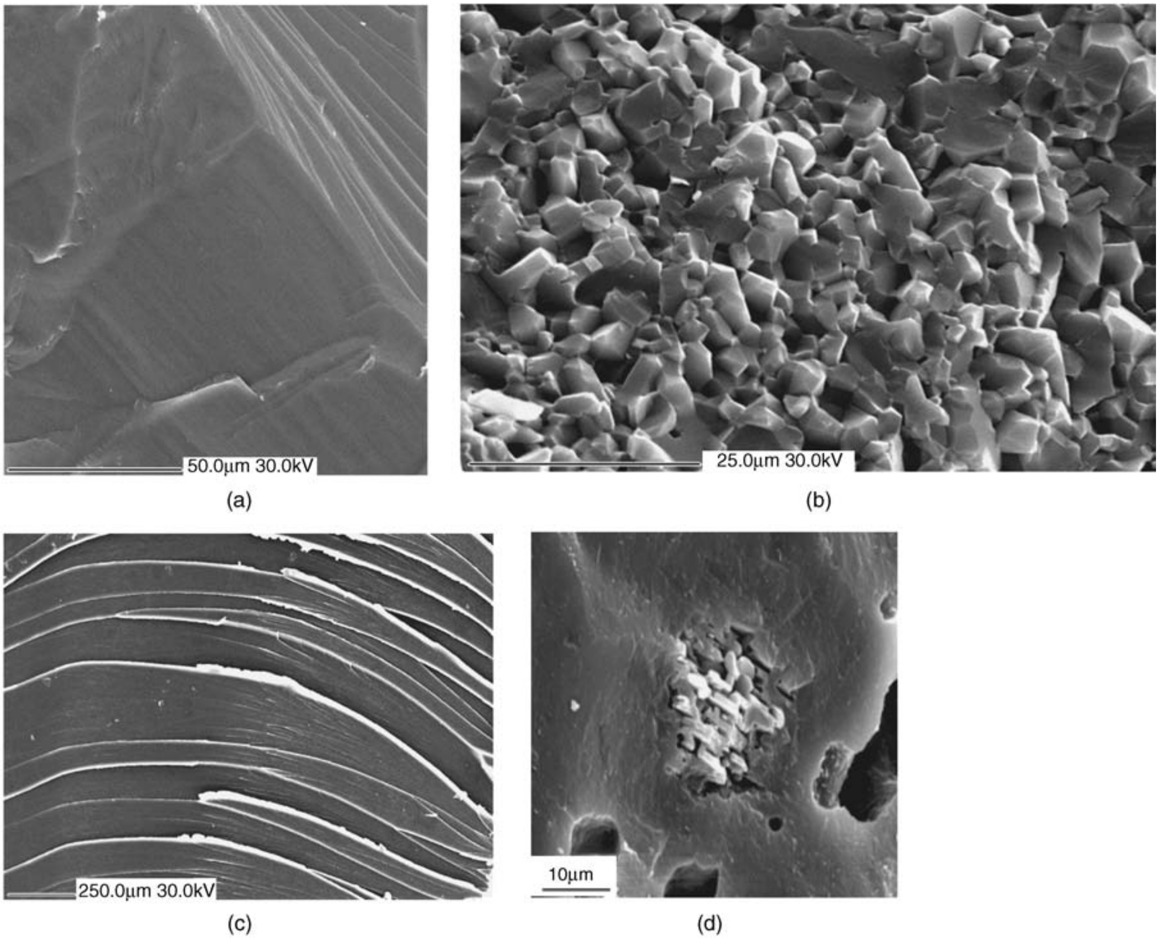


Figure 10.15 Fracture: (a) cleavage fracture of single crystalline silicon; (b) transgranular fracture of a fine-grained polycrystalline alumina ceramic; (c) conchoidal fracture of brittle poly(methyl methacrylate); (d) fracture of porcelain showing a cluster of crystallites and pores in a glass-like matrix.

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 (Section 10.2.1). The reasoning, published in the first quarter of the 20th century, runs as follows. The stress is zero at the equilibrium spacing, r_0. 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:

$$\sigma_c = \sqrt{\frac{E\,\gamma}{r_0}}$$

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 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 $E/100$ to $E/1000$. 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:

$$K_t = 1 + 2\sqrt{\frac{a}{\rho}}$$

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 10.16). For a relatively long crack with a sharp tip:

$$K_t \approx 2\sqrt{\frac{a}{\rho}}$$

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:

$$\sigma_c = \sqrt{\frac{E\gamma}{2a}}$$

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 time 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*:

$$\sigma_c = \sqrt{\frac{2\gamma E}{\pi a}}$$

where γ is the surface energy of the solid, E is Young's modulus and a is the length of a surface

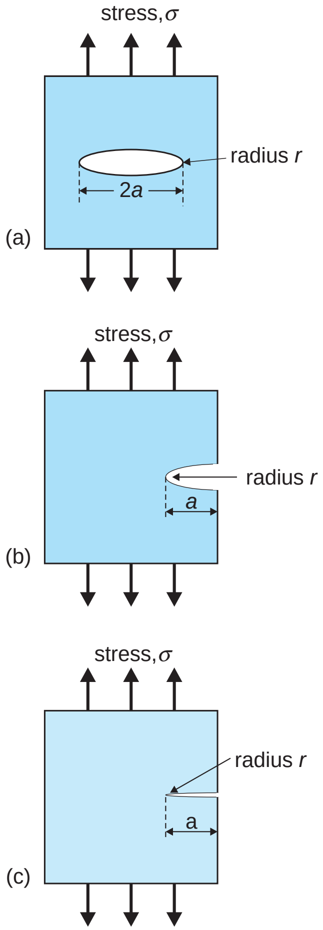


Figure 10.16 Stress at (a) an elliptical pore in a solid; (b) an elliptical groove in a solid; (c) a sharp crack in a solid.

crack or half the length of an internal crack (Figure 10.16). 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_0:

$$\rho_c \approx 2.5\, r_0$$

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 10.3.8).

10.3.2 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 10.4). 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 10.17a), *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.

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 10.17b).

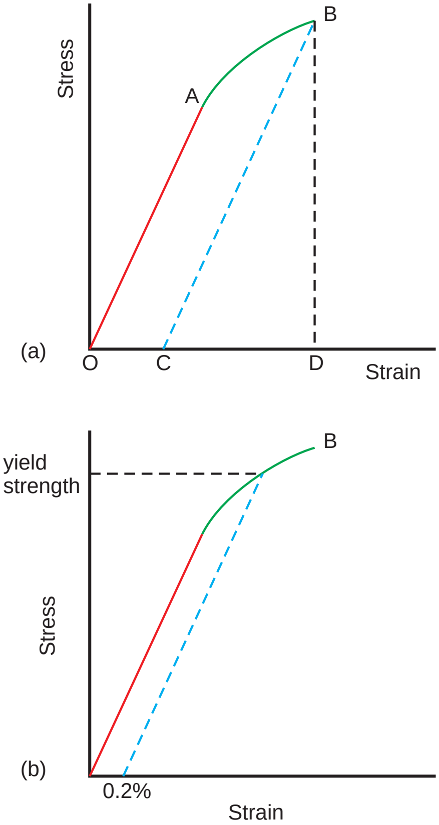


Figure 10.17 Plastic deformation: (a) permanent distortion of the material, OC; (b) the yield strength is determined by the intersection of an offset line, here at 0.2%, with the engineering stress–strain curve.

10.3.3 Dislocation movement and plastic deformation

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 (see Section 8.5). 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 10.18a,b). The slices are usually of 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 10.18c,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.

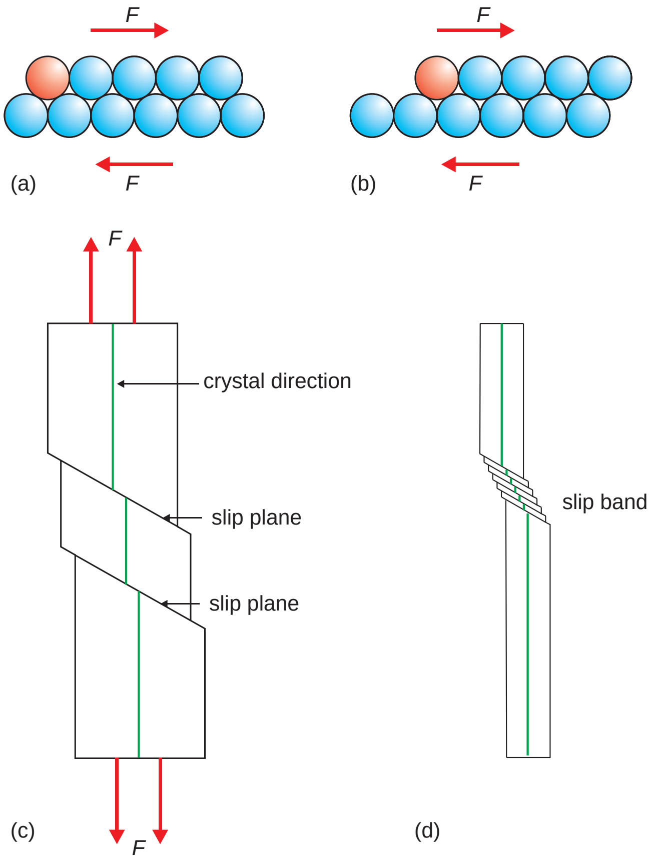


Figure 10.18 Slip: (a, b) a shear stress results in slip when atom planes slide over each other; (c) in a rod, slip is characterised by diagonal planes across which atoms in the crystal have sheared; (d) slip planes are often aggregated into narrow regions, a slip band.

The amount of energy required to move a plane of atoms from one stable position to another (Figure 10.18a) 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 (Section 3.5.1) 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 step in the crystal profile (Figure 10.19). 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 is now recognised as the process that successfully explains the deformation properties of metals.

The movement of dislocations is constrained by crystallography. 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*.

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 10.20). In these planes, the greatest linear density of atoms occurs along the ⟨110⟩ directions. In such a metal there are 4 different {111} planes and 3 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.

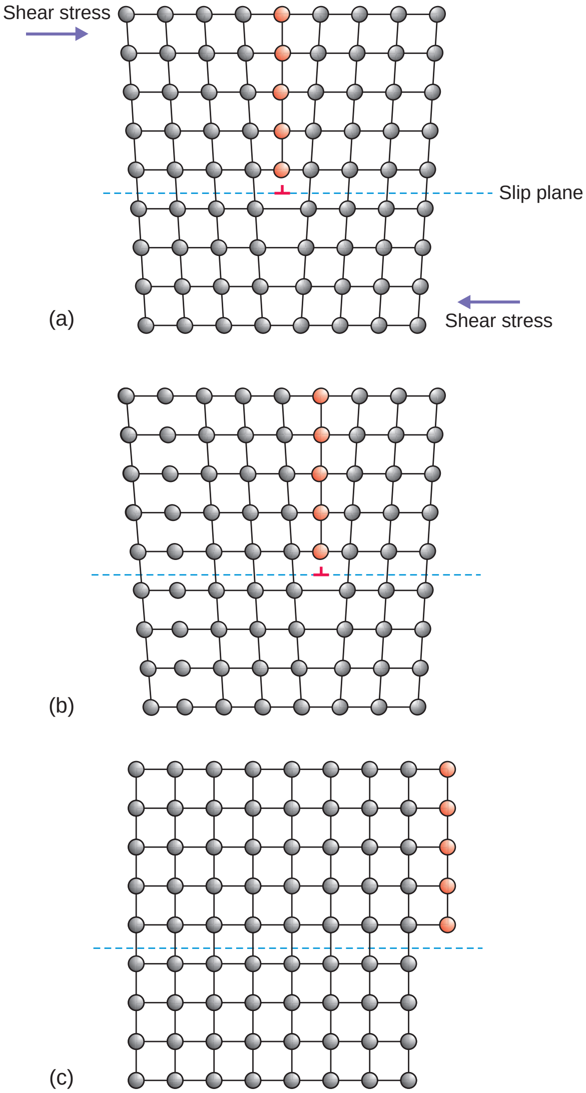


Figure 10.19 Application of a shear stress to a crystal containing an edge dislocation (a, b) can allow the dislocation to move out of the crystal to leave a surface step (c).

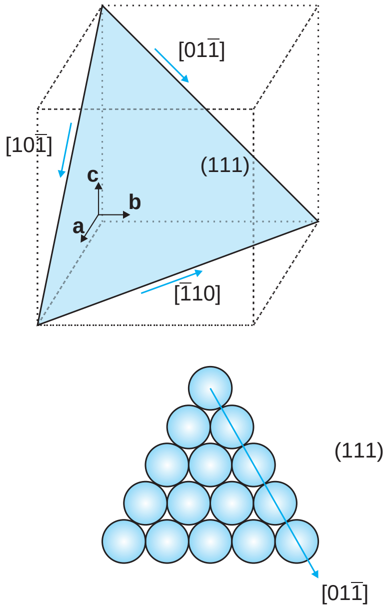


Figure 10.20 (a) The (111) plane in a crystal of a metal with the face-centred cubic (A1, copper) structure, shaded. The directions along each of the edges are given. (b) The same (111) plane represented as a packing of atoms. One direction, $[0\ 1\ \bar{1}]$, is marked.

10.3.4 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 lose 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 20th century and the onset of the 21st 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 is of importance to be able to estimate roughly 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 difficulty of slip in ceramics at normal temperatures. The preferred slip plane in ionic crystals with the halite structure, such as NaCl or LiF, is {110} and the slip direction used is $\langle 1\bar{1}0\rangle$ (Figure 10.21). For the more metallic halite structure solids such as TiC, the slip system is similar to that in face-centred cubic metals, $\{111\}\langle 1\bar{1}0\rangle$.

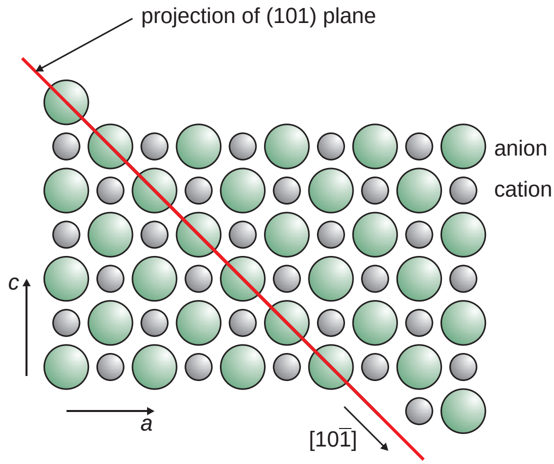


Figure 10.21 A {110} slip plane in the halite (NaCl, B1) structure.

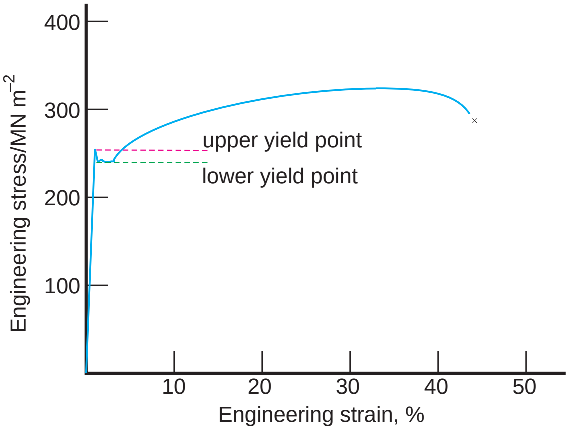


Figure 10.22 Schematic engineering stress–strain curve for a carbon steel.

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 10.22). 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 in 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.

10.3.5 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 10.4c). 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 10.23). 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.

10.3.6 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:

$$\%\,\text{elongation} = 100\left(\frac{l - l_0}{l_0}\right)$$

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 10.24). 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.

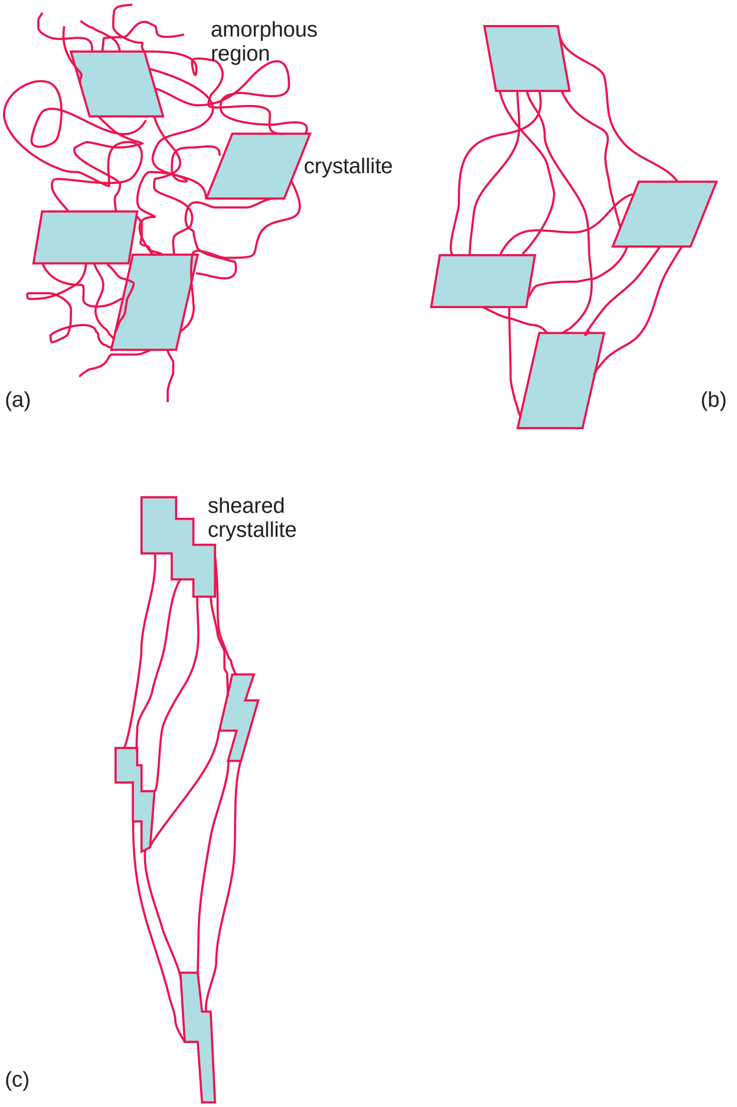


Figure 10.23 Deformation of a semi-crystalline polymer: (a) unstressed state; (b) on initial loading, the molecules in the amorphous regions elongate; (c) when the polymer chains can no longer accommodate the stress, the crystalline regions deform.

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:

$$\%\ \text{reduction in area} = 100\left(\frac{A_0 - A_\text{f}}{A_0}\right)$$

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 10.24e and 10.25).

The term ductility is generally reserved for metals, and in the case of polymers, ductility is replaced by *elongation*. The elongation is simply measured

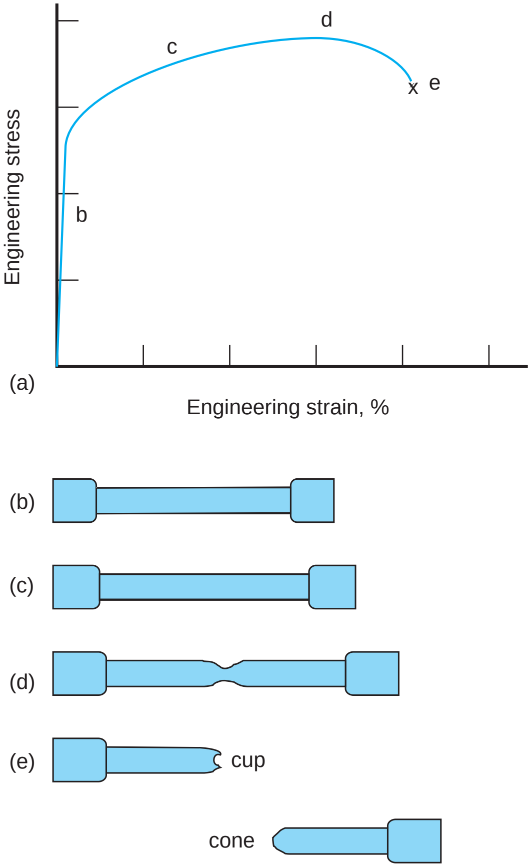


Figure 10.24 Metal ductility: (a) schematic tensile engineering stress–strain curve for a ductile metal; (b–e) the corresponding shape of the test piece during the test.

as the length of the polymer after stretching, l, divided by the original length, l_0, often given as a percentage:

$$\%\,\text{elongation} = 100\left(\frac{l}{l_0}\right)$$

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 10.26). The term *percentage elongation at fracture* used for metals is generally replaced by *percentage elongation at break* for polymers.

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 form a long tear. As with metals, the edges of the tear have a pulled-out, fibrous texture.

10.3.7 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 by the strain set up in the crystal structure by the 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. 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.

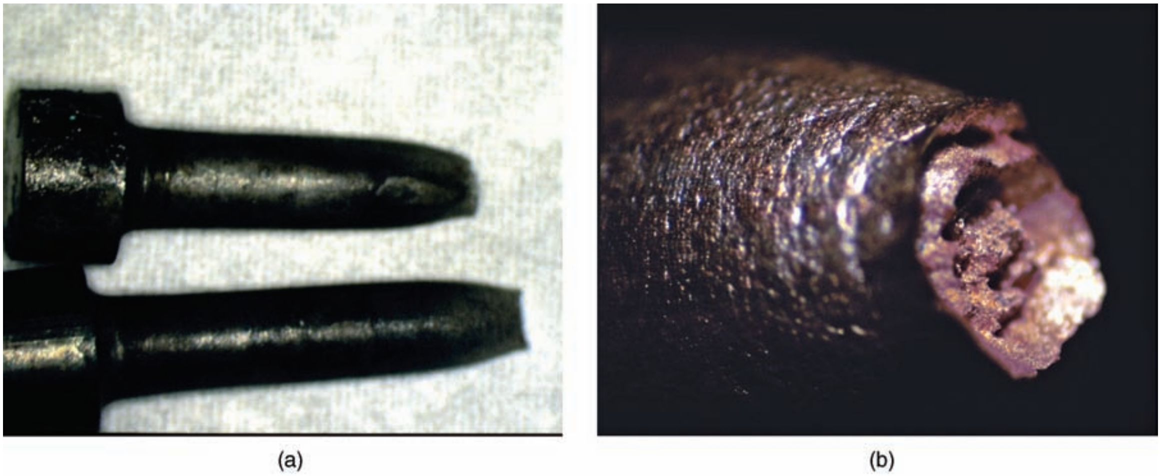


Figure 10.25 (a) A fractured semi-ductile steel tensile test specimen, showing necking and cup and cone fracture. (b) Detail of cup and cone fracture surface from a copper tensile test specimen. The neck diameter is approximately 2 mm.

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 10.27a). 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 10.27b). 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 10.27c–e). A Frank-Read source can continually emit dislocation loops during stress, thus significantly multiplying the dislocation density in a crystal.

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,

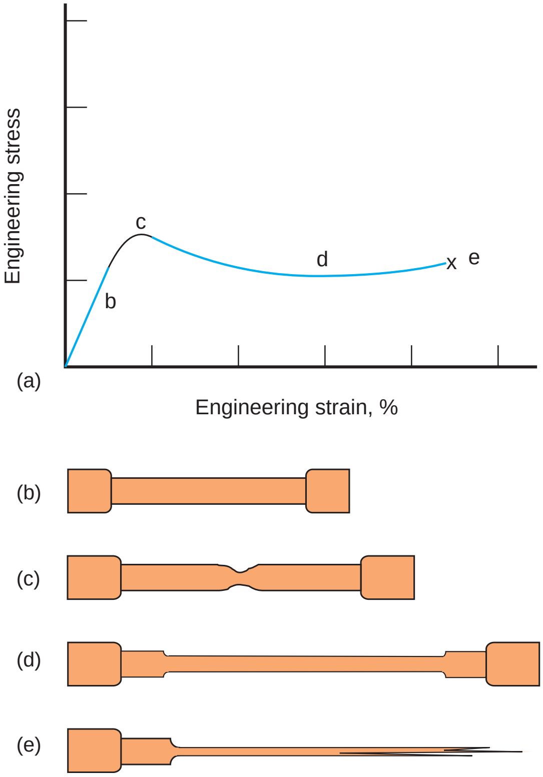


Figure 10.26 Polymer elongation: (a) schematic tensile engineering stress–strain curve for a polymer; (b–e) the corresponding shape of the test piece during the test.

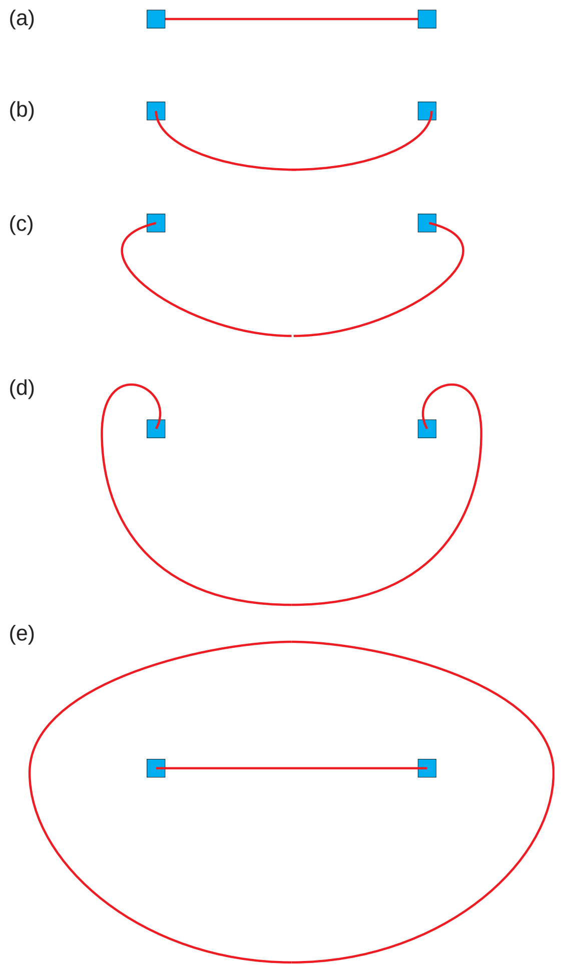


Figure 10.27 Dislocation multiplication at a Frank-Read source: a dislocation pinned at each end (a) can respond to stress by bowing (b), eventually generating a dislocation loop plus the original pinned dislocation (c–e).

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 the block copolymer strategy. In the case of metals and ceramics, the two phases remain separate, and the material is called a composite (Section 10.6).

10.3.8 Computation of deformation and fracture

Deformation, fracture and strengthening are of vital importance. Attempts to quantify brittle fracture in the early years of the 20th century, described above, 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 become 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 2.1.4) 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 can be subjected to forces or pressures causing deformation, and in this way elastic constants can 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 2.3.6). This powerful computational technique, mainly associated with band structure calculations, can provide values for the lattice parameters 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 20th 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.

10.4 Time-dependent properties

10.4.1 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 sinusoidal changes to completely irregular patterns (Figure 10.28). 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 versus log N that shows a distinct change of slope, called a *knee* (Figure 10.29). 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. For most non-ferrous alloys and pure metals such as copper or aluminium, the curve is smooth up to fracture (Figure 10.29).

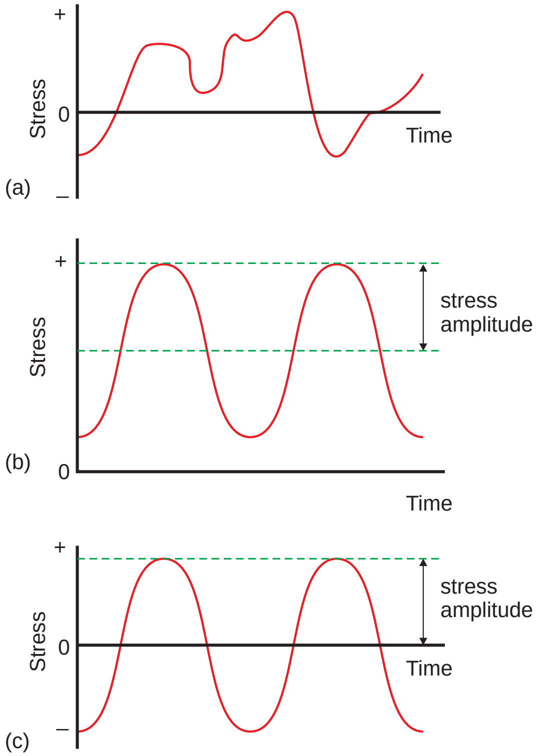


Figure 10.28 Stress cycles: (a) irregular; (b) positive sinusoidal cycle; (c) positive and negative sinusoidal cycle.

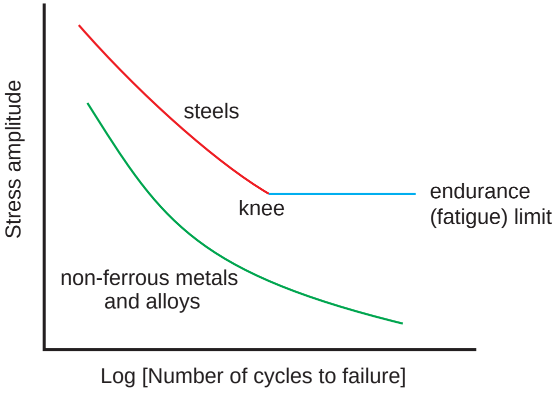


Figure 10.29 Stress amplitude plotted against log [number of cycles to failure] for typical steels and many non-ferrous metals and alloys.

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 (Figure 10.30). 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. 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 the whole part fails catastrophically, by ductile or brittle failure, depending upon the nature of the material.

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.

10.4.2 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 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* (Figure 10.31). 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 10.32).

10.4.2.1 Primary 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:

$$\varepsilon = \alpha \log t$$

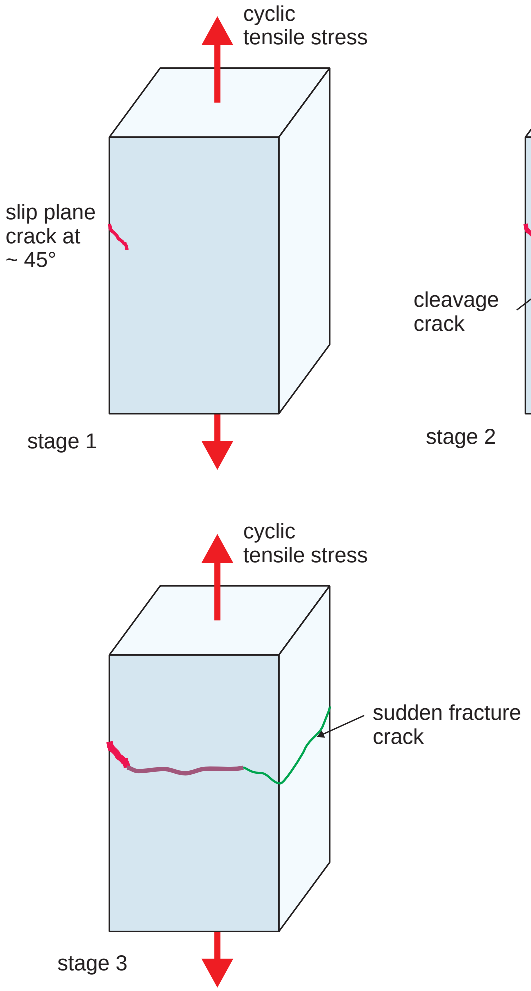


Figure 10.30 Fatigue failure: stage 1, initial crack formation on a slip plane; stage 2, crack growth along cleavage planes; stage 3, sudden failure due to rapid crack propagation.

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 to 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.

10.4.2.2 Secondary creep

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

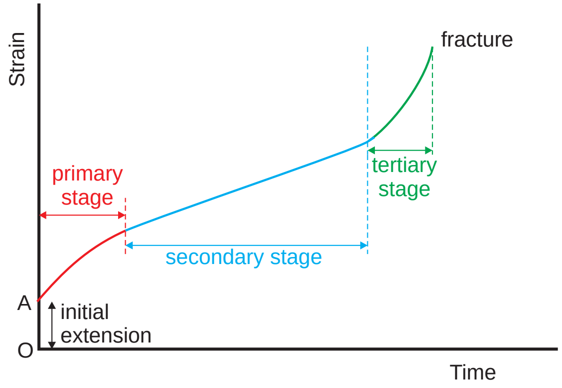


Figure 10.31 A normal creep curve for a solid subjected to a constant load at constant temperature.

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, 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, is 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:

$$K = \frac{d\varepsilon}{dt} = \frac{\pi^2 D\,\sigma^n}{\sqrt{(bN)}\,G^n\,k_B T}$$

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 10.33a). 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:

$$K = \frac{d\varepsilon}{dt} = \frac{A\sigma\,\Omega\,D_b\,\delta}{k_B\,T\,d^3}$$

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).

At relatively higher temperatures, diffusion of atoms within the grains (bulk diffusion) becomes more important in both metals and ceramics. This is called *Herring-Nabarro creep* (Figure 10.33b). In

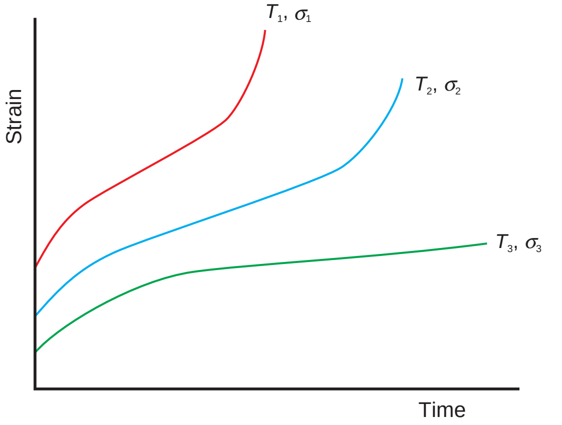


Figure 10.32 The effect of temperature on creep: when the stresses σ_1, σ_2 and σ_3 are equal, the temperature values lie in the sequence $T_1 > T_2 > T_3$; when temperatures T_1, T_2 and T_3 are equal, the stress values lie in the sequence $\sigma_1 > \sigma_2 > \sigma_3$.

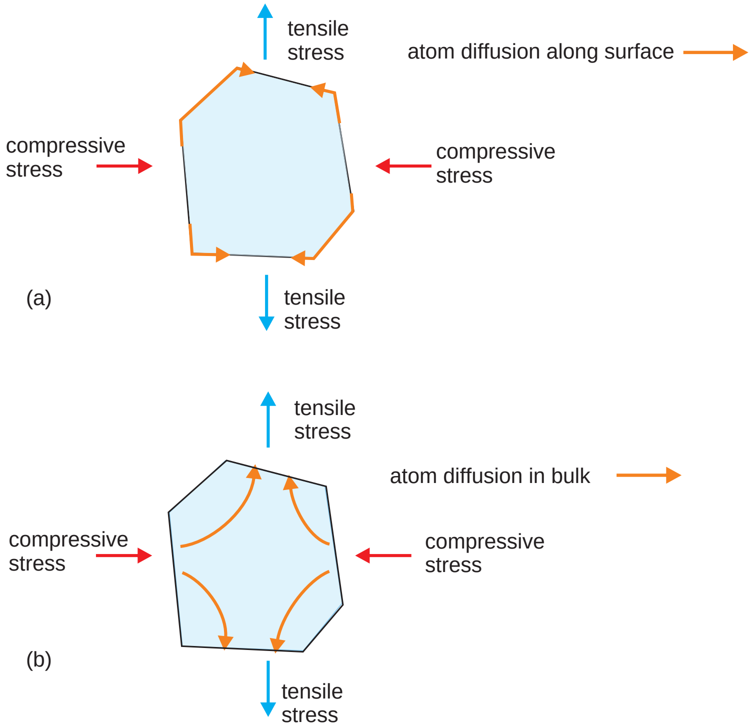


Figure 10.33 Creep mechanisms: (a) Coble creep, in which atoms diffuse along grain boundary surfaces; (b) Herring-Nabarro creep, in which atoms diffuse within grains. Note that vacancy diffusion will be in the opposite directions to that of atom diffusion.

this model, the slope of the secondary creep curve is given by:

$$K = \frac{d\varepsilon}{dt} = \frac{A\sigma\, \Omega\, D}{k_{\mathrm{B}}\, T\, d^2}$$

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 Herring-Nabarro 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 to 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.

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:

$$K = \frac{d\varepsilon}{dt} = A\sigma^n \exp\left(\frac{-E}{RT}\right)$$

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.

10.4.2.3 Tertiary creep

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.

10.5 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.

10.5.1 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 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 (Section 5.3.7, Figure 5.23). The carbon sheets are about 0.335 nm apart, and are linked by weak van der Waals bonds giving a large intersheet spacing called the van der Waals gap (Figure 10.34a). 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 (Section 9.3.2, 9.3.4), 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 10.34b). 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.

Molybdenum disulphide, MoS_2, is another layer structure widely used as a solid lubricant (Figure 10.34c). It is composed of MoS_2 layers

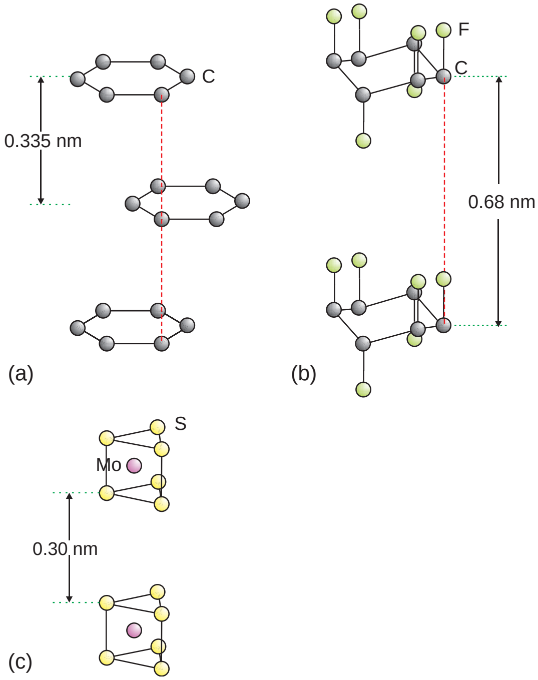


Figure 10.34 The structures of solid lubricants: (a) graphite; (b) graphite fluoride, CF_x; (c) molybdenum disulphide, MoS_2.

weakly linked by van der Waals bonds. The Mo atoms are surrounded by six sulphur 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 disulphide 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). A solid lubricant that is widely used in the temperature range from approximately 540°C to 900°C is the eutectic mixture of calcium fluoride, CaF_2, and barium fluoride, BaF_2. The melting point of the eutectic is 1022°C.

10.5.2 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 counterintuitive mechanical property was first noticed in foam-like structures (Figure 10.35). In these, the links or bonds between elements of structure are of fixed length, but the links between them 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

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 10.36a), the material becomes thinner when stretched (Figure 10.36b). However, if the bonds between the blocks have a re-entrant geometry (Figure 10.36c), a tensile force will cause the blocks to move apart and the material will become fatter (Figure 10.36d). 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.

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.

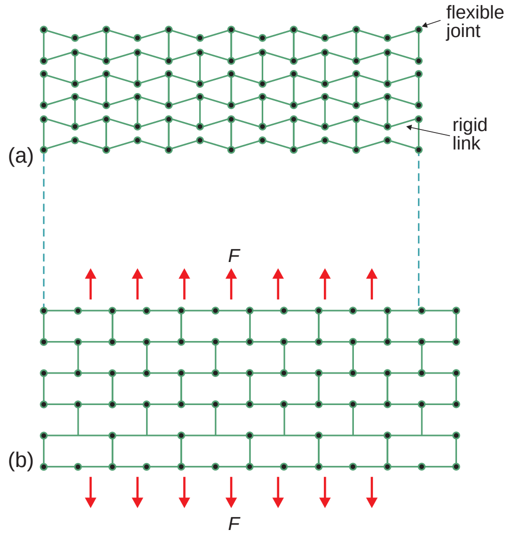


Figure 10.35 An auxetic network in which the bonds between elements of structure (lines) are of fixed length, but the links (circles) are flexible: (a) initial state; (b) under an applied force, *F*.

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 $\langle 111 \rangle$, towards the corners of the unit cell from the central atom. Weaker bonds lying along the cell edges link the copper atoms (Figure 10.37a). 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 10.37b). In this case, the $\langle 111 \rangle$ 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 10.37c).

The arrangement of the atoms perpendicular to (111) is rather different (Figure 10.37d). 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 10.37e).

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.

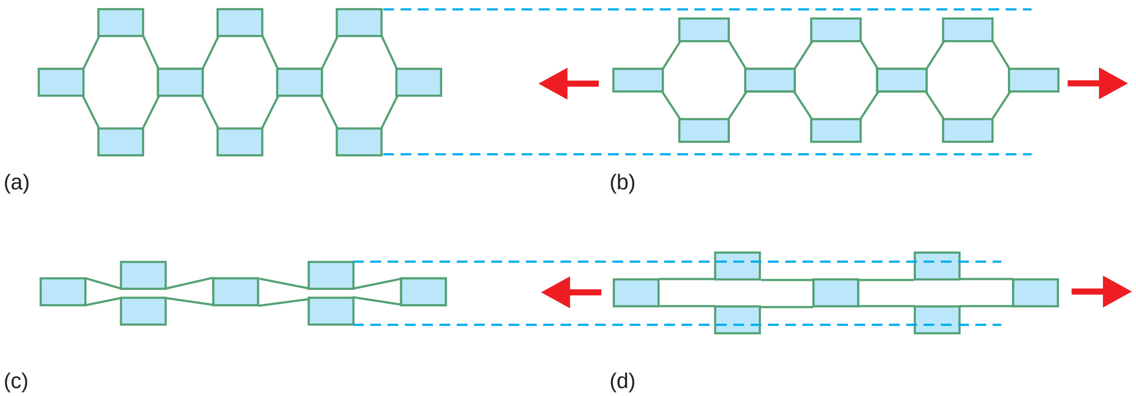


Figure 10.36 A solid composed of rigid blocks of crystal linked by bonds of constant length, but with flexible links to the crystallites: (a, b) a normal solid becomes thinner under tension; (c, d) a material with re-entrant bonds is auxetic, and becomes thicker under tension.

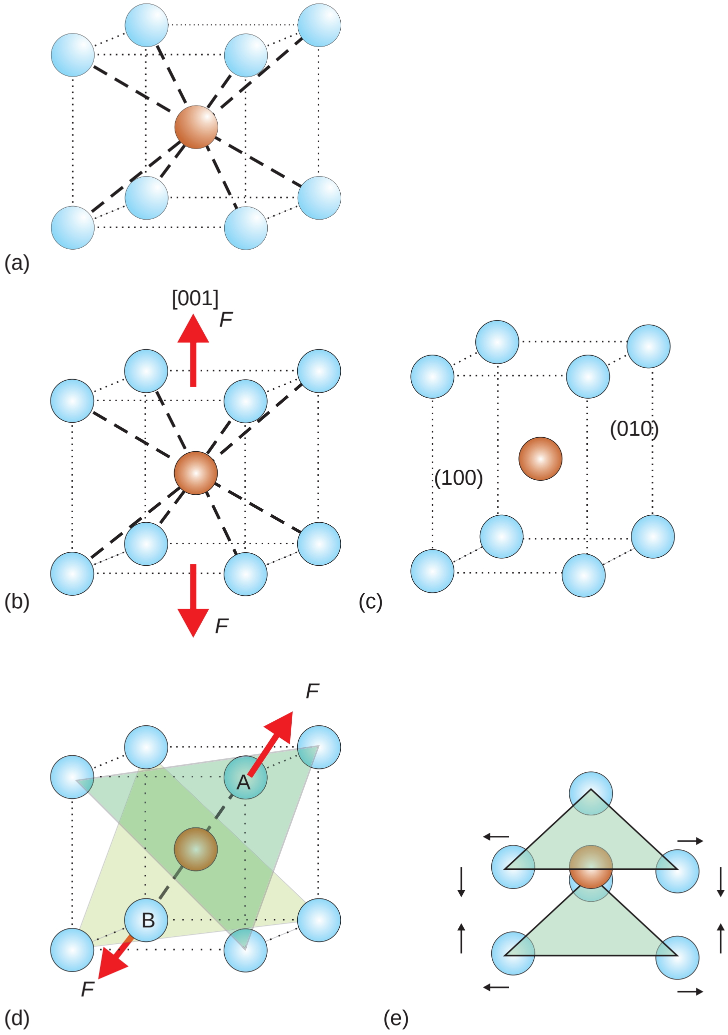


Figure 10.37 Auxetic β-brass: (a) the CsCl structure of β-brass; strong bonds are drawn as heavy broken lines, and weak bonds as light dotted lines; (b, c) the application of a force along [001] results in contraction perpendicular to (100) and (010) faces; (d) the (111) planes; a force along [111] acts directly along strong bonds causing the triangles of zinc atoms above and below the central copper atom to come together slightly (e), which causes the structure to expand perpendicular to [111].

10.5.3 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 (Figure 10.38).

The technique most widely used is the small-scale equivalent to the hardness test, called *nanoindentation*. It has the advantage that the properties

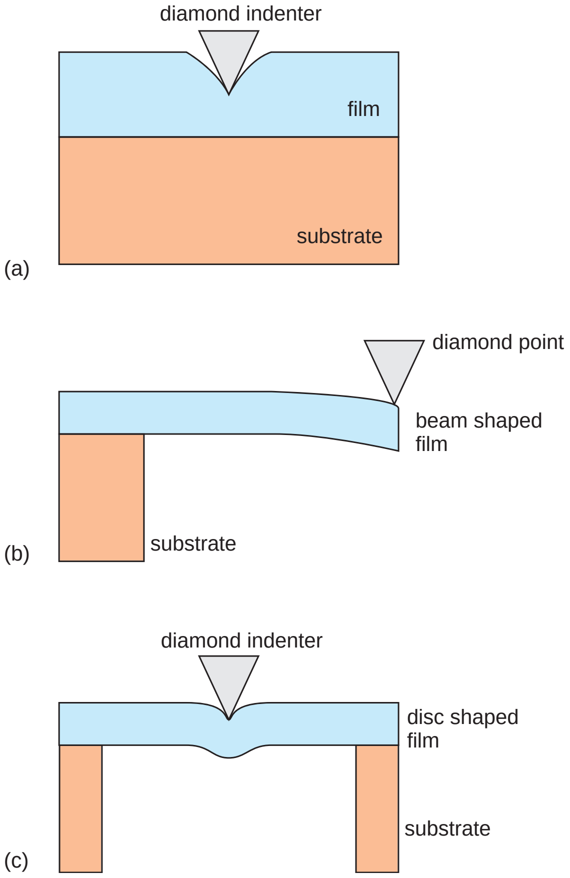


Figure 10.38 Methods of measuring the mechanical properties of thin films: (a) nanoindentation; (b) beam deflection; (c) disc deflection. (Redrawn following Pharr and Oliver [23], see Further Reading.)

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 those employed in large-scale testing (Figure 10.8). A diamond indenter with a triangular pyramidal shape, called a *Berkovich indenter*, is used (Figure 10.39a). 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.

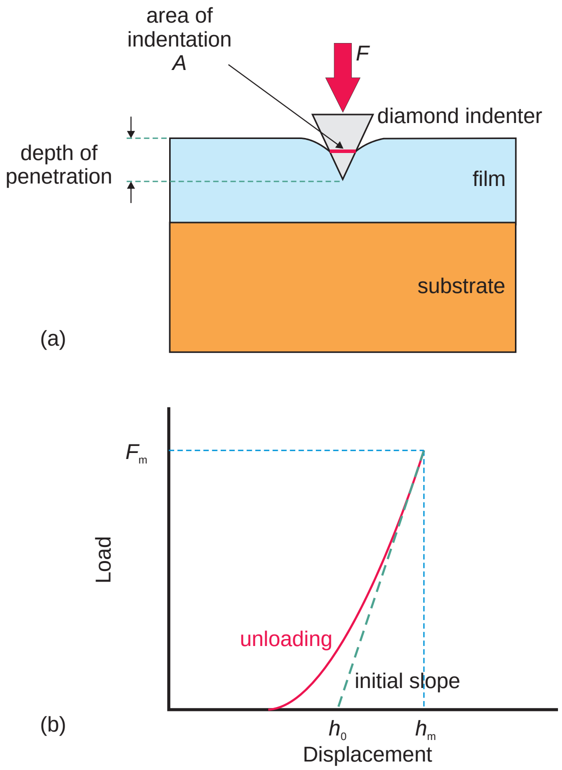


Figure 10.39 Nanoindentation: (a) indentation of a thin film; (b) schematic unloading curve for a nanoindentation experiment. (Redrawn following Pharr and Oliver [23], see Further Reading.)

This means that interpretation of the results is less obvious than for large-scale experiments.

A representative load–displacement curve for unloading shows the maximum load applied, F_m, and the corresponding maximum displacement, h_m (Figure 10.39b). The initial slope of the unloading curve, dF/dh, is a measure of the initial stiffness, S_i, and is given by:

$$S_i = \frac{dF}{dH} = \frac{F_\mathrm{m}}{h_\mathrm{m} - h_0}$$

where h_0 is the extrapolation of the initial slope to $F = \text{zero}$.

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:

$$H = \frac{F_\mathrm{m}}{A}$$

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:

$$S_\mathrm{i} = \frac{2}{\sqrt{\pi}} E_\mathrm{r} \sqrt{A}$$

where A is the area of indentation at maximum load, F_m, and E_r is the *reduced Young's modulus*, given by:

$$\frac{1}{E_\mathrm{r}} = \left(\frac{1 - \nu_\mathrm{fi}^2}{E_\mathrm{fi}}\right) - \left(\frac{1 - \nu_\mathrm{in}^2}{E_\mathrm{in}}\right)$$

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 soft and ductile metal. However, under some circumstances, the nanowires exhibited brittle fracture, and showed no necking or thinning. In these instances (111) twins in the gold were responsible for brittle fracture. These different mechanisms were confirmed by molecular dynamics simulations.

10.6 Composite materials

Composites are solids made of more than one material type, designed to have enhanced properties compared with the separate materials themselves (Section 6.5). The mechanical properties of composite materials are often difficult to obtain, because of the complex microstructures found, especially in biological structures, which may be small and severely affected by the state of dehydration of the sample.

10.6.1 Young's modulus of large particle composites

The Young's (elastic) modulus of a composite containing large particles depends upon the volume fraction of the constituent phases. The modulus falls between an upper and a lower limit.

$$E_\mathrm{c}(\text{upper limit}) = E_\mathrm{m} V_\mathrm{m} + E_\mathrm{p} V_\mathrm{p}$$

$$E_\mathrm{c}(\text{lower limit}) = \frac{E_\mathrm{m} E_\mathrm{p}}{E_\mathrm{m} V_\mathrm{p} + E_\mathrm{p} V_\mathrm{m}}$$

where E_c, E_m and E_p are the Young's moduli of the composite, matrix and particles, 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 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.

10.6.2 Young's 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:

$$l_c = \frac{\sigma_f d}{2\tau_c}$$

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 $l > 15\, l_c$ the material is termed a *continuous fibre* composite.
- If $l < 15\, l_c$ 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.

10.6.2.1 Longitudinal load on a continuous and aligned fibre composite

To obtain a value of the Young's modulus of a composite in which the fibres are both continuous and aligned parallel to each other, 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 10.40a). This is termed the *isostrain 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

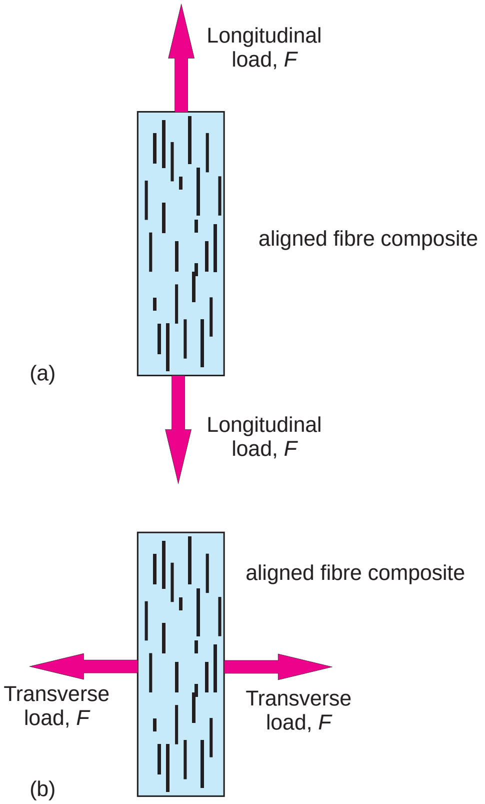


Figure 10.40 Composite materials: (a) a longitudinal load applied to an aligned fibre composite; (b) a transverse load applied to an aligned fibre composite.

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:

$$E_{\mathrm{cl}} = E_{\mathrm{m}} V_{\mathrm{m}} + E_{\mathrm{f}} V_{\mathrm{f}}$$

where E_{cl} is the Young's modulus of the composite under a longitudinal load, E_{m} is the Young's modulus of the matrix, E_{f} is the Young's modulus of the fibre, and V_{f} and V_{m} are the volume fraction, (V_{f} is the volume of fibre/total volume, V_{m} is the volume of matrix/total volume). For two components:

$$V_{\mathrm{f}} = 1 - V_{\mathrm{m}}$$

$$E_{\mathrm{cl}} = E_{\mathrm{m}}(1 - V_{\mathrm{f}}) + E_{\mathrm{f}} V_{\mathrm{f}}$$

This is the same equation as *the upper bound for large particle composites*. The fraction of the load carried by each component is given by:

$$\frac{F_{\mathrm{f}}}{F_{\mathrm{m}}} = \frac{E_{\mathrm{f}} V_{\mathrm{f}}}{E_{\mathrm{m}} V_{\mathrm{m}}}$$

where F_{f} is the load carried by the fibre and F_{m} is the load carried by the matrix.

10.6.2.2 Transverse load on a continuous and aligned fibre composite

When a load is applied in a transverse direction to a composite that contains continuous and aligned fibres (Figure 10.40b), the iso-strain condition is unreasonable, and it is more appropriate to assume an *iso-stress condition* to apply:

$$\sigma_{\mathrm{c}} = \sigma_{\mathrm{m}} = \sigma_{\mathrm{f}} = \sigma$$

where σ represents the stress and the subscripts have the same meaning as above. In this case:

$$E_{\mathrm{ct}} = \frac{E_{\mathrm{m}} E_{\mathrm{f}}}{E_{\mathrm{f}} V_{\mathrm{m}} + E_{\mathrm{m}} V_{\mathrm{f}}} = \frac{E_{\mathrm{m}} E_{\mathrm{f}}}{(1 - V_{\mathrm{f}}) E_{\mathrm{f}} + E_{\mathrm{m}} V_{\mathrm{f}}}$$

where E_{ct} is the Young's modulus in a transverse direction for the composite. This is the same equation as the *lower bound for large particle composites.*

10.6.3 Young's modulus of a two-phase system

Many ordinary solids, such as ceramics, are made up of several phases. Strictly speaking, these are not composite materials, but similar reasoning can be applied to obtain the Young's modulus and other mechanical properties of such systems. Although the equations for a solid composed of several phases with a complex microstructure are frequently unwieldy, simpler equations exist for well-defined geometries.

A ceramic body composed of two phases, one of which is distributed as particles within the matrix of the other, has a Young's modulus given by:

$$E_{\mathrm{c}}(\text{upper limit}) = E_{\mathrm{m}} V_{\mathrm{m}} + E_{\mathrm{p}} V_{\mathrm{p}}$$

$$E_{\mathrm{c}}(\text{lower limit}) = \frac{E_{\mathrm{m}} E_{\mathrm{p}}}{E_{\mathrm{m}} V_{\mathrm{p}} + E_{\mathrm{p}} V_{\mathrm{m}}}$$

where E_{c}, E_{m} and E_{p} are the moduli of the ceramic, matrix and particles, respectively, and V_{m} and V_{p} are the corresponding volume fractions. These equations are identical to those for large particle composites given above.

A ceramic body composed of layers aligned parallel to a uniaxial stress, in which the strain is shared equally by the two phases, the iso-strain condition (the *Voigt model*), has a Young's modulus:

$$E_{\mathrm{c}}(\text{upper limit}) = E_{\mathrm{m}} V_{\mathrm{m}} + E_{\mathrm{s}} V_{\mathrm{s}}$$

The value of Young's modulus is often called the *Voigt bound*, is identical to that for a continuous aligned fibre composite under a longitudinal load, and gives Young's modulus when the load is applied parallel to the sheets. Similarly, if the stress is applied perpendicular to the layers, and an iso-stress condition applies (the *Reuss model*), the Young's modulus is:

$$E_c(\text{lower limit}) = \frac{E_m E_s}{E_m V_s + E_s V_m}$$

The value of Young's modulus, often called the *Reuss bound*, is identical to that for transverse loading on a fibre composite, and gives a value for Young's modulus normal to the layers. In both of these equations, E_c, E_m and E_s are the moduli of the ceramic, matrix and sheets, respectively, and V_m and V_s are the corresponding volume fractions. Recent studies on limpet teeth show that they behave as fibre-reinforced ceramic composites.

The presence of pores in ceramics usually leads to weakness. The Young's modulus of a body with a Poisson's ratio of 0.3, containing isolated closed pores, is described by the equation:

$$E_c = E_0\left(1 - 1.9V_p + 0.9V_p^2\right)$$

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*, 3rd edn. Taylor and Francis, Boca Raton, FL.

Bryant, L.C. and Bewlay, B.P. (1995) The Coolidge process for making ductile tungsten. *Materials Research Society Bulletin*, **XX** (August): 67.

Eberhart, M.E. (1999) Why things break. *Scientific American*, **281** (October): 44.

Freiman, S. and Mecholsky, J.J., Jr. (2012) *Fracture of Brittle Materials: Testing and Analysis*. John Wiley and Sons, Ltd., Hoboken.

Griffith, A.A. (1920) *Phil. Trans. Royal. Soc. Lond.*, **A221**: 163–98. (This is the original paper describing the Griffith's theory of brittle fracture).

Hecker, S.S. and Ghosh, A.K. (1976) The forming of sheet metal. *Scientific American*, **235** (November): 100.

Lawn, B. (1993) *Fracture of Brittle Solids*, 2nd edn. Cambridge University Press, Cambridge. (This is now a classical book about fracture.)

McQueen, H.J. and Tegart, W.J.M. (1975) The deformation of metals at high temperatures. *Scientific American*, **232** (April): 116.

Semiatin, S.L. and Lahoti, G.D. (1981) The forging of metals. *Scientific American*, **245** (August): 82.

Sun, C.T. and Lin, Z.-H. (2012) *Fracture Mechanics*. Elsevier Academic Press, Oxford.

General computational methods:

Teter, D.M. (1998) Computational alchemy, the search for new superhard materials. *Materials Research Society Bulletin*, **23**: 22.

Atomistic simulations:

Buehler, M.J. (2008) *Atomistic Modelling of Materials Failure: Deformation and Fracture of Brittle*, *Ductile and Nanoscale Materials*. Springer Science+Business Media, New York.

Buehler, M.J. and Gao, H. (2006) Dynamical fracture instabilities due to local hyperelasticity at crack tips. *Nature*, **439**: 307–10.

Holland, D. and Marder, M. (1998) Ideal brittle fracture of silicon studied with molecular dynamics. *Phys. Rev. Lett.*, **80**: 746–9.

Vatne, I.R., Ǿstby, E., Thaulow, C. and Farkas, D. (2011) Quasicontinuum simulation of crack propagation in bcc-Fe. *Mat. Sci. Eng.*, **A528**: 5122–34.

Density functional methods:

Huag, Z., Feng, J. and Pan, W. (2012) Theoretical investigations of zircon-type YVO_4. *J. Solid State Chem.*, **185**: 42–8, and references therein.

Finite element methods:

Knipprath, C., Bond, I.P. and Trask, R.S. (2012) Biologically inspired crack delocalisation in a high strain-rate environment. *J. Roy. Soc. Interface*, **9**: 665–76.

Panagiotopoulou, O., Wilshin, S.D., Rayfield, E.J., Shefelbine, S.J. and Hutchinson, J.R. (2012) What makes an accurate and reliable subject-specific finite element model? A case study of an elephant femur. *J. Roy. Soc. Interface*, **9**: 351–61.

Zienkiewicz, O.C., Taylor, R.L. and Zhou, J.Z. (2005) *The Finite Element Method: Its Basis and Fundamentals, 6th edition*. Elsevier Butterworth Heinemann, Amsterdam.

Nanoscale methods:

Xu, B., *et al.* (1999) Making negative Poisson's ratio microstructures by soft lithography. *Adv. Mats.*, **11**: 1186–9.

Mitschke, H., *et al.* (2011) Finding auxetic frameworks in periodic tessellations. *Adv. Mats.*, **23**: 2669–74.

Lu, Y., Song, J., Huang, J.Y. and Lou, J. (2011) Fracture of sub-20nm ultrathin gold nanowires. *Adv. Funct. Mats.*, **21**: 3982–7.

Pharr, G.M. and Oliver, W.C. (1992) Measurement of thin film mechanical properties using nanoindentation. *Materials Research Society Bulletin*, **XV11** (July): 28.

Composites:

Lu, D. and Barber, A.H. (2012) Optimal nanoscale composite behaviour in limpet teeth. *J. Roy. Soc. Interface*, **9**: 1318–24.

Problems and exercises

Quick quiz

1 A material under a tensile force is:
(a) Stretched.
(b) Twisted.
(c) Compressed.

2 Torsional forces occur in a material that is:
(a) Sheared.
(b) Twisted.
(c) Compressed.

3 Both tensile and compressive forces occur in a rod that is:
(a) Bent.
(b) Twisted.
(c) Stretched.

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.

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.

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.

7 The initial (linear) part of an engineering stress–engineering strain curve represents:
(a) Plastic deformation
(b) The tensile strength.
(c) Elastic deformation.

8 In comparison to 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.

9 Permanent deformation of a solid when all stress is removed is a sign of:
(a) Plastic deformation.
(b) Elastic deformation.
(c) Tensile deformation.

10 The defects mainly held responsible for plastic deformation are:
(a) Point defects.
(b) Precipitates.
(c) Dislocations.

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.

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.

13 The specific strength of a material is:
(a) The tensile strength/specific gravity.
(b) The tensile strength/volume.
(c) The tensile strength/weight.

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.

15 The lateral dimensional change that accompanies a longitudinal strain is:
(a) Poisson's modulus.
(b) Poisson's ratio.
(c) Poisson's strain.

16 Fracture of a polycrystalline solid that takes place between the crystallites is:
(a) Intergranular fracture.
(b) Transgranular fracture.
(c) Cleavage.

17 Conchoidal fracture is displayed by brittle solids that are:
(a) Polycrystalline.
(b) Amorphous.
(c) Natural (biomaterials).

18 The Griffith theory of brittle fracture postulates that the fracture is due to:
(a) Dislocations.
(b) Precipitates.
(c) Small cracks.

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.

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.

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.

22 Ceramics can be strengthened by:
(a) Immobilising dislocations.
(b) Removing Griffith flaws.
(c) Increasing the degree of crystallinity.

23 One of the following does *not* describe hardness:
(a) Knoop.
(b) Brinell.
(c) Poise.

24 Materials that fail after repeated cycles of stress are said to suffer:
(a) Fatigue failure.
(b) Creep failure.
(c) Wear failure.

25 Fatigue failure does not occur at stress levels below:
(a) The persistence limit.
(b) The fatigue limit.
(c) The endurance limit.

26 The gradual elongation of a material under a constant load is called:
(a) Fatigue.
(b) Creep.
(c) Yield.

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.

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.

29 The temperature dependence of creep is often described by:
(a) An Arrhenius law.
(b) A linear law.
(c) A parabolic rate law.

30 Solid lubricants often have:
(a) Layer structures.
(b) Amorphous structures.
(c) Liquid crystal structures.

31 Materials that expand when under tension are called:
(a) Nanomaterials.
(b) Eutactic materials.
(c) Auxetic materials.

32 The hardness of thin films is measured using:
(a) A Vickers indenter.
(b) A Knoop indenter.
(c) A Berkovich indenter.

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

10.1 A weight of 500 kg is hung from a 2 cm diameter rod of brass. What is the engineering stress?

10.2 A weight of 3500 kg is hung from a 1.5 cm diameter rod of nickel. What is the engineering stress?

10.3 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.

10.4 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.

10.5 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.

10.6 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.

10.7 A cast iron rod of length 200 mm and dimensions 10 × 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.

10.8 A zinc bar of length 125 mm and dimensions 5 × 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.

10.9 Calculate Poisson's ratio for a bar of metal originally 10 × 10 × 100 mm, which is extended to 101 mm, if there is no change in the overall volume of the sample.

10.10 A copper bar of 10 mm square section is subjected to a tensile load that increases its length from 100 mm to 102 mm. The value of Poisson's ratio for copper is 0.343. Calculate the new dimensions of the bar.

10.11 A brass rod of 12.5 mm diameter is subjected to a tensile load that increases its length from 150 mm to 151.5 mm. The value of Poisson's ratio for brass is 0.350. Calculate the new diameter.

10.12 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.

10.13 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, and (b) the change in diameter of the rod.

10.14 A niobium bar of dimensions 15 mm square 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.

10.15 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.

10.16 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 and (b) the percentage elongation.

10.17 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 and (b) the percentage elongation.

10.18 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 table. The maximum load applied was 810 N and the final length between the gauge marks was 54 mm.

(a) Plot an engineering stress versus engineering strain curve.

(b) Determine Young's modulus of the alloy.

(c) Determine the tensile strength.

(d) Determine the 0.2% offset yield strength of the alloy.

(e) Determine 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

10.19 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 versus engineering strain curve.

(b) Determine Young's modulus of the alloy.

(c) Determine the tensile strength.

(d) Determine the 0.1% offset yield stress of the alloy.

(e) Determine 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

10.20 Figure 10.41 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.

10.21 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

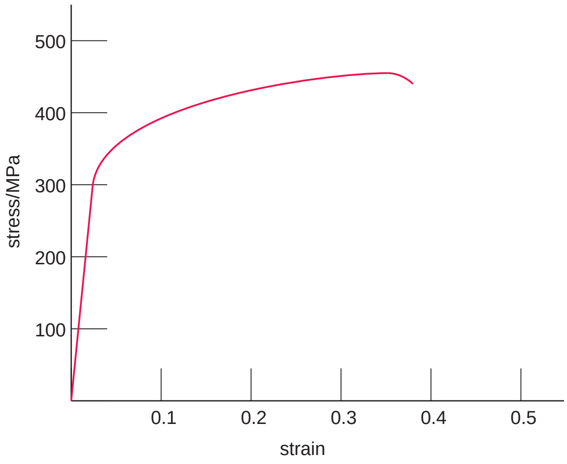


Figure 10.41 The engineering stress–engineering strain curve of a carbon steel.

10.22 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 × 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?

10.23 Using the Griffith criterion, [$\sigma_c = (2\gamma E/\pi a)^{½}$], 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 Jm^{-2}.

10.24 A glass plate has to withstand a stress of $10^8\,\mathrm{Nm}^{-2}$. Using the data in the previous question, what will be the critical crack size for this to be achieved?

10.25 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 $6 \times 10^6\,\mathrm{Nm}^{-2}$ 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.

10.26 Determine (a) the upper bound and (b) 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.

10.27 An aluminium alloy is to be strengthened by the incorporation of beryllium oxide (BeO) particles. Calculate (a) the upper and (b) 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 $2698\,\mathrm{kg\,m}^{-3}$. Young's modulus of BeO is 301.3 GPa and its density is $3010\,\mathrm{kg\,m}^{-3}$.

10.28 Compute Young's modulus of a composite consisting of continuous and aligned glass fibres of 50% volume fraction in an epoxy resin matrix under (a) longitudinal and (b) transverse loading. Young's modulus of the glass fibres is 76 GPa and that of the resin is 3 GPa.

10.29 Compute Young's modulus of a composite consisting of continuous and aligned carbon fibres of 60% weight fraction in an epoxy resin matrix under (a) longitudinal and (b) transverse loading. Young's modulus of the carbon fibres is 290 GPa and the density is $1785\,\mathrm{kg\,m}^{-3}$. Young's modulus of the resin is 3.2 GPa and its density is $1350\,\mathrm{kg\,m}^{-3}$.

10.30 Determine (a) the Voigt and (b) 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.

10.31 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 (a) parallel and (b) perpendicular to the layers. Young's modulus and density of MgO are 210.3 GPa and $3580\,\mathrm{kg\,m}^{-3}$, and for silica are 72.4 GPa and $2650\,\mathrm{kg\,m}^{-3}$.

10.32 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.

10.33 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?

11

Insulating solids

- What is a dielectric?
- Why can quartz be used for telling the time?
- What is 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 11.1).

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 historical division is useful in broad-brush terms and is retained here. In this chapter the insulators are described. In Chapter 13 those materials that are conductors of electricity are discussed.

11.1 Dielectrics

11.1.1 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*. 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 11.2). In a vacuum (or air, in practice), we find:

$$q = \frac{\varepsilon_0 AV}{d}$$

where q is the charge on the capacitor, ε_o is a constant, the *permittivity of free space*, A is the area of the plates and d their separation. The capacitance, c, of

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.
© 2013 John Wiley & Sons, Ltd. Published 2013 by John Wiley & Sons, Ltd.

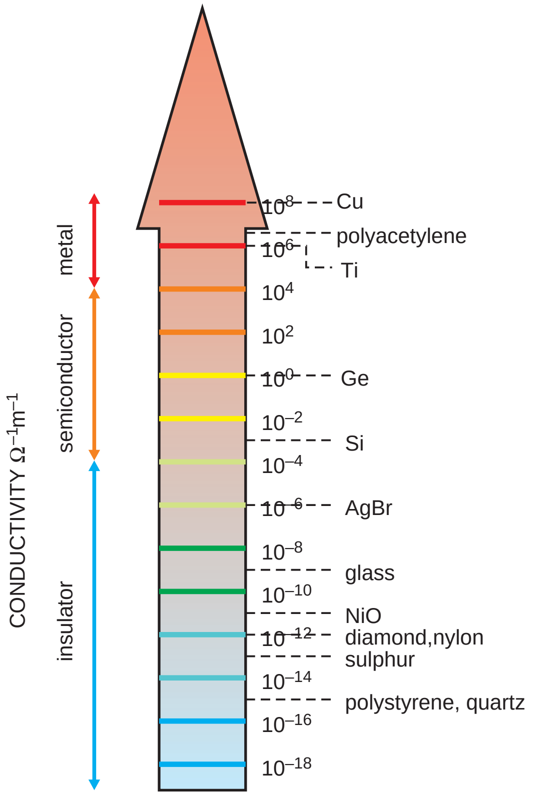


Figure 11.1 The range of electronic conductivity in solids.

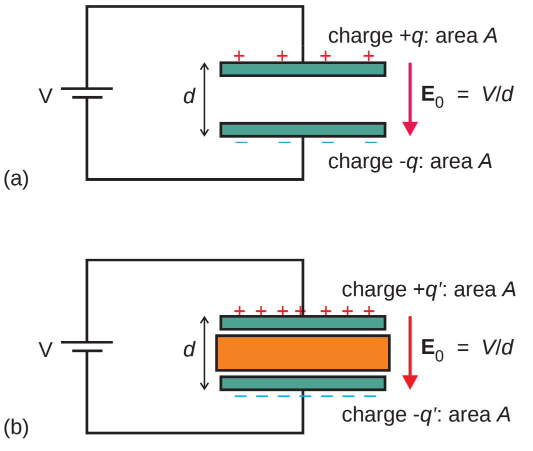


Figure 11.2 Capacitors: (a) charges will accumulate on metal plates due to an applied voltage, *V*; (b) a slab of dielectric (insulator) inserted between the plates will cause the charges on the plates to change.

the arrangement is equal to the ratio of the charge on either of the metallic foils, *q*, to the potential difference, *V*, between them:

$$c = \frac{q}{V} = \frac{\varepsilon_0 A}{d}$$

If the region between the plates is now filled with a dielectric the charge and the capacitance increase by an amount ε_r:

$$q' = \frac{\varepsilon_0\, \varepsilon_r A V}{d}$$

$$c' = \frac{\varepsilon_0\, \varepsilon_r A}{d}$$

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 11.1.2). 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 11.3). The polarisation of the dielectric, **P**, is a vector quantity, defined as the electric dipole moment per unit

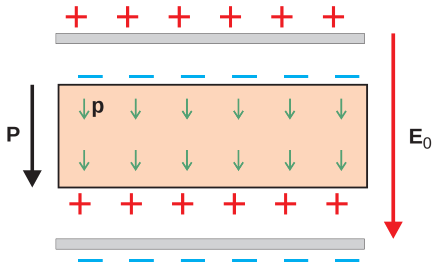


Figure 11.3 In an applied electric field, $\mathbf{E}_0$, an insulator gains a surface charge due to the formation of internal dipoles, $\mathbf{p}$, that induce an observable polarisation, $\mathbf{P}$.

volume. The polarisation vector points from the negative surface to the positive surface and the vectors $\mathbf{P}$ and $\mathbf{p}$ are parallel.

In the case of a solid that is isotropic (uniform in all directions), for example, a glass or cubic crystal, it is found that at ordinary field strengths $\mathbf{P}$ is proportional and parallel to the applied electric field, $\mathbf{E_0}$, and:

$$\mathbf{P} = \varepsilon_0\,\chi\,\mathbf{E}_0 \tag{11.1}$$

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 with 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 *electric susceptibility* is related to the relative permittivity by the equation:

$$\begin{aligned} \chi &= (\varepsilon_r - 1) \\ \mathbf{P} &= (\varepsilon_r - 1)\varepsilon_0\,\mathbf{E}_0 \end{aligned} \tag{11.2}$$

For ordinary electric field strengths, the electric dipole moment $\mathbf{p}$, 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, $\mathbf{P}$, is:

$$P = \sum_j N_j p_j = \sum_j N_j \alpha_j E_{\text{loc}}$$

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 whilst 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 11.3.10), which may show high deformation. (Note that electrostriction is a property of all dielectrics and is quite different from piezoelectricity, Section 11.2.)

11.1.2 Polarisability

The polarisation of an insulating solid is derived from the atomic constituents that make up the

[1] The SI units of α are C m^2/V. The most commonly quoted units for polarisability in the literature are non-SI polarisability volumes, α', m^3, cm^3 or Å^3. To convert these units, note that: α (C m^2/V) = 1.11265×10^{-10} α' (m^3) = 1.11265×10^{-16} α' (cm^3) = 1.11265×10^{-30} α' (Å^3).

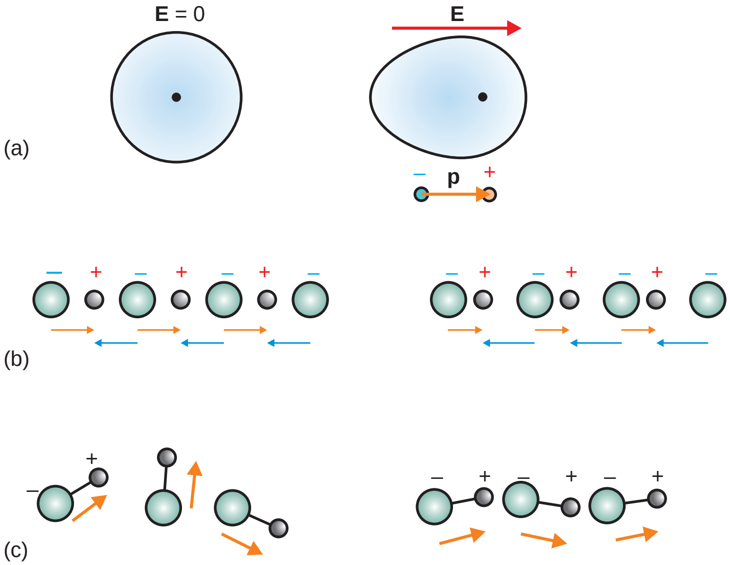


Figure 11.4 Polarisation: (a) electronic; (b) ionic; (c) orientational. Dipoles are shown as arrows.

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 11.4a), 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 11.4b). 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 11.4c) 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 and the electrode regions 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:

$$\alpha_t = \alpha_e + \alpha_i + \alpha_d + \alpha_s + \ldots$$

11.1.3 Polarisability and relative permittivity

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, $\mathbf{E}_{\text{loc}}$, is not the same as the applied field, $\mathbf{E}_0$, but will also include contributions from internal fields $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$, and so on, arising from the induced and permanent dipoles in the structure.

$$\mathbf{E}_{\text{loc}} = \mathbf{E}_0 + \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 + \ldots$$

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:

$$\mathbf{E}_{\text{loc}} = \mathbf{E}_0 + \frac{\mathbf{P}}{3\varepsilon_0} \qquad (11.3)$$

Using the three equations (11.1), (11.2) and (11.3), it is possible to derive the most widely used relationship between relative permittivity and polarisability, the *Clausius-Mossotti relation*, equation (11.4), usually written:

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{N\alpha}{3\,\varepsilon_0} \qquad (11.4)$$

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:

$$\alpha = 3\,\varepsilon_0 V_m \frac{\varepsilon_r - 1}{\varepsilon_r + 2}$$

Because the polarisability volume, α', is often used rather than the SI polarisability, α, a commonly encountered form of the Clausius-Mossotti equation is:

$$\alpha' = \frac{3\,V_m}{4\pi}\left(\frac{\varepsilon_r - 1}{\varepsilon_r + 2}\right)$$

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 equation (11.4) by M/ρ, where M is the molar mass of the material and ρ is its density, to obtain:

$$\frac{(\varepsilon_r - 1)M}{(\varepsilon_r + 2)\rho} = \frac{N_A\alpha}{3\,\varepsilon_0} = P_m$$

where N_A is the Avogadro constant, given by MN/ρ.

11.1.4 The frequency dependence of polarisability and relative permittivity

As the total polarisability of a material is made up of 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 both 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 11.5).

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 is best described by using

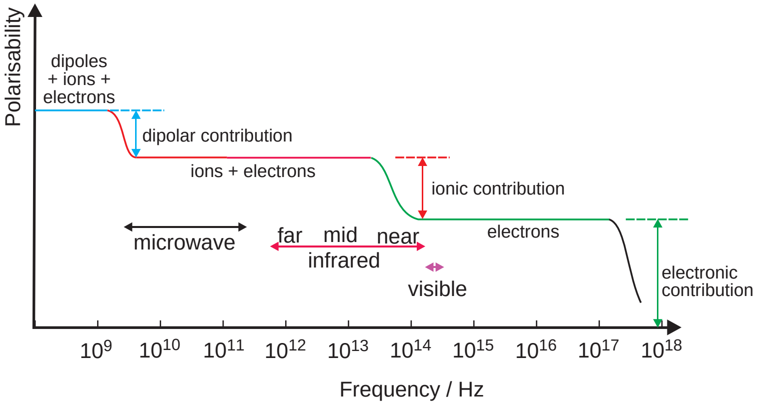


Figure 11.5 The contribution of dipole orientation, ions and electrons to the overall polarisability of a solid, schematic.

the complex dielectric constant:

$$\varepsilon_r = \varepsilon' - i\varepsilon''$$

The loss tangent, tan δ, then specifies the phase lag, where:

$$\tan \delta = \frac{\varepsilon''}{\varepsilon'}$$

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:

$$n^2 = \varepsilon_r \qquad (11.5)$$

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 11.1). With this in mind, substitution of the relationship given in equation (11.5) into the Clausius-Mossotti equation yields the *Lorentz–Lorenz* equation:

$$\frac{n^2 - 1}{n^2 + 2} = \frac{N\alpha_e}{3\varepsilon_0}$$

11.1.5 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 it is usual to define a set of orthogonal axes in the crystal and refer the applied electric field to these axes: conveniently taken to coincide with the crystallographic axes for tetragonal and orthorhombic systems. In hexagonal systems one axis is taken to coincide with the crystallographic **c**-axis and the other two are normal to the **c**-axis. In monoclinic and triclinic crystals it is still possible to define three Cartesian axes, although the relationship between these and the crystallographic axes is not simple. The relative permittivity (and the refractive index) is then quoted as three values, corresponding to the polarisations projected onto the axes (Table 11.1).

The relative permittivity of polycrystalline samples does not reflect this because the crystallographic

Table 11.1 The relative permittivity and refractive index of some crystals

Compound	Symmetry	Relative permittivity	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	$\sim 5 \times 10^5$	—	—	—
Corundum Al_2O_3	Hexagonal	perp. c, 9.34	10^2–10^9	perp. c, 1.761	3.101
		along c, 11.54		along c, 1.769	3.129
Beryl $Be_3Al_2Si_6O_{18}$	Hexagonal	perp. c, 6.86	10^3	perp. c, 1.589	2.525
		along c, 5.95		along c, 1.582	2.503
Rutile TiO_2	Tetragonal	along a, b, 86	10^4–10^6	along a, b, 2.609	6.807
		along c, 170		along c, 2.900	8.410

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 (Section 14.4.2), 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:

$$\alpha\,(\text{compound}) = \sum \alpha\,(\text{components})$$

For example, for an oxide mineral:

$$\alpha\,(\text{mineral}) = \sum \alpha\,(\text{component oxides})$$

$$\alpha\,(Mg_2SiO_4) = 2\alpha\,(MgO) + \alpha\,(SiO_2)$$

A more extended form of the additivity rule is obtained if the 'components' are actual ions or atoms. In the example above we would then write:

$$\alpha(Mg_2SiO_4) = 2\alpha(Mg^{2+}) + \alpha(Si^{4+}) + 4\alpha(O^{2-})$$

11.2 Piezoelectrics, pyroelectrics and ferroelectrics

11.2.1 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 in the absence of both electric fields and mechanical stresses. 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 pyroelectric effect will be seen. 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 11.6):

- If polarisation, $\mathbf{P}$, changes with applied electric field, $\mathbf{E}$, we have a *dielectric*.

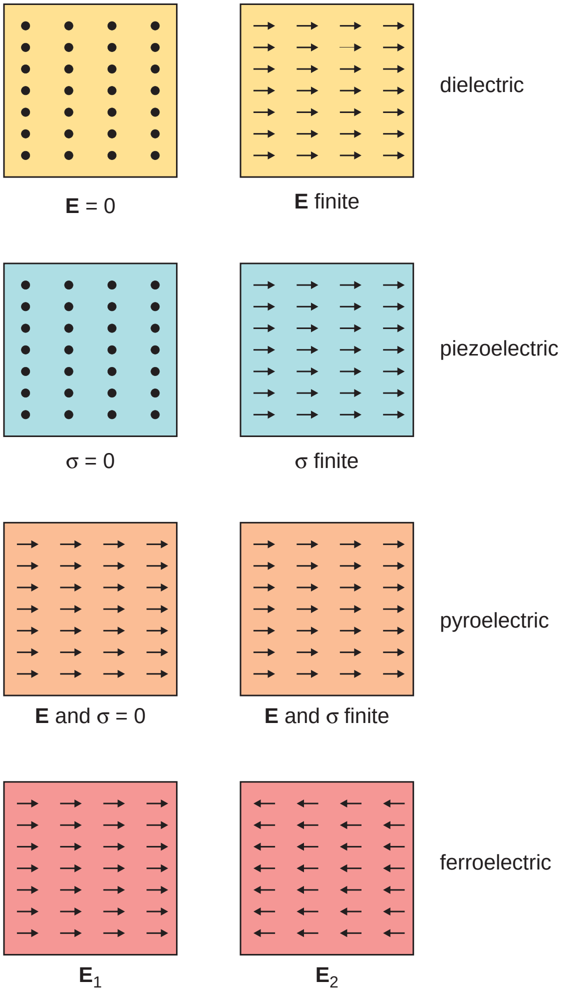


Figure 11.6 Schematic relationship between insulating solids: $\mathbf{E}$ is an applied electric field, and σ is an applied stress.

- In some dielectrics the polarisation, $\mathbf{P}$, 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, $\mathbf{E}$, 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*.

11.2.2 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, suppose that a dipole arises in a distorted tetrahedron of anions surrounding a central cation. The observed effect will then 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 (Section 5.1.3). 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 a centre of symmetry (Figure 11.7).

The relationship between the measured polarisation **P** and the stress applied, σ, can to a *first approximation* be 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 (Section 11.3.8).

A *pyroelectric* crystal possesses a spontaneous polarisation, $\mathbf{P}_s$, at all times, even in the absence of an electric field. The polarisation is due to elementary dipoles in the crystal that are aligned to give an observable external bulk polarisation (Figure 11.6). 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

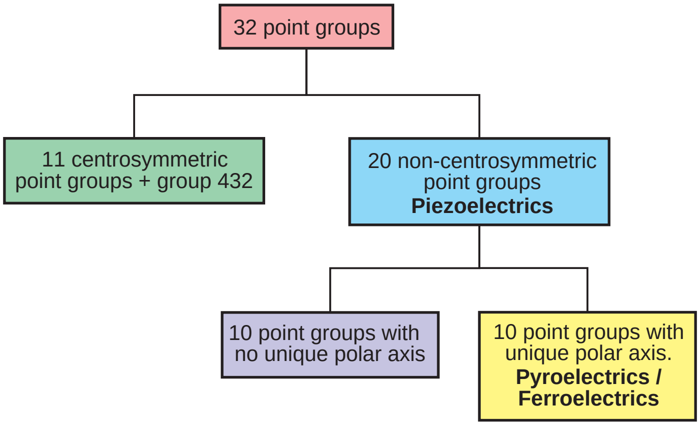

32 point groups
11 centrosymmetric point groups + group 432
20 non-centrosymmetric point groups **Piezoelectrics**
10 point groups with no unique polar axis
10 point groups with unique polar axis. **Pyroelectrics / Ferroelectrics**

Figure 11.7 The relationship between point group symmetry and piezoelectric and pyroelectric properties.

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 11.7).

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} $\mathrm{C\,m^{-2}\,K^{-1}}$. This is simpler than the corresponding situation for a piezoelectric because the polarisation can only lie along the unique polar axis, with the one exception noted.

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 11.8). The tetrahedra are slightly flattened, which gives rise to electric dipoles, all of which lie parallel to the **c**-axis. Materials showing the pyroelectric effect are used as infrared radiation detectors.

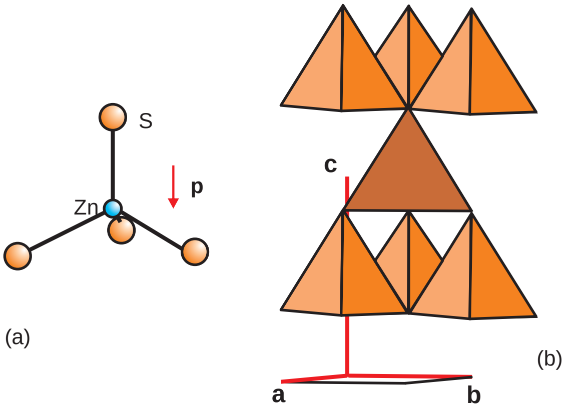


Figure 11.8 The structure of hexagonal ZnS (wurtzite): (a) a ZnS_4 tetrahedron, with an electric dipole, **p**, parallel to the **c**-axis; (b) the stacking of ZnS_4 tetrahedra in the unit cell. ZnO (zincite) is isostructural.

11.2.3 Piezoelectric mechanisms

A piezoelectric crystal may develop surface charges upon 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 11.9a). 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 11.9b) will deform all the bonds equally, and will not give rise to any dipoles. When applied to a crystal an observable piezoelectric

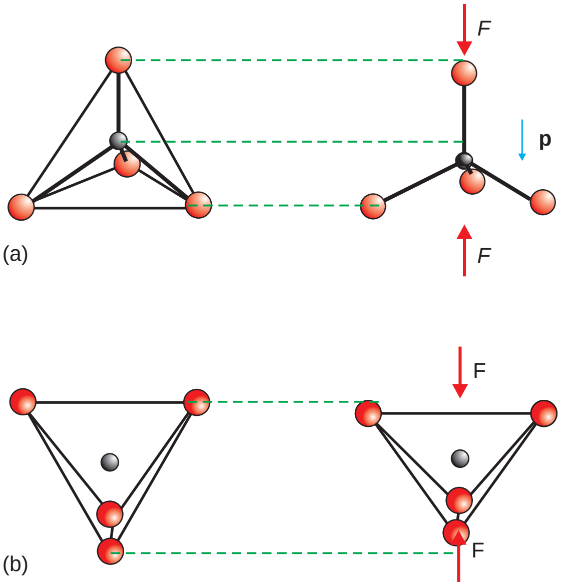


Figure 11.9 Piezoelectricity: (a) a force applied to a tetrahedron along a bond gives rise to a dipole due to distortion; (b) a force applied perpendicular to a tetrahedron edge does not.

effect will only be observed if the symmetry of the unit cell is suitable.

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 11.10a). 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 11.10b).

11.2.4 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 11.11a). The degree of polarisation produced depends upon the direction of the stress with respect to the crystallographic structure. For example, an electric field applied along X will cause

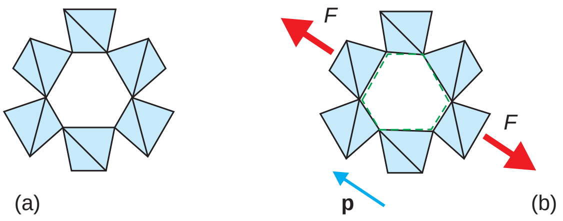


Figure 11.10 Piezoelectricity in quartz: (a) part of the structure of room-temperature α-quartz, projected down the **c**-axis; (b) application of a force to the structure creates a distortion (dotted lines) so that internal dipoles no longer cancel. (Note that the tetrahedra are arranged in a helix, not in rings.)

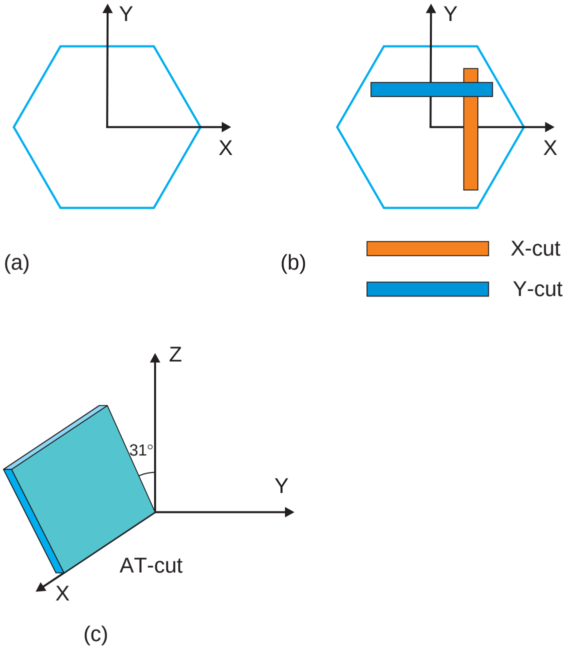


Figure 11.11 Quartz crystals: (a) section though a crystal normal to the crystallographic **c**-axis; (b) X-cut and Y-cut crystal plates; (c) AT-cut crystal plate. Note that X and Y are not crystallographic axes, but Z is parallel to the crystallographic **c**-axis.

the crystal to elongate or contract along Y. Similarly a stress along Y will generate a polarisation along X. Similar relationships apply to other directions in the crystal.

In their simplest form, oscillators consist of a slice of quartz with metal electrodes plated onto opposite faces. Utilizing the inverse piezoelectric effect, the application of a varying voltage to the electrodes will cause the quartz to expand and contract, so 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 afforded a means of measuring time superior to the best mechanical clocks, and so 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 1 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* is a slice with faces parallel to Y and Z and normal to X, with a temperature coefficient of frequency variation of -22×10^{-6} per °C. A *Y-cut* is a slice with faces parallel to X and Z, perpendicular to Y, with a temperature coefficient of frequency variation of $+90 \times 10^{-6}$ per °C (Figure 11.11b). In making a temperature-stable oscillator, the aim is to make a cut that 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 11.11c). Other thermally stable cuts are now known and are used for certain applications.

11.2.5 Piezoelectric polymers

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, strongly polar bonds such as carbon–fluorine, carbon–chlorine, and carbon–nitrogen, and hydrogen bonds. Polar carbon–fluorine bonds are found in polymers such as poly(vinyl fluoride) $[CH_2–CHF]_n$, known as PVF (Figure 11.12a), and poly(vinylidene fluoride) $[CH_2–CF_2]_n$, known as PVF_2 (Figure 11.12b). The negative end of the dipole is located on the fluorine atom ($C \leftarrow F$), and a smaller dipole is found on the carbon–hydrogen bond ($C \rightarrow H$). In PVF, the overall dipole moment is greatest in the isotactic form of the polymer, in which all of the

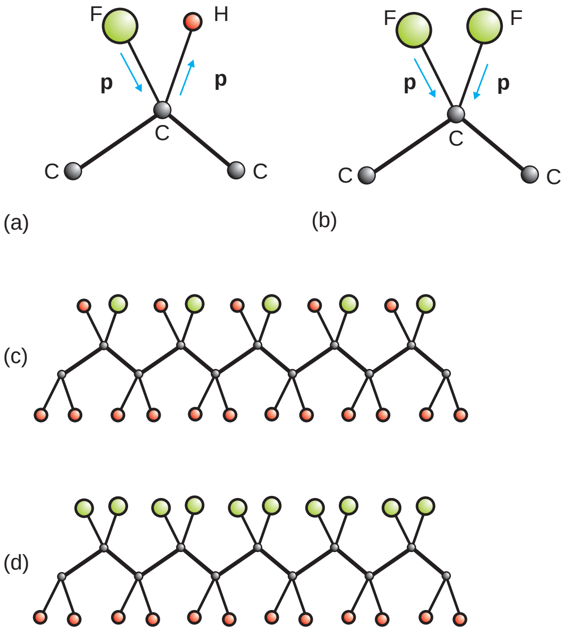


Figure 11.12 (a) Dipoles present in a tetrahedral unit of PVF, poly(vinyl fluoride) $[CH_2–CHF]_n$. (b) Dipoles present in a tetrahedral unit of PVF_2, poly(vinylidene fluoride) $[CH_2–CF_2]_n$. (c) Isotactic structure of a polymer chain of PVF. (d) Isotactic structure of a chain of PVF_2.

fluorine atoms are on the same side of the carbon–carbon backbone (Figure 11.12c). 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 $C \leftarrow F$ dipole opposed by a $C \rightarrow H$ dipole. Again, the isotactic form of the polymer has the highest net dipole moment (Figure 11.12d). Defects in the chain, especially caused by irregular linking of the monomer units during polymerisation, reduce the overall dipole moment of the chains.

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 11.13a). In odd polymers, such as nylon 5, the dipoles are aligned to give an observable polarisation (Figure 11.13b).

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 11.14). 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.

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. One type of reaction involves ion mechanosensitive channels, which open in response to tension in the cell membrane to allow ions to pass. It has recently been found that some mechanosensitive channels are built from piezoelectric proteins that respond to stress by polarising. This causes a change in their conformation so as to allow ion passage.

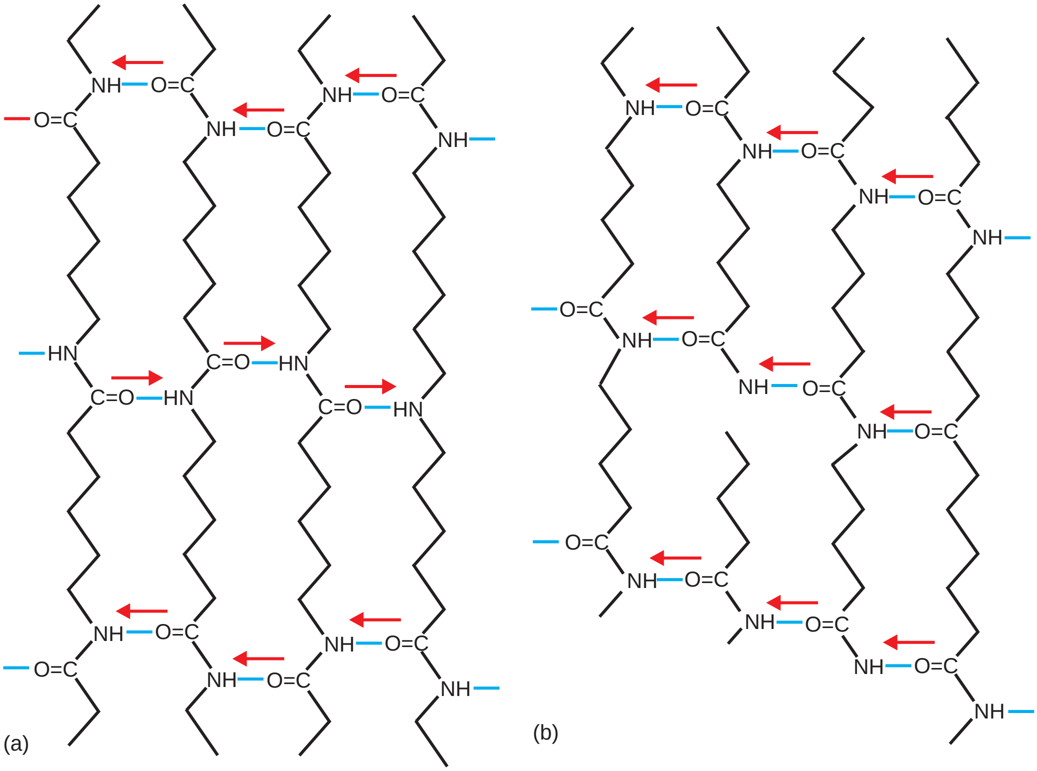


Figure 11.13 (a) The electric dipoles present in chains of an even nylon, nylon 6; no overall dipole moment is observed. (b) The electric dipoles present in chains of an odd nylon, nylon 5; the dipoles add to produce an observed dipole moment.

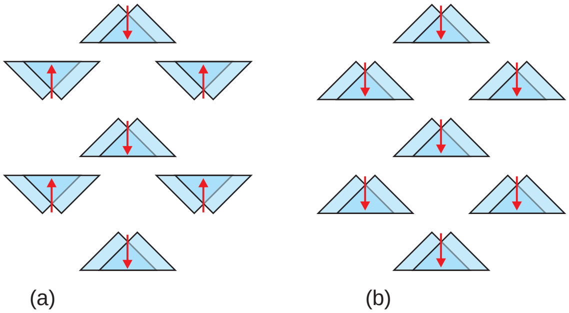


Figure 11.14 The crystal structure (schematic) of two forms of PVF_2, viewed down the polymer chains, shown as double triangles. The electric dipoles in the chains are drawn as arrows. (a) In a centrosymmetric structure the electric dipoles cancel and the material is a non-piezoelectric. (b) In a non-centrosymmetric structure the electric dipoles add and the material is piezoelectric.

11.3 Ferroelectrics

11.3.1 Ferroelectric 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 11.2). 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 dipoles.

The process can be illustrated by a simple model. Suppose we have a rectangular array of anions with centred cations (Figure 11.15a). This arrangement is

Table 11.2 Ferroelectrics and antiferroelectrics

Compound	Formula	Curie temperature/K	Spontaneous polarisation/C m^{-2}	Relative permittivity*
Hydrogen-bonded compounds				
Rochelle salt	NaK(COO.CHOH)$_2$.4H$_2$O	298	0.01	5×10^3
Triglycine sulphate	(NH$_2$CH$_2$COOH)$_3$.H$_2$SO$_4$	322	0.03	2×10^3
Potassium dihydrogen sulphate	KH$_2$PO$_4$	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$_2$NaNb$_5$O$_{15}$	833	0.40	6×10^4
Antiferroelectrics				
Tungsten trioxide	WO$_3$	1010	0	300*
Ammonium dihydrogen phosphate	NH$_4$H$_2$PO$_4$	148	0	57, 10*
Lead hafnate	PbHfO$_3$	476	0	200*
Lead zirconate	PbZrO$_3$	503	0	150*
Sodium niobate	NaNbO$_3$	627	0	700, 70*

*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.

non-ferroelectric. However, when relatively small cations are involved, the structure gains 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 11.15b). Repetition of this motif builds a structure with an aligned dipole array characteristic of a ferroelectric (Figure 11.15c,d).

The cation displacement can take place in one of two equivalent 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 11.16). 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 11.17).

Domain walls are not atomically smooth, as figuratively represented, but have a thickness of between 0.5–1 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.

11.3.2 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

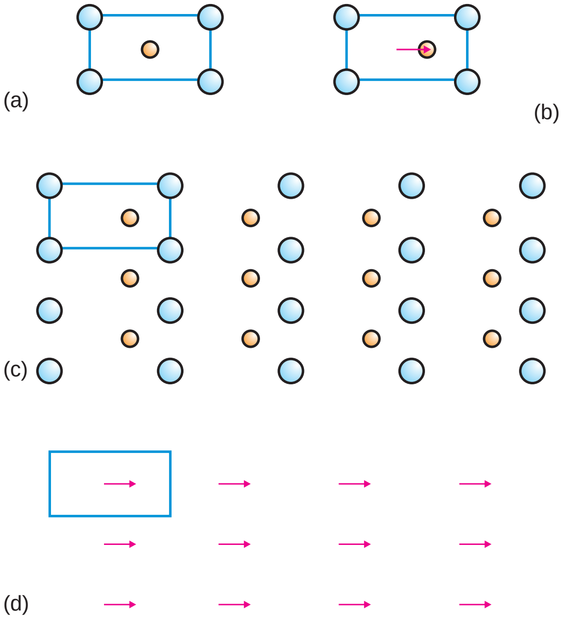


Figure 11.15 Ferroelectric crystal formation (schematic): (a) a cation-centred unit cell; (b) cation displacement creates an electric dipole in each cell; (c) the structure as an array of unit cells; (d) the structure as an array of electric dipoles.

enough to overcome the energy barrier between alternative configurations. This corresponds to the segment O–A on a graph of the polarisation versus field (Figure 11.18). As **E** increases, cations will start to gain sufficient energy to overcome this

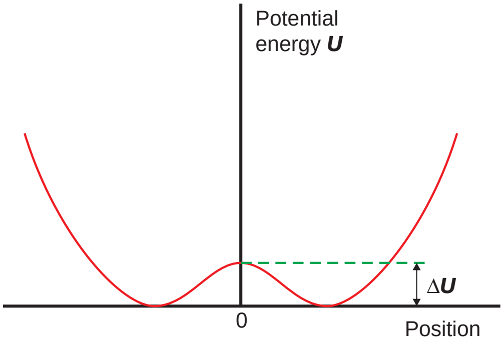


Figure 11.16 Variation of potential energy versus position for cation displacement from the centre of a surrounding anion polyhedron, schematic.

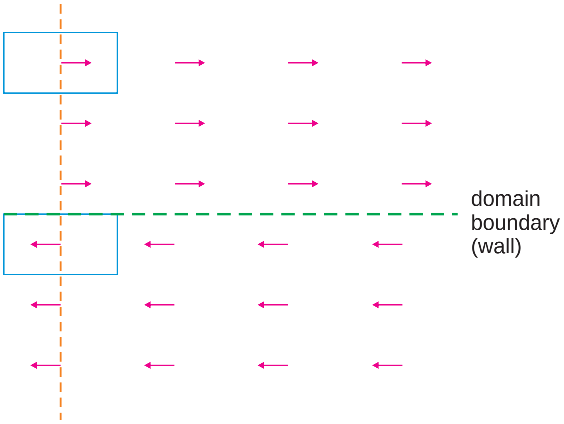


Figure 11.17 Domains due to the differing alignment of dipoles in adjacent regions of a crystal. The regions are separated by a domain boundary or wall, which extends over approximately a nanometre.

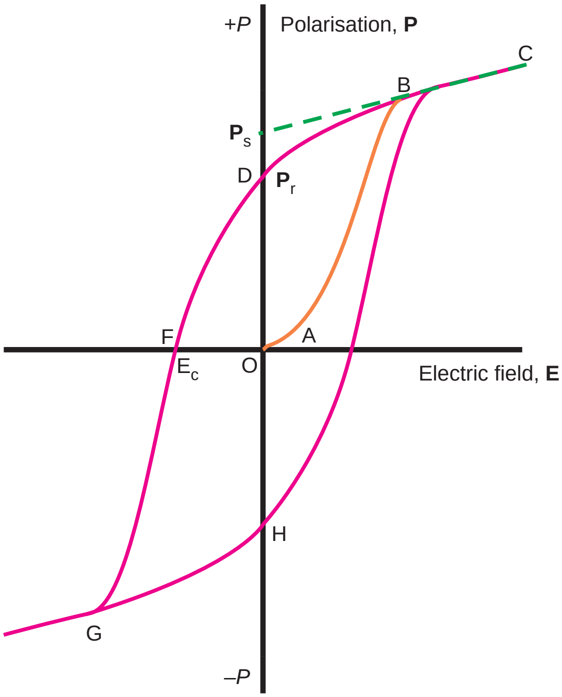


Figure 11.18 Hysteresis behaviour of the polarization, **P**, versus the applied electric field, **E**, for a ferromagnetic crystal. As the field takes values between $+E$ and $-E$, the polarisation, P, takes values between $+P$ and $-P$.

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 11.18). 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 11.18).

On reducing and then reversing the applied electric field the converse takes place but the **P**–**E** path will trace a new path 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 11.18). 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 definition of *hysteresis* is 'to lag behind'). Crystals that exhibit hysteresis and a domain structure are called *ferroic* materials.

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* $\mathbf{P}_s$.

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. For example, the nucleation and growth of ferroelectric domains in perovskite structure $BiFeO_3$ has been observed when subjected to an electric field applied by a metallic probe (Figure 11.19a).

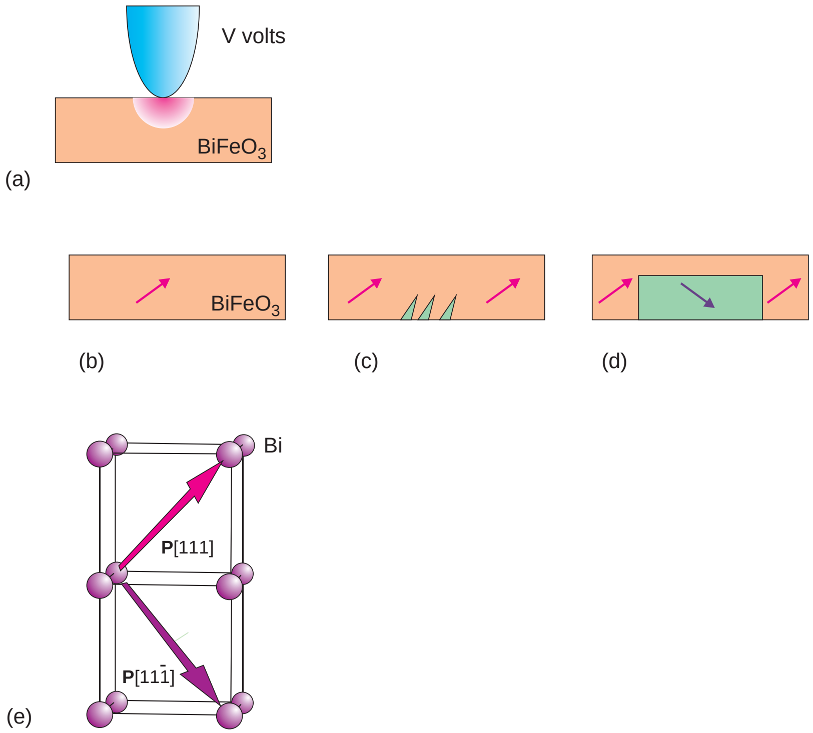


Figure 11.19 Domain growth in ferroelectric $BiFeO_3$: (a) metal probe with voltage V applied to a film of $BiFeO_3$; (b–d) domain growth, initiated at the lower surface of the film; (e) the change in polarisation direction with respect to the pseudocubic unit cell of $BiFeO_3$. (Adapted from Nekon *et al.* [17], see Further Reading.)

The samples consisted of single (001) sheets of $BiFeO_3$ with the polarisation vector along the [111] direction. Application of an electric field caused domains to be nucleated at the lower film surface as arrow-shaped incursions into the film. These amalgamated into a single domain as the field strength increased (Figure 11.19b–d). The polarisation in the new domains switched to $[11\bar{1}]$ (Figure 11.19e).

11.3.3 Antiferroelectrics

Ferroelectricity is governed by two types of factors: (i) chemical bonds, which are short-range forces; and (ii) dipolar interactions, which are long-range (Section 3.1). 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 11.2, Figure 11.20).

The balance between ferroelectric and antiferroelectric states is delicately poised, and some antiferroelectrics readily transform to ferroelectric states. For example, lead zirconate, $PbZrO_3$, 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. The Zr^{4+} cations are readily replaced by Ti^{4+} cations and a solid solution forms between $PbZrO_3$ and the similar $PbTiO_3$. A substitution of three or four per cent changes the phase from an antiferroelectric to a ferroelectric phase. The crystal structure also changes slightly from orthorhombic (antiferroelectric) to rhombohedral (ferroelectric), although both are only slightly distorted versions of the cubic perovskite structure (Figure 5.33b,c).

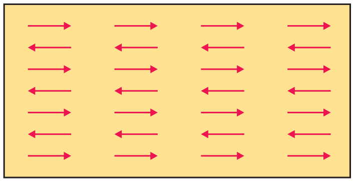


Figure 11.20 Antiparrallel arrangement of electric dipoles in an antiferroelectric material, schematic.

11.3.4 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 (Figure 11.21a and Sections 8.2–8.4). In either case the relative permittivity rises to a sharp peak in the neighbourhood of T_C (Figure 11.21b)

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*:

$$\varepsilon_r = \frac{C}{T - T_C}$$

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 $1/\varepsilon_r$ versus T:

$$\frac{1}{\varepsilon_r} = \frac{T}{C} - \frac{T_C}{C}$$

Ideally the graph is linear, with a slope of $1/C$ and an intercept on the T-axis of T_C (Figure 11.22). Frequently the point of intercept, T_0, is slightly different from the measured value of T_C, and the Curie–Weiss

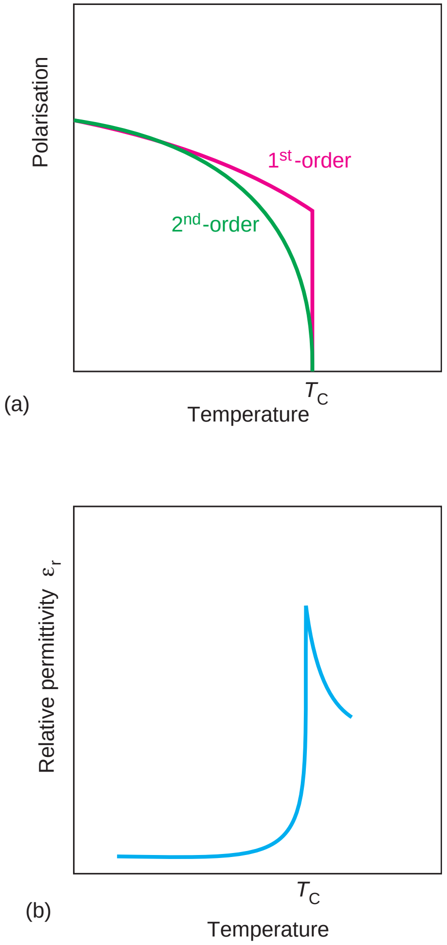


Figure 11.21 Ferroelectric characteristics near to the Curie temperature: (a) the polarisation of a ferroelectric approaches zero sharply in a first-order transition, or gradually for a second-order transition; (b) the relative permittivity rises to a sharp peak.

equation is often written in the form:

$$\varepsilon_r = \frac{C}{T - T_0}$$

where T_0 is the *extrapolated Curie temperature* (Figure 11.22). 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.

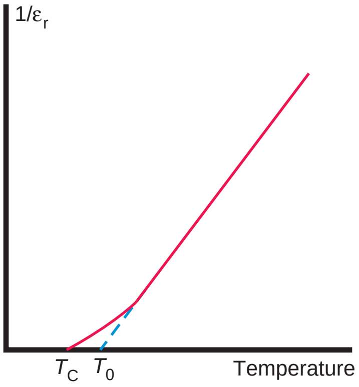


Figure 11.22 The Curie–Weiss behaviour of a ferroelectric solid above the Curie temperature.

11.3.5 Ferroelectricity due to hydrogen bonds

Hydrogen bonds are formed when a hydrogen atom sits between two electronegative atoms in an off-centre position (Section 3.1). At temperatures below the Curie temperature the hydrogen atoms are ordered on one side of the hydrogen bond or the other (Figure 11.23a). As the hydrogen in a hydrogen bond can occupy two equally stable positions, it is not difficult to see a possible origin for

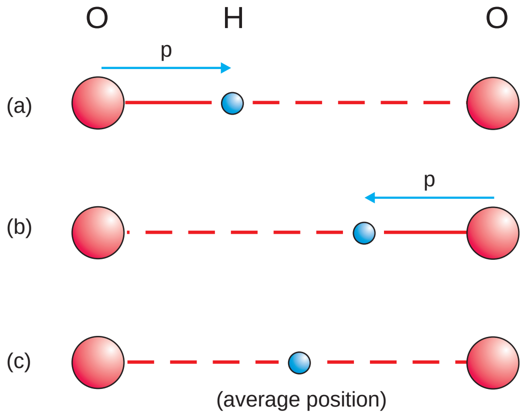


Figure 11.23 Hydrogen bonds (shown dashed) in ferroelectrics: (a, b) below the Curie temperature, T_C, hydrogen atoms in hydrogen bonds lie to one side or the other of the centre; (c) above T_C the hydrogen atoms are, on average, central.

ferroelectric switching. A sufficiently high electric field will swap the dipole direction by causing the hydrogen ion to jump to the alternative position (Figure 11.23b). At temperatures higher than the Curie temperature, atomic vibrations induced by thermal energy overcome the barrier between the two alternative positions 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 11.23c). The compounds in this group are ordered at lower temperatures and become disordered at higher temperatures, and display an order–disorder transition (Section 8.4).

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 sulphate, $(NH_2CH_2COOH)_3.H_2SO_4$. However, the interaction of the hydrogen bonds with other features of the crystal structure usually makes each compound unique. This feature 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 11.24a) 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. Below 123 K the hydrogen atoms order (Figure 11.24b) so that each (PO_4) tetrahedron in the high-temperature form is converted into a $[PO_2(OH)_2]$ tetrahedron

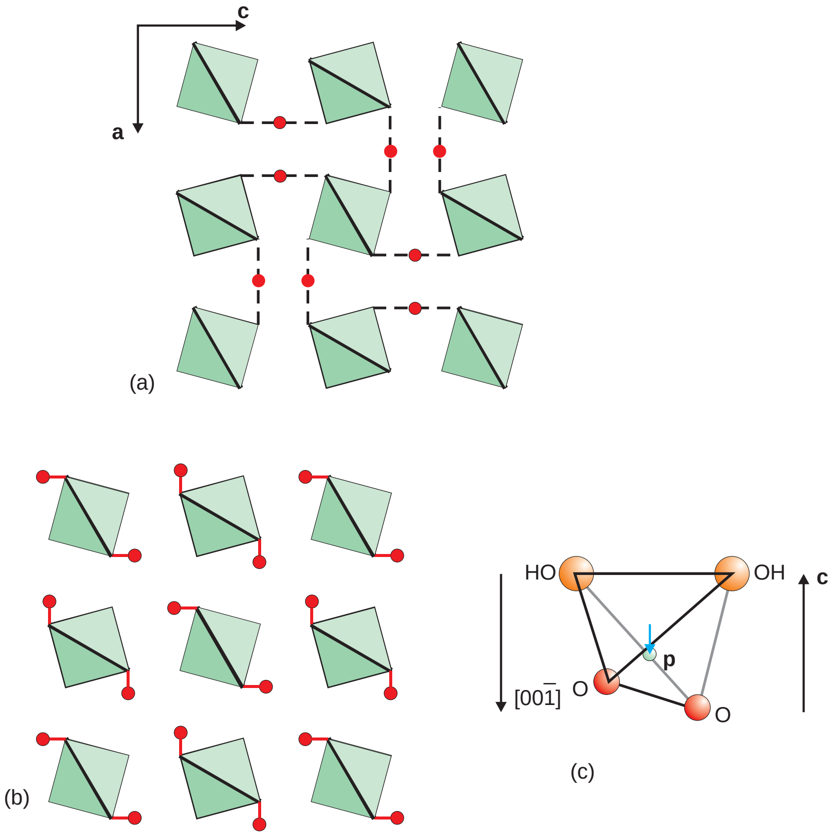


Figure 11.24 Ferroelectricity in KH_2PO_4: (a) skeleton of the structure projected down [001]; the PO_4 tetrahedra project as squares; hydrogen bonds, dotted lines; hydrogen atoms, grey circles; potassium atoms are omitted; (b) low-temperature structure with ordered H atoms; (c) the displacement of the P atoms in the PO_4 tetrahedra, as a result of the H-atom ordering and the subsequent formation of (OH) groups, induces an electric dipole, **p**, parallel to the **c**-axis.

(Figure 11.24c). 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 phosphorus and hydrogen atoms are closely linked. When an external field is applied, both the hydrogen and phosphorus atoms switch in concert.

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 $NH_4H_2PO_4$ is an antiferroelectric below the transition temperature.

11.3.6 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 (Section 8.4).

Sodium nitrite, $NaNO_2$, is an example of this behaviour. The structure is similar to that of NaCl, and if the NO_2 groups were spherical instead of shaped like blunt arrowheads, it would be identical (Figure 11.25). They point along the **b**-axis with their planes perpendicular to the **a**-axis. The dipoles in each NO_2^- group point along the **b**-axis. In an applied electric field the 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.

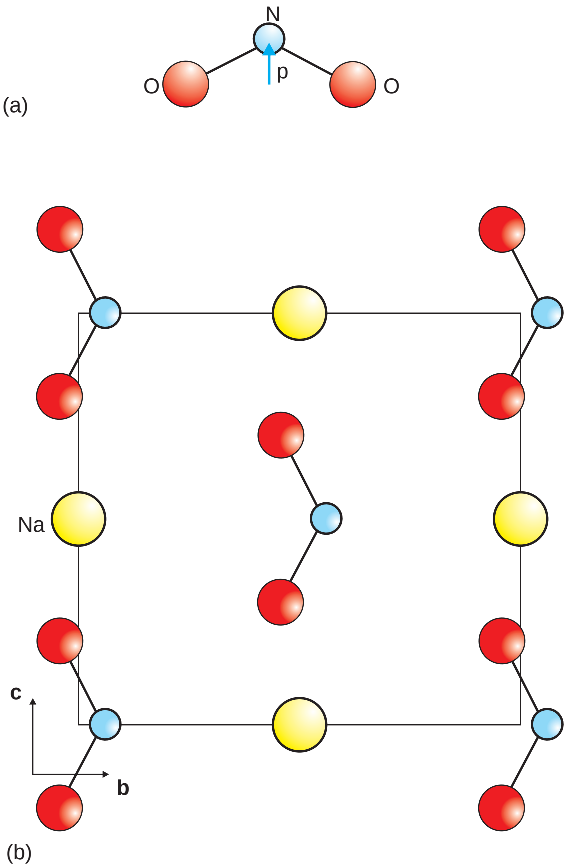


Figure 11.25 Ferroelectricity in $NaNO_2$: (a) the structure of a planar nitrite (NO_2^-) group, the electric dipole, **p**, in each unit points towards the N atom; (b) the low-temperature structure of sodium nitrite, $NaNO_2$, projected down [100]. The dipoles are aligned along the **b**-axis. In the high-temperature paraelectric form the dipoles are arranged at random along +**b** and −**b**.

As the NO_2^- ion is fairly rigid, the rotation mechanism operates.

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 pointing along −**b**. This transition is 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.

11.3.7 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, 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 (Figure 5.33b,c). A displacive transition (Section 8.3.1) 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 11.26a,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 K 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 11.26c). The O^{2-}—Ti^{4+} bond lengths parallel to the **c**-axis are now 0.22 nm 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 11.26d) 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 ferroelectric. There is no preference as to which of the original cubic axes becomes the polar direction, and so this can take one of six equivalent directions, parallel to ± **x**, ± **y** or ± **z**. On cooling, a ferroelectric domain pattern forms, reflecting this symmetry.

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 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 $\langle 100 \rangle$ direction in all structures rather than apparently different directions when referred to the true tetragonal, orthorhombic or rhombohedral cells.

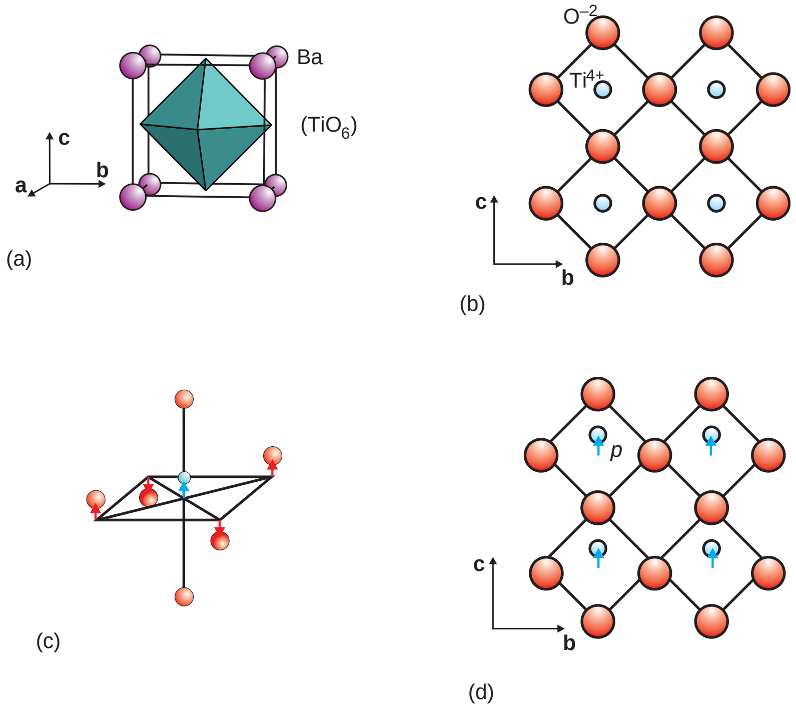


Figure 11.26 The barium titanate structure: (a) the cubic perovskite structure of paraelectric high-temperature cubic barium titanate, $BaTiO_3$, with the Ba^{2+} ions at the unit cell corners and a TiO_6 octahedron at the unit cell centre; (b) projection of the structure down [100], with the Ba^{2+} ions omitted; (c) octahedral distortion and Ti^{+4} ion displacement in the tetragonal ferroelectric phase, exaggerated; (d) projection of the tetragonal form, with the ionic displacements exaggerated and the Ba^{2+} ions omitted. The electric dipoles, **p**, that are generated by the off-centre displacement of the Ti^{4+} ions, point along the **c**-axis of the tetragonal unit cell.

11.3.8 Poling and polycrystalline ferroelectric solids

A ferroelectric crystal 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.

11.3.9 Doping and modification of properties

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 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, or *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 11.27). The diagrams are also of value because highly desirable dielectric properties are found in the neighbourhood of *morphotropic phase boundaries* (MPBs), which are boundaries within a broad solid solution range where the structure changes, often occurring within solid solutions formed between structurally dissimilar end members. For example, the boundary between the rhombohedral and tetragonal phases in the $PbZrO_3$–$PbTiO_3$ (PZT) system (Figure 11.27) is an MPB.

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

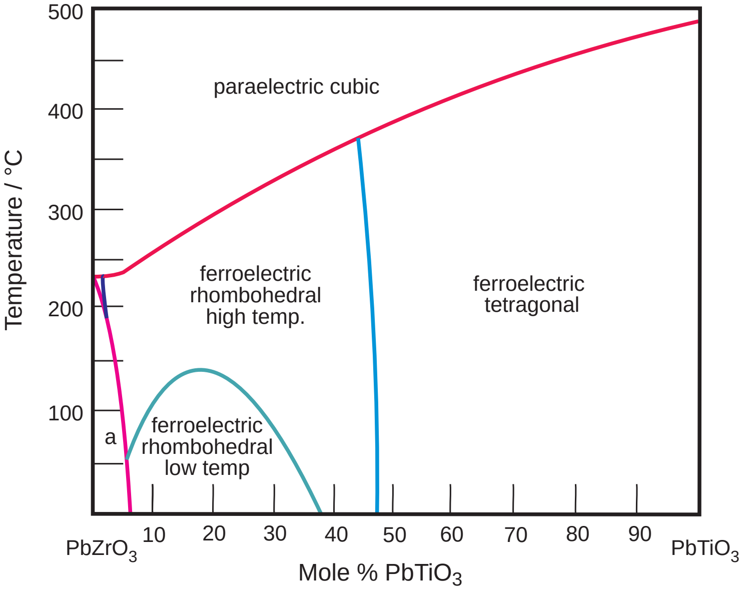


Figure 11.27 The phase diagram of the $PbZrO_3$–$PbTiO_3$ (PZT) system. The antiferroelectric phase region is found in the region marked a.

polycrystalline ceramic samples with desired piezoelectric and ferroelectric characteristics.

11.3.10 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 above in the following ways.

1. 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 11.28a).
2. 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 11.28b). There is no Curie–Weiss behaviour above T_m.
3. The hysteresis loop tends to be narrow with a low value for the remnant polarisation (Figure 11.28c).
4. 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.

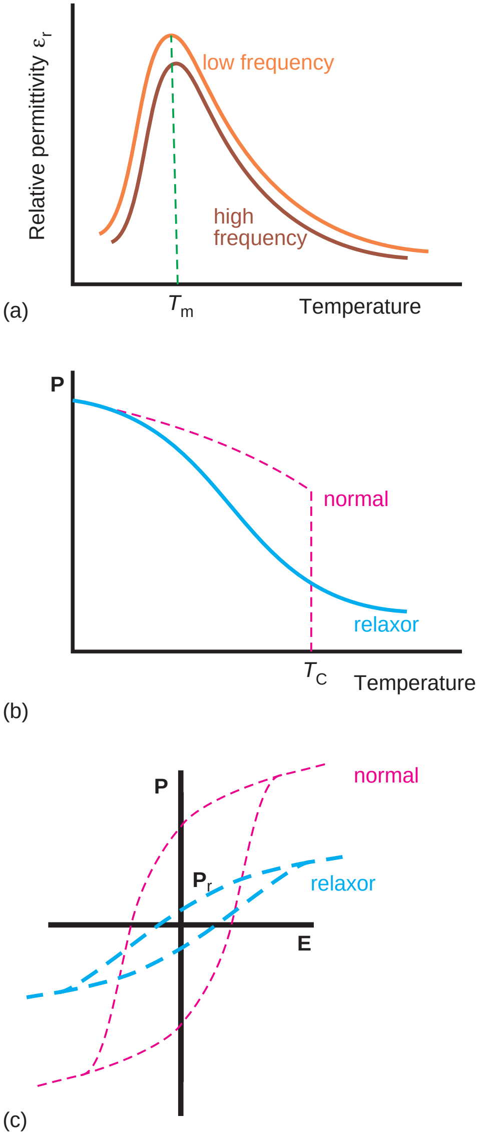


Figure 11.28 Relaxor ferroelectrics schematic: (a) variation of relative permittivity with temperature and measurement frequency; (b) variation of polarisation with temperature; (c) variation of polarisation with electric field (hysteresis loop).

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 B1 and B2 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.

11.3.11 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. This means that distortions such as polyhedral rotation may be blocked in the bulk but become allowed at surfaces. 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 whether new features will emerge as a result of surface interference.

Evidence now suggests that ferroelectric properties persist in small particles prepared by milling of bulk material. It appears, though, that at dimensions below about 10 unit cells, say 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 1 or 2 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 11.29). 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

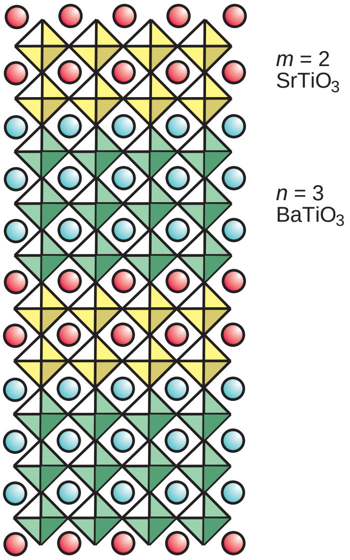


Figure 11.29 A ferroelectric superlattice composed of $m = 2$ unit cells of $SrTiO_3$ and $n = 3$ unit cells of $BaTiO_3$.

$CaTiO_3$ 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, producing an enhanced polarisation of the TiO_6 octahedra in the $BaTiO_3$ component.

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 thickness reduces, but surprisingly, at the lowest values, ferroelectricity recovers. In bulk $PbTiO_3$ octahedral tilt is suppressed and in $SrTiO_3$ oxygen rotation is suppressed. 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 [2] in Further Reading). In essence, improper ferroelectricity is ferroelectricity that is not due to the normal polarisation of the structure (Section 11.2) 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 changed this, and now improper ferroelectrics are becoming widely studied.

11.3.12 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 (Section 10.1.2) 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 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 number of bits 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 11.16) 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 11.30a). 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 the 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 surrounding crystal 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,

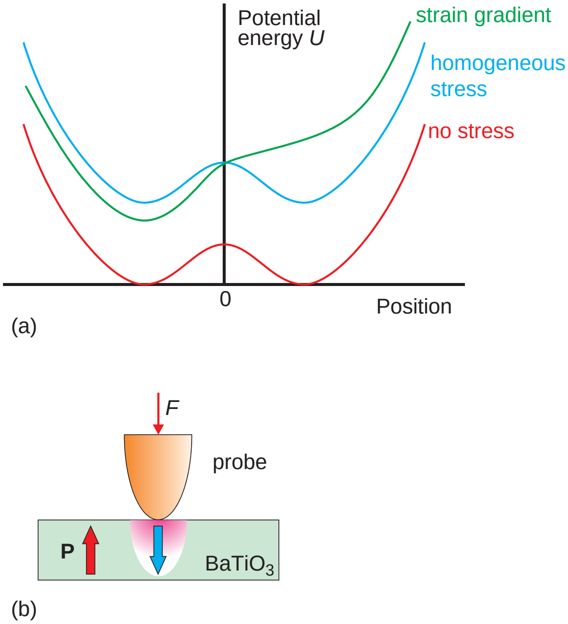


Figure 11.30 The flexoelectric effect: (a) evolution of the potential energy curve for a ferroelectric under a homogeneous stress and strain gradient; (b) switching via mechanical stress imposed by the tip of an atomic probe microscope. (Adapted from Lu *et al.* (2012), see Further Reading.)

creating a domain with dimensions of the order of 30 nm diameter (Figure 11.30b). The resultant mechanically written domain pattern can be both read and erased electrically, thus generating a mechanically written permanent memory store.

Further reading

General:

Kingery, W.D., Bowen, H.K. and D. R. Uhlmann, D.R. (1976) Chapter 18, in *Introduction to Ceramics*, 2nd edn. John Wiley & Sons, Ltd., Chichester.

Lines, M.E. and Glass, A.M. (2001) *Principles and Applications of Ferroelectrics and Related Materials*. Oxford Classics, Oxford University Press, Oxford.

Introductory crystallography with respect to the dielectric properties described in:

Bloss, F.D. (1971) Chapter 11, in *Crystallography and Crystal Structures*. Holt Rinehart and Winston, New York.

Megaw, H.D. (1973) Chapter 15 in *Crystal Structures*. W.B. Saunders, Philadelphia.

Newnham, R.E. (1975) Chapter 4, in *Structure–Property Relations*. Springer, Berlin.

Tilley, R.J.D. (2006) Chapter 4, in *Crystals and Crystal Structures*. John Wiley & Sons, Ltd., Chichester.

Specific materials are described in:

Coste, B., *et al.* (2011) Piezo proteins are pore-forming subunits of mechanically activated channels. *Nature*, **483**: 176–81.

Withers, R.L., Thompson, J.G. and Rae, A.D. (1991) The crystal chemistry underlying ferroelectricity in $Bi_4Ti_3O_{12}$, Bi_3TiNbO_9 and Bi_2WO_6. *J. Solid State Chem.*, **94**: 404.

Newnham, R.E. (1997) Molecular mechanisms in smart materials. *Materials Research Society Bulletin*, **22**: 20.

Dunn, P.E. and Carr, S.H. (1989) A historical perspective on the occurrence of piezoelectricity in materials. *Materials Research Society Bulletin*, **XIV** (February): 22.

Haertling, G.H. (1999) Ferroelectric ceramics: history and technology. *J. Amer. Ceram. Soc.*, **82**: 797–818.

Shannon, R.D. and Oswald, R.A. (1991) Dielectric constants . . . and the oxide additivity rule. *J. Solid State Chem.*, **95**: 313.

Nanoparticles, thin films and superlattices:

Ahn, C.H., Rabe, K.M. and Triscone, J.-M. (2004) Ferroelectric superlattices. *Science*, **303**: 488–91.

Back, S.H., *et al.* (2011) Giant piezoelectricity on Si for hyperactive MEMS. *Science*, **334**: 958–61.

Bousquet, E., *et al.* (2008) Improper ferroelectricity in perovskite oxide artificial superlattices. *Nature*, **452**: 732–6.

Dawber, M., *et al.* (2007) Tailoring the properties of artificially layered ferroelectric superlattices. *Adv. Mater.*, **19**: 4153–9.

Nekon, C.T., *et al.* (2011) Domain dynamics during ferroelectric switching. *Science*, **334**: 968–71.

Rondinelli, J.M. and Fennie, C.J. (2012) Octahedral rotation-induced ferroelectricity in cation ordered perovskites. *Adv. Mater.*, **24**: 1961–8.

Rørvik, P.H., Grande, T. and Einarsrud, M.-A. (2011) One-dimensional nanostructures of ferroelectric perovskites. *Adv. Mater.*, **23**: 4007–34.

Yang, H.Y., *et al.* (2012) Emergent phenomena at oxide interfaces. *Nature Materials*, **11**: 103–13.

Flexoelectric effect

Lu, H., *et al.* (2012) Mechanical writing of ferroelectric polarization. *Science*, **336**: 59–61.

Problems and exercises

Quick quiz

1 The relative permittivity of a material is also called the:
(a) Dielectric constant.
(b) Dielectric permittivity.
(c) Dielectric susceptibility.

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.

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.

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.

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.

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.

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.

8 A spontaneous polarisation is NOT found in:
(a) Piezoelectric solids.
(b) Ferroelectric solids.
(c) Pyroelectric solids.

9 A solid that has a switchable spontaneous polarisation is:
(a) A pyroelectric.
(b) A piezoelectric.
(c) A ferroelectric.

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.

11 The crystal structure of pyroelectric crystals must contain:
(a) A polar axis.
(b) A centre of symmetry.
(c) Switchable permanent dipoles.

12 Ferroelectric crystals must possess:
(a) Hydrogen bonds.
(b) Switchable dipoles.
(c) Polar groups.

13 Hysteresis is characteristic of:
(a) Pyroelectric crystals.
(b) Piezoelectric crystals.
(c) Ferroelectric crystals.

14 The dipoles in an antiferroelectric crystal are:
(a) In parallel rows.
(b) In antiparallel rows.
(c) In randomly aligned rows.

15 The paraelectric state of a ferroelectric is:
(a) The high-temperature phase.
(b) The low-temperature phase.
(c) The antiferroelectric phase.

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.

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.

18 Hydrogen bonding is NOT the cause of ferroelectricity in:
(a) Triglycine sulphate.
(b) Rochelle salt.
(c) Sodium nitrite.

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.

20 The process by which polycrystalline solids can be made ferroelectric is:
(a) Annealing.
(b) Sintering.
(c) Poling.

Calculations and questions

11.1 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?

11.2 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?

11.3 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?

11.4 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?

11.5 The dipole moment of a water molecule is 6.2×10^{-30} C m, 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.

11.6 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.

11.7 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; sulphur 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.

11.8 Derive the Clausius–Mosotti equation using equations:

$$\mathbf{P} = (\varepsilon_r - 1)\,\varepsilon_0\,\mathbf{E}_0$$
$$\mathbf{P} = N\,\alpha\,\mathbf{E}_{loc}$$
$$\mathbf{E}_{loc} = \mathbf{E}_0 + \frac{\mathbf{P}}{3\varepsilon_0}$$

11.9 The following data are for a single crystal of MgO. Estimate (a) the electronic and ionic polarisability volumes, (b) the electronic and ionic polarisabilities, of this material. Relative permittivity, 9.65; refractive index, 1.736; unit cell, cubic, $a = 0.4207$ nm, $Z = 4$ formula units of MgO.

11.10 The following data are for a single crystal of α-quartz: average relative permittivity, 4.477; average refractive index, 1.5485; unit cell, hexagonal, $a = 0.49136$ nm, $c = 0.54051$ 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.

11.11 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 11.9 and 11.10, assuming that only ionic and electronic polarisations are important.

11.12 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 principal refractive indices are 1.635, 1.651, 1.670. Estimate the electronic polarisability of the mineral.

11.13 Estimate (a) the polarisability and (b) 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.

11.14 The relative permittivity of the garnet $Y_3Fe_5O_{12}$ was measured as 15.7. Calculate (a) the polarisability volume and (b) 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 $Y_3Fe_5O_{12}$.

11.15 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.

11.16 Use the additivity rule to estimate (a) the 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.

11.17 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}.

11.18 Show that the units for the direct piezoelectric coefficient, d, of $\mathrm{C\,N^{-1}}$ and $\mathrm{m\,V^{-1}}$ are equivalent.

11.19 The value of the piezoelectric coefficient for quartz is given as $2.3\,\mathrm{pC\,N^{-1}}$. Calculate the polarisation of a plate of dimensions $10\,\mathrm{cm} \times 5\,\mathrm{cm} \times 0.5\,\mathrm{mm}$ when a mass of 0.5 kg is placed on it.

11.20 An electret film has a piezoelectric coefficient of $170\,\mathrm{pC\,N^{-1}}$. Calculate the change in thickness when 500 V are applied across a film 0.1 mm thick.

11.21 The semiprecious gemstone tourmaline, with an approximate formula $\mathrm{CaLi_2Al_7(OH)_4(BO_3)_3Si_6O_{18}}$, has a pyroelectric coefficient, π_i, of $4 \times 10^{-6}\,\mathrm{C\,m^{-2}\,K^{-1}}$. The unique polar axis is the crystallographic **c**-axis. What is the change in polarisation caused by a change of temperature of 100°C?

11.22 The measured relative permittivity, ε_r, of a ceramic sample of $\mathrm{PbZrO_3}$ as a function of temperature, T°C, is given below. Determine (a) the Curie temperature, (b) the Curie constant for this sample.

ε_r	130	142	166	222	360	420	472	556
T/°C	50	100	150	200	225	230	234	235
ε_r	775	3200	3000	2840	2440	1620	1240	840
T/°C	236	238	240	242	250	275	300	350

11.23 The measured relative permittivity, ε_r, of a ceramic sample of $\mathrm{Cd_2Nb_2O_7}$ as a function of temperature, T°C, is given below. Determine (a) the extrapolated Curie temperature, (b) the Curie constant, C, for this sample.

ε_r	4500	4125	3750	3500	3225	3000	2800	2600
T/°C	−80	−75	−70	−65	−60	−55	−50	−45
ε_r	2465	2280	2115	2000	1860	1750	1630	1560
T/°C	−40	−35	−30	−25	−20	−15	−10	−5

11.24 The measured relative permittivity, ε_r, of a crystal of triglycine sulphate as a function of temperature, T°C, is given below. Determine (a) the Curie temperature, (b) the Curie constant for triglycine sulphate.

ε_r	120	190	280	400	540	730	1300	~7000
T/°C	40	41	42	43	44	45	46	47
ε_r	1100	830	700	590	520	460	420	380
T/°C	48	49	50	51	52	53	54	55
ε_r	250	180	130	110				
T/°C	60	65	70	75				

11.25 Zinc oxide, which is isostructural with wurtzite, figure 11.8, has a hexagonal unit cell with $a = 0.3250\,\mathrm{nm}$, $c = 0.5207\,\mathrm{nm}$, $Z = 2$. The atom positions are:

Zn: $\frac{1}{3}, \frac{2}{3}, 0; \frac{2}{3}, \frac{1}{3}, \frac{1}{2}$
O: $\frac{1}{3}, \frac{2}{3}, 0.3825; \frac{2}{3}, \frac{1}{3}, 0.8525$

Estimate the maximum spontaneous polarisation of ZnO, assuming that the structure is ionic.

11.26 Calculate the maximum spontaneous polarisation of a crystal of sodium nitrite given that the unit cell is orthorhombic, with $a = 0.360\,\mathrm{nm}$, $b = 0.575\,\mathrm{nm}$, $c = 0.535\,\mathrm{nm}$, $Z = 2$ formula units of $\mathrm{NaNO_2}$. The dipole moment of each N—O bond is $0.5 \times 10^{-30}\,\mathrm{C\,m^{-1}}$ and the O–N–O angle is 115°.

11.27 (a) Calculate the dipole moment of a $\mathrm{TiO_6}$ octahedron in tetragonal $\mathrm{BaTiO_3}$, $a = 0.3997\,\mathrm{nm}$, $c = 0.4031\,\mathrm{nm}$, assuming that the compound is fully ionic and the $\mathrm{Ti^{4+}}$ ions are displaced by 0.012 nm along the **c**-axis of the unit cell.

(b) Determine the maximum spontaneous polarisation under these conditions.

11.28 (a) Calculate the dipole moment of an $\mathrm{NbO_6}$ octahedron in tetragonal $\mathrm{KNbO_3}$, $a = 0.4002\,\mathrm{nm}$, $c = 0.4064\,\mathrm{nm}$, assuming that the compound is fully ionic and the $\mathrm{Nb^{5+}}$ ions are displaced by 0.017 nm along the **c**-axis of the unit cell.

(b) Determine the maximum spontaneous polarisation under these conditions.

11.29 (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.

11.30 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 piezoelectric, as is quartz?

12

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 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.

12.1 Magnetic materials

12.1.1 Characterisation of magnetic materials

The weak magnetic properties of most solids can be measured using a *Gouy balance* (Figure 12.1). 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.)

Diamagnetic and paramagnetic substances are characterised by their *susceptibility* to the magnetic

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

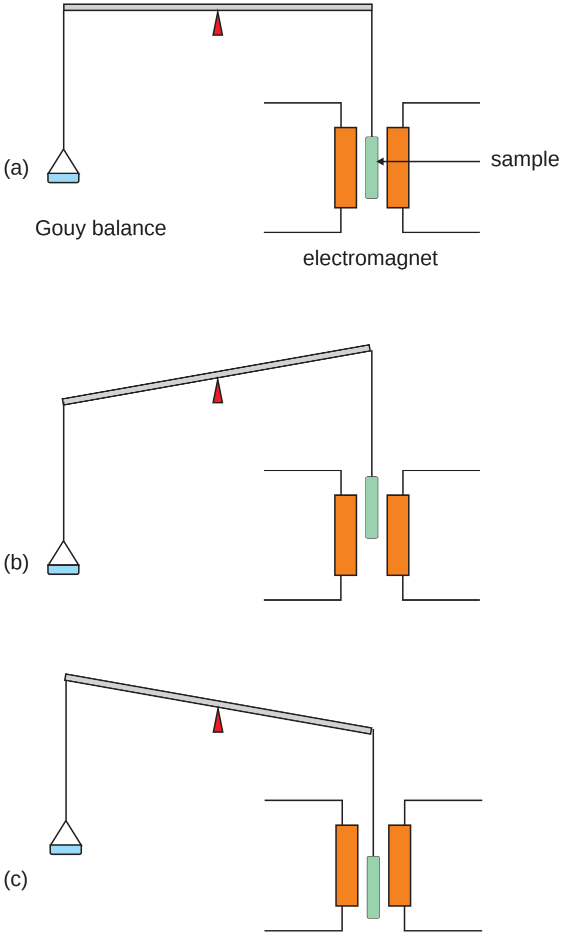


Figure 12.1 Magnetic materials in a Gouy balance: (a) electromagnet off; (b) a diamagnetic substance; (c) a paramagnetic substance.

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.

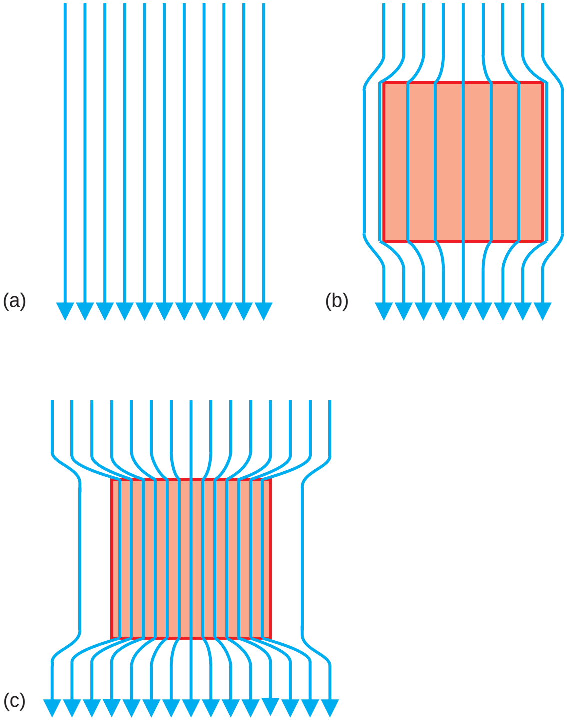


Figure 12.2 Weak magnetic materials in an external magnetic field: (a) no solid present; (b) a diamagnetic solid; (c) a paramagnetic solid.

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 12.2).

12.1.2 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:

$$\mu = [\text{current in loop}] \times [\text{area of loop}] = I\pi r^2$$

where I is the current and r is the radius of the loop.

The Earth acts as a magnetic dipole, with a dipole moment of approximately $8 \times 10^{22}\,\mathrm{Am}^2$. 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 vacuum permeability. The magnetic field within the solenoid is given by:

$$\mathbf{H} = In$$

where I is the current in the solenoid and n is the number of turns of the coil per meter.

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:

$$\begin{aligned} \mathbf{B} &= \mathbf{B}_0 + \mu_0\mathbf{M} \\ &= \mu_0\mathbf{H} + \mu_0\mathbf{M} \\ &= \mu_0(\mathbf{H} + \mathbf{M}) \end{aligned} \qquad (12.1)$$

where **M** is the magnetisation of the sample. If **M** opposes $\mathbf{B}_0$, **B** is less than $\mathbf{B}_0$ and the sample is diamagnetic. If **M** is in the same direction as $\mathbf{B}_0$, **B** is increased and the substance is a paramagnetic or ferromagnetic. Equation (12.1) is also written:

$$\mathbf{B} = \mu_0\mathbf{H} + \mathbf{J}$$

where **J** is the *magnetic polarization*.

For many paramagnetic materials, **M** is small and it is possible to write:

$$\mathbf{B} \approx \mu_0\mathbf{H}$$

For many ferromagnetic materials, **B** is much greater than **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, **H**, by the expression:

$$\mathbf{M} = \chi\mathbf{H} \qquad (12.2)$$

where the constant of proportionality, χ, is the magnetic susceptibility.

Combining equations (12.1) and (12.2) gives:

$$\begin{aligned} \mathbf{B} &= \mu_0(1 + \chi)\mathbf{H} \\ \text{hence } \mu &= \mu_0(1 + \chi) \\ \text{and } \quad \mu &= (1 + \chi) \end{aligned}$$

12.1.3 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* (Section 1.2).

A spinning electron has a magnetic dipole moment, the *Bohr magneton*, μ_B, given by:

$$\mu_B = \frac{eh}{4\pi m}$$

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:

$$\mathbf{m}_{\text{orbital}} = \mu_B\sqrt{l(l+1)}$$

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, . . . $\pm l$. 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:

$$\mathbf{m}_{\text{spin}} = g\mu_B\sqrt{s(s+1)}$$

where *g* is *the free electron g-value*, equal to 2.002319, and *s* is $^1/_2$. 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 $\pm^1/_2$. 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 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 *Russell-Saunders coupling* (Sections 1.3.1, 1.3.2). In summary, individual electron spin quantum numbers, *s*, are added to give a many-electron total spin quantum number, *S*, and 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*. More details are given below for the lanthanoids and 3d transition metals.

The total magnetic dipole moment of the atom is given by:

$$\mathbf{m}_{\text{atom}} = g_J\mu_B\sqrt{J(J+1)} \qquad (12.3)$$

where g_J is the *Landé g-factor*, given by:

$$g_J = 1 + \frac{J(J+1) - L(L+1) + S(S+1)}{2J(J+1)} \qquad (12.4)$$

The method of calculating *S*, *L* and *J* is described in Section 1.3.2. 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, . . . $\pm J$, 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. These are particularly found in the transition metals, with incompletely filled d shells, and the lanthanoids and actinoids, which have incompletely filled f shells.

12.1.4 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 (Figure 12.3a,b). The imposition of a magnetic field generates weak magnetic dipoles that are only present for as long as the field persists. 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.

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 12.3c,d). 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.

Because of thermal agitation it would be expected that paramagnetic susceptibility would vary with temperature (Figure 12.4a). This is given by the *Curie Law*:

$$\chi = \frac{C}{T}$$

where χ is the magnetic susceptibility, T is the temperature (K) and C is the *Curie constant*. Curie Law

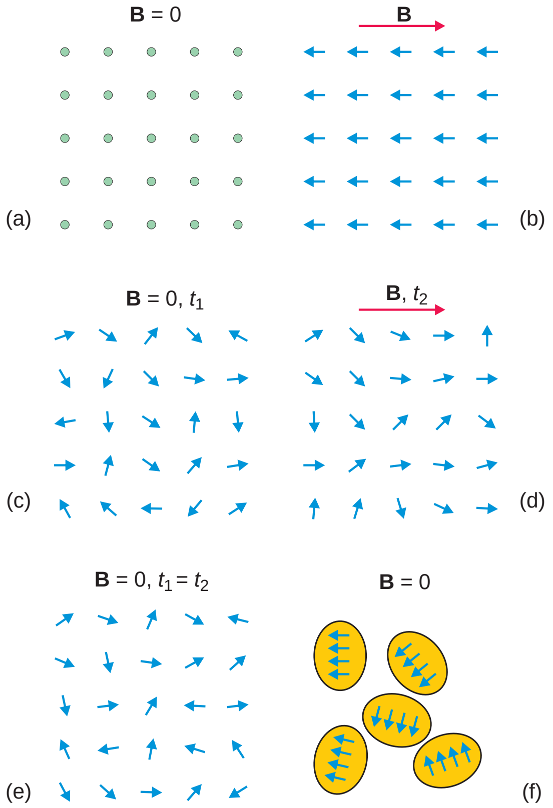


Figure 12.3 The effect of an applied magnetic flux density, **B**, on a solid: (a, b), a diamagnetic solid; (c, d) a paramagnetic solid; (e) a spin glass, similar to (c), but below a temperature T_f the orientation of the dipoles changes slowly; (f), a cluster glass with oriented dipoles in small volumes below a temperature T_f; (g) a ferromagnetic solid; (h) an antiferromagnetic solid; (i) a canted magnetic solid; (j) a ferrimagnetic solid.

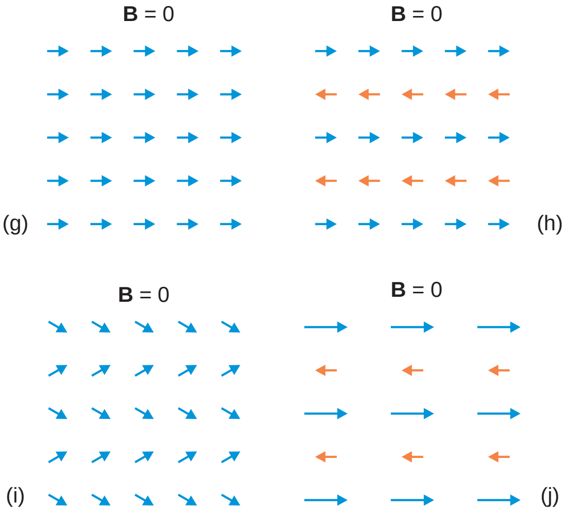


Figure 12.3 (*Continued*)

dependence in a solid is indicative of the presence of isolated paramagnetic ions or atoms in the material.

Interacting magnetic dipoles can produce a variety of magnetic properties in a solid. The interactions can increase sufficiently that at low temperatures 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 a *spin glass* (Figure 12.3e). The phase $Bi_4V_{1.7}Ni_{0.3}O_{10.55}$ forms a spin glass when the room temperature paramagnetic state is cooled to a low temperature. In other materials, the interactions below T_f are strong enough for local ordering, but because of the crystal structure, the localised regions are restricted and no long-range order occurs (Figure 12.3f). This arrangement is called a *cluster glass*, and is found at low temperatures in compounds such as Prussian Blue, $KFe_2(CN)_6$, which has a magnetic freezing point of approximately 25 K.

Ferromagnetic materials are those in which the magnetic dipoles align parallel to each other over considerable distances in the solid (Figure 12.3g). 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 actinoides with unpaired f electrons.

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.

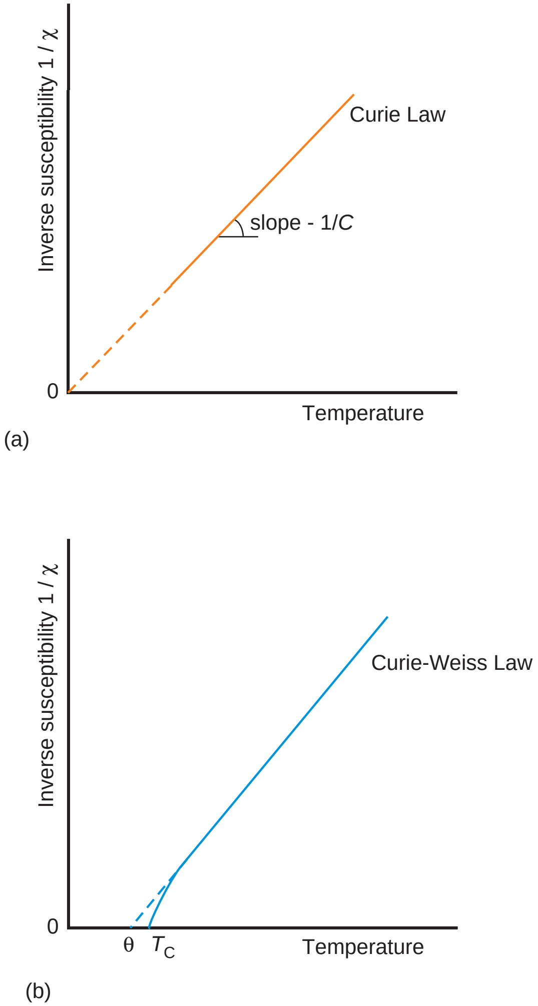


Figure 12.4 The temperature dependence of the reciprocal magnetic susceptibility: (a) Curie Law behaviour of a paramagnetic solid; (b) Curie–Weiss Law behaviour of a ferromagnetic solid above the Curie temperature; (c) Curie–Weiss Law behaviour of an antiferromagnetic solid; (d) a ferrimagnetic solid.

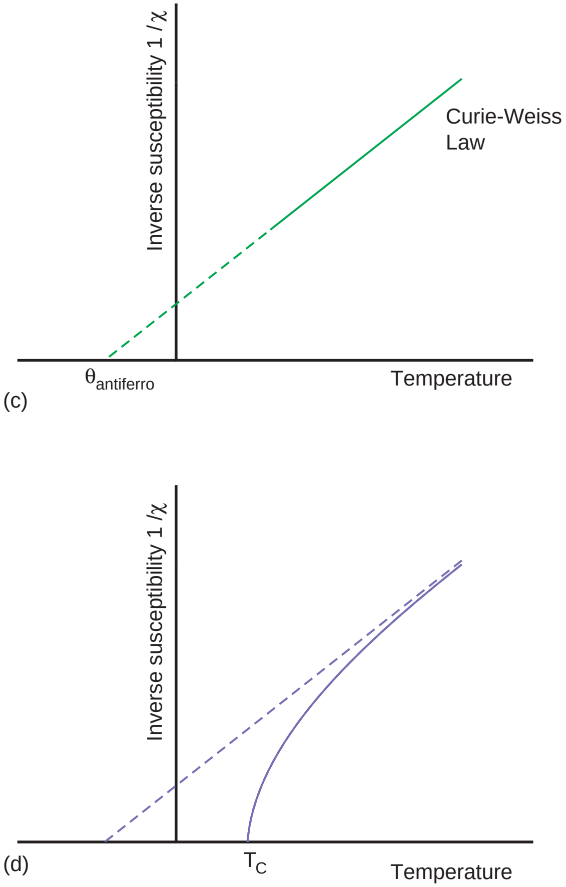


Figure 12.4 (*Continued*)

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*:

$$\chi = \frac{C}{T - \theta}$$

The Curie–Weiss constant, θ, is positive, has the dimensions of temperature, and a value usually close to, but not quite identical to, the Curie temperature, T_C (Figure 12.4b).

It is energetically favourable in some materials for the elementary magnetic dipoles to align in an antiparallel fashion (Figure 12.3h). 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 above. In this case, the Curie-Weiss constant, θ, is negative (Figure 12.4c).

Ferromagnetic ordering and antiferromagnetic ordering represent extremes of dipole orientation. In a number of solids, neighbouring magnetic dipoles are not aligned parallel to one another, but at an angle, referred to a *canted* arrangement (Figure 12.3i). Such canted arrangements can be thought of as an intermediate configuration between ferromagnetic and antiferromagnetic states. For example, a ferromagnetic ordering of magnetic dipoles arranged on a square lattice (Figure 12.5a) can be transformed by canting of alternate layers (Figure 12.5b) into an antiferromagnetic configuration (Figure 12.5c).

An important group of solids have two different magnetic dipoles present, one of greater magnitude than the other. When these line up in an antiparallel arrangement (Figure 12.3j) 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:

$$\frac{1}{\chi} = \frac{T}{C} + \frac{1}{\chi_0} + \frac{\xi}{T - \theta}$$

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/χ *versus* temperature is a straight line except near the Curie temperature T_C (Figure 12.4d).

Note that, with the exception of diamagnetic materials, these magnetic states are temperature-sensitive. Ordered magnetic dipole arrays usually

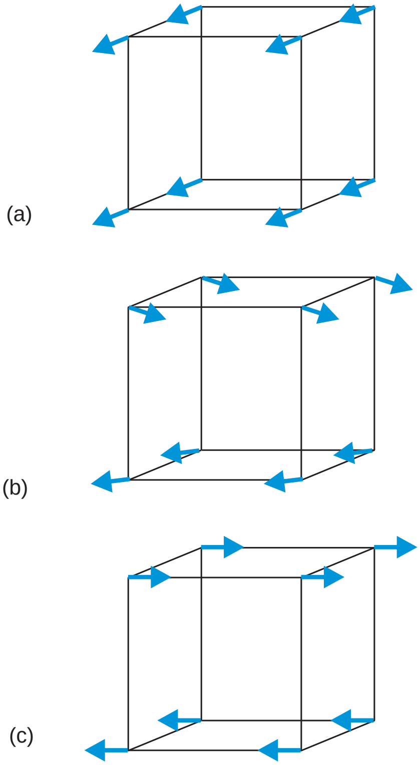


Figure 12.5 Ordered arrays of magnetic dipoles arranged on a square lattice: (a) ferromagnetic configuration; (b) canted antiferromagnetic configuration; (c) antiferromagnetic configuration.

give way to paramagnetic disorder at higher temperatures. All of these ordering patterns can be explained to some extent by models involving the overlap or interaction of atomic orbitals on neighbouring atoms (see below). However, the prediction of which of the several possibilities might actually be found at any temperature is difficult to determine. In recent years, however, computations, especially via density functional theory (Section 2.3.6), have been able to successfully account for the ground state magnetic structure of a number of solids, especially the important magnetic oxides. Additionally, although magnetic solids are generally thought to be inorganic compounds or metals, there is much current interest in organic and molecular magnetic materials.

12.2 Paramagnetic materials

12.2.1 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 equations (12.3) and (12.4).

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 of 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, are in good agreement with measurements (Table 12.1).

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, equation (12.4) reduces to:

$$g_J = 1 + \frac{S(S+1) + S(S+1)}{2S(S+1)} = 2$$

To a good approximation, equations (12.3) and (12.4) can then be replaced by a *spin-only formula*:

$$\mathbf{m} = 2\mu_B\sqrt{S(S+1)}$$

where S is the total spin quantum number, equal to $s_1 + s_2 + s_3 \ldots$ 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 $\frac{1}{2}$:

$$S = s_1 + s_2 + s_3 \ldots s_n = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} \ldots = \frac{n}{2}$$

Table 12.1 Calculated and observed magnetic dipole moments for the lanthanoids

Ion	Configuration	S	L	J	m_{calc}*	m_{meas}*
La^{3+}, Ce^{4+}	f^0	0	0	0	0	Diamagnetic
Ce^{3+}, Pr^{4+}	f^1	1/2	3	5/2	2.54	~2.5
Pr^{3+}	f^2	1	5	4	3.58	3.5
Nd^{3+}	f^3	3/2	6	9/2	3.62	3.5
Pm^{3+}	f^4	2	6	4	2.68	—
Sm^{3+}	f^5	5/2	5	5/2	0.84	1.5
Sm^{2+}, Eu^{3+}	f^6	3	3	0	0	3.4 (Eu^{3+})
Eu^{2+}, Gd^{3+}, Tb^{4+}	d^7	7/2	0	7/2	7.94	~8.0
Tb^{3+}	f^8	3	3	6	9.72	9.5
Dy^{3+}	f^9	5/2	5	15/2	10.63	10.6
Ho^{3+}	f^{10}	2	6	8	10.61	10.4
Er^{3+}	f^{11}	3/2	6	15/2	9.59	9.5
Tm^{3+}	f^{12}	1	5	6	7.57	7.3
Yb^{3+}	f^{13}	1/2	3	7/2	4.54	4.5
Yb^{2+}, Lu^{3+}	f^{14}	0	0	0	0	Diamagnetic

* The units are Bohr magnetons. The calculated values are from equation (12.3).

The spin-only formula can then be written:

$$\mathbf{m} = \mu_B\sqrt{n(n+2)} \qquad (12.5)$$

The magnetic dipole moments calculated in this way are generally in good agreement with observation (Table 12.2). In cases where the experimental value is considerably different to the spin-only value, orbital effects are presumed to be significant.

Table 12.2 The magnetic properties of the 3d transition metal ions

Ion	Configuration	S	$\mathrm{m}_{\mathrm{calc}}$*	$\mathrm{m}_{\mathrm{meas}}$*
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

* The units are Bohr magnetons. The calculated values are from equation (12.5).

12.2.2 High and low spin: crystal field effects

An important feature of 3d transition metal ions is that those ions with configurations between $3\mathrm{d}^4$ and $3\mathrm{d}^7$ have two apparent spin states, *high spin* (HS) and *low spin* (LS). For example, Fe^{3+}, with a $3\mathrm{d}^5$ 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 12.6). In an isolated atom or ion, the d orbitals all have the same energy. Suppose that the ion is now placed into an oxygen octahedron, made up by six negative O^{2-} ions (Figure 12.7a). 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

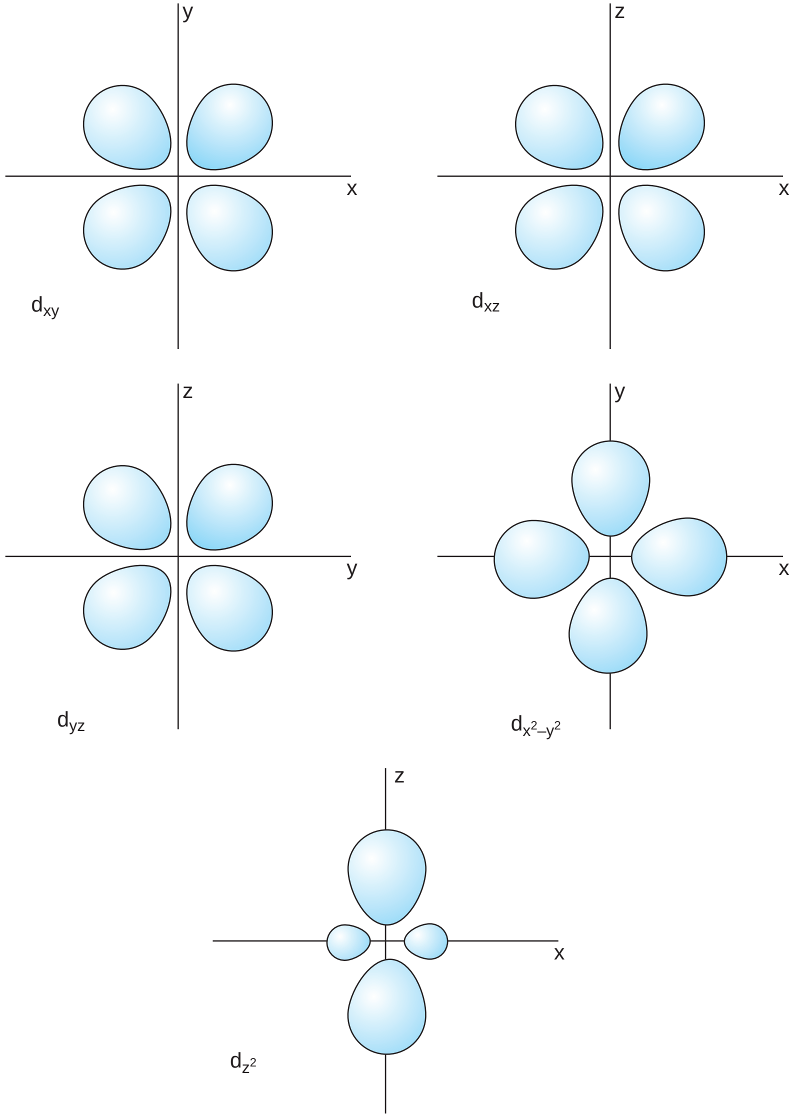


Figure 12.6 The five d orbitals. One group, d_{xy}, d_{xz} and d_{yz}, have lobes of electron density lying between the axes. The other group, $d_{x^2-y^2}$ and d_{z^2}, have electron density lying along the axes.

pointing directly towards the O^{2-} ions ($d_{x^2-y^2}$ and d_{z^2}) will be raised in energy most (Figure 12.8). 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 the first pair. This lower energy group, labelled t_{2g}, contains three orbitals of equal energy.

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

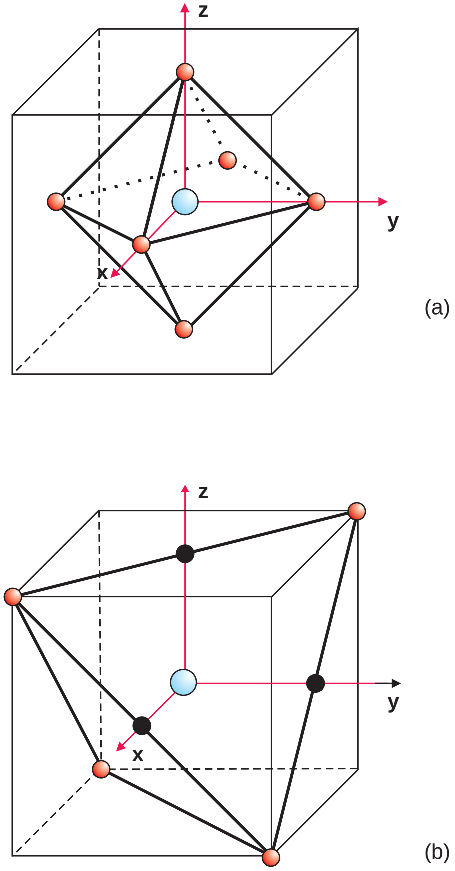


Figure 12.7 (a) A cation (large sphere) surrounded by six point charges arranged at the vertices of an octahedron experiences an octahedral crystal field. (b) A cation surrounded by four point charges arranged at the vertices of a tetrahedron experiences a tetrahedral crystal field.

charges are small and further away from the cation, the splitting is smaller.

The form of the crystal field splitting depends upon the geometry of the surrounding anions. For example, when a transition metal ion is surrounded by a tetrahedron of oxygen ions (Figure 12.7b), 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 12.9). In the case of the ions $3d^1$ to $3d^3$, the electrons keep apart and retain parallel spins, in accordance with Hund's rules (Section 1.2.2). 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 high-spin (HS) and low-spin (LS) configurations respectively. Low-spin and high-spin 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 to d^7 ions will thus depend upon whether they are in a high-spin or low-spin configuration.

This procedure can be repeated for ions in a tetrahedral geometry. However, only high-spin configurations are found because the crystal field splitting is always much smaller than that in octahedral geometry.

12.2.3 Temperature dependence of paramagnetic susceptibility

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 (Section 12.1.3):

$$M_j = J, J-1, J-2, \ldots -J$$

The $2J+1$ values of M_J give rise to a set of energy levels of separation:

$$\Delta E = g_J \mu_B B$$

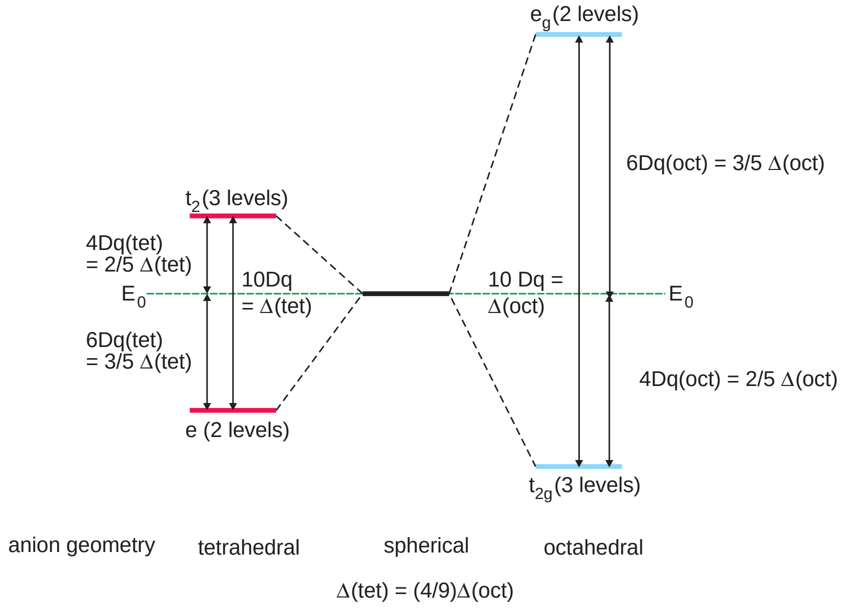


Figure 12.8 The crystal field splitting, Δ(tet) or 10Dq(tet), and Δ(oct) or 10Dq(oct), of the energy of the five d orbitals in tetrahedral and octahedral crystal fields, with respect to a spherical distribution of surrounding charges.

where g_J is the Landé g-factor, equation (12.4), and B is the magnitude of the magnetic flux density (Figure 12.10). 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. It is found that the value of $\mathbf{M}$ is:

$$\mathbf{M} = N g_J \mu_B J B_J(x) \qquad (12.6)$$

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:

$$x = \frac{g_J \mu_B \, J \, B}{k_B T}$$

[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:

$$\mathbf{M} = N\,\mathbf{m}\,L(x) = N\,\mathbf{m}\left(cothx - \frac{1}{x}\right)$$

where

$$x = \frac{mB}{k_B T}$$

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:

$$\mathbf{M} = \frac{N g_J^2 \mu_B^2 B J(J+1)}{3 k_B T}$$

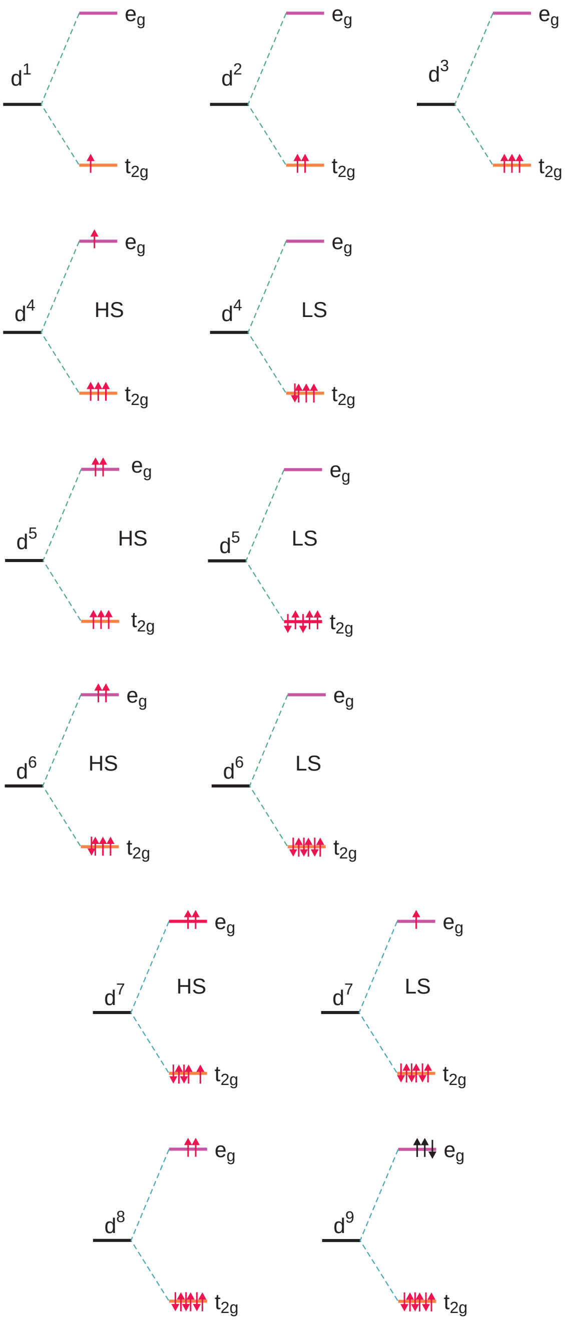


Figure 12.9 The electron configurations possible for d^n cations in an octahedral crystal field. For the ions d^4 to d^7, two configurations are possible, high spin (HS) or low spin (LS).

Substituting $\mathbf{B} = \mu_0\mathbf{H}$ for paramagnetic materials and rearranging gives:

$$\frac{\mathbf{M}}{\mathbf{H}} = \chi = \frac{N\mu_0 g_J^2 \mu_B^2 J(J+1)}{3k_B T}$$

where χ is the magnetic susceptibility. This has the same form as the Curie equation if the Curie constant is given by:

$$C = \frac{N\mu_0 g_J^2 \mu_B^2 J(J+1)}{3k_B}$$

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 $\chi_{mass} = \chi_v/\rho$, where ρ is the density.

The susceptibility of one mole of the substance, the molar susceptibility, χ_m, is given by $\chi_v V_m$, where V_m is the molar volume of the material.

12.2.4 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 the previous sections, 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 2.3). 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

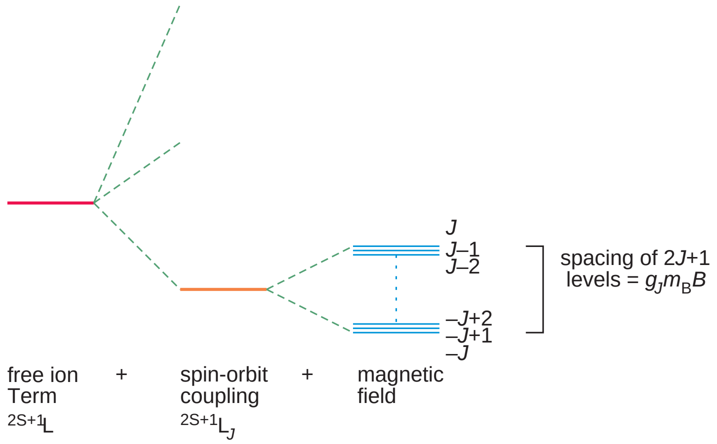


Figure 12.10 The splitting of atomic energy levels in a magnetic field: the Russell-Saunders free ion term (left) becomes more complex when the spin and orbital contributions couple (centre). In a magnetic field each spin-orbit level splits into a further $2J+1$ equally spaced levels.

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 12.11a). 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, that is, 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 12.11b).

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 vacuum 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.

12.3 Ferromagnetic materials

12.3.1 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 12.3). 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 20th 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 12.5).)

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

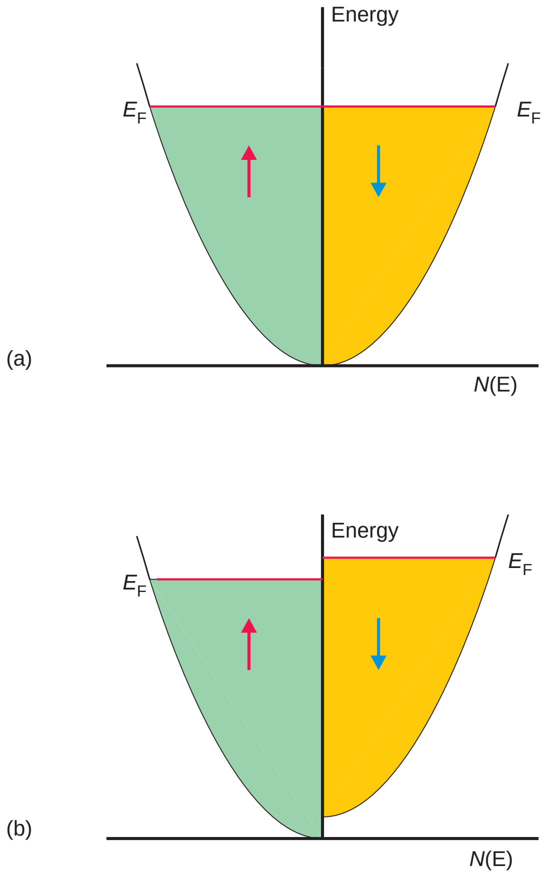


Figure 12.11 (a) The density of states for electrons in a metal: (a) in the absence of a magnetic field; (b) in a magnetic field.

Table 12.3 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

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 (12.1) 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 equation (12.6):

$$\mathbf{M} = Ng_J\mu_B JB_J(x)$$

but the Brillouin function, $B_J(x)$, is now a function of $\mathbf{M}$:

$$x = \frac{g_J\mu_B\,J\,B}{k_BT} = \frac{g_J\mu_B\,J\mu_0(\mathbf{H} + \lambda\mathbf{M})}{k_BT}$$

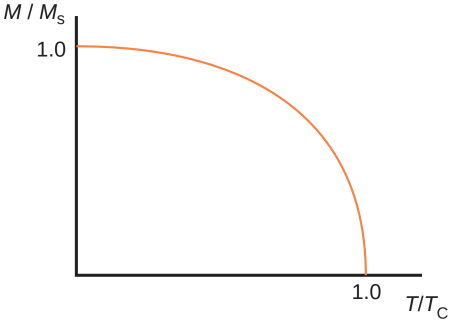


Figure 12.12 Variation of the relative spontaneous magnetisation, $\mathbf{M}/\mathbf{M}_s$, as a function of relative temperature, T/T_C. When $T = 0$ the magnetisation is equal to the saturation magnetisation, $\mathbf{M}_s$. When $T = T_c$, the Curie temperature, the spontaneous magnetisation is zero.

This equation cannot be solved analytically, and in the past graphical solutions have been utilised. Nowadays it is easier to use a computer to plot the results (Figure 12.12). There are three special cases to note.

1. $T = 0\,\text{K}$, $\mathbf{M} = \mathbf{M}_s = N g_J \mu_B J$
2. $T = T_C$, $\mathbf{M} = 0$

$$T_C = \frac{N\mu_0\, g_J^2 \mu_B^2\, J(J+1)\lambda}{3k_B} = \lambda C \qquad (12.7)$$

3. $T \gg T_C$

$$\chi = \frac{N\mu_0\, g_J^2 \mu_B^2\, J(J+1)}{3k_B(T - \lambda C)} = \frac{C}{(T - T_C)} \qquad (12.8)$$

The theory gives good agreement with the observations. In particular equation (12.7) 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 12.8 shows that ferromagnetic compounds obey the Curie–Weiss Law well above a transition temperature, T_C, at which the material loses its ferromagnetic properties.

12.3.2 Exchange energy

Three 3d transition metals, iron, cobalt and nickel, are ferromagnetic metals (Table 12.4). The band model (Section 2.3) leads to the expectation that all metals would show Pauli paramagnetism, so 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.

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, as far as possible.

Table 12.4 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 energy decrease due to the preference for parallel spins is called the *exchange energy*. This idea has been encountered as Hund's First Rule (Section 1.2.2), which says that 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 that 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 12.13). 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.

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 12.14. 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 12.4) 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

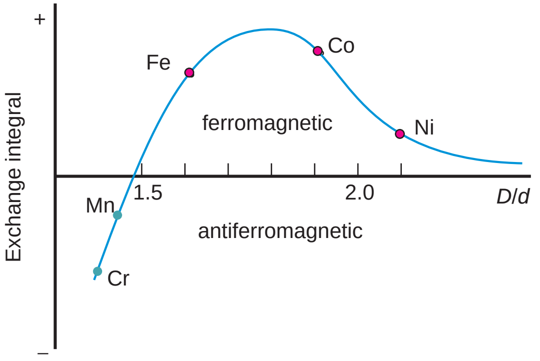


Figure 12.13 The Bethe–Slater curve for the magnitude of the exchange integral as a function of *D*/*d*. *D* is the separation of the atoms in a crystal and *d* the diameter of the 3d orbital.

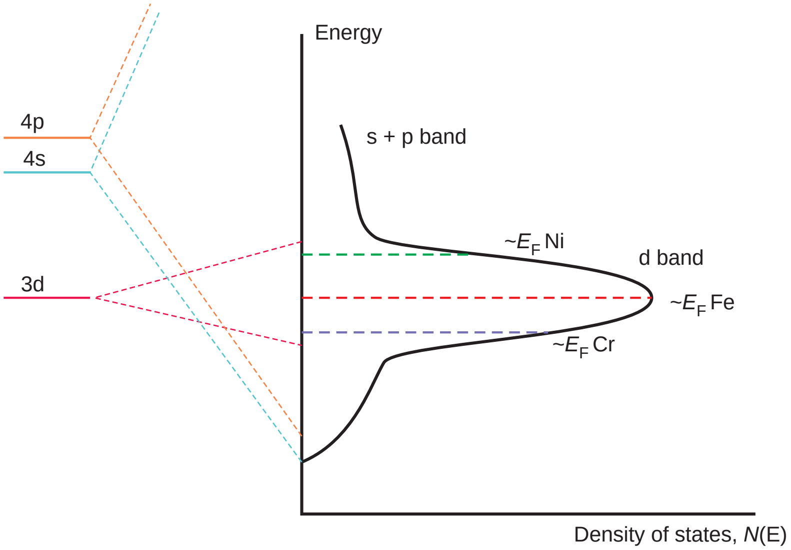


Figure 12.14 Schematic diagram of energy band overlap of 3d, 4s and 4p orbitals for the 3d transition metals.

electrons (1.72 holes), and of nickel contains 9.4 electrons (0.6 holes).

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 (Section 4.2.3). 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.

12.3.3 Domains

All ferromagnetic[1] substances remain magnetised to some extent after the external field is removed, and this feature characterises ferromagnetism. Some materials 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

[1] Ferromagnetic and ferrimagnetic materials (Section 12.5) 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.

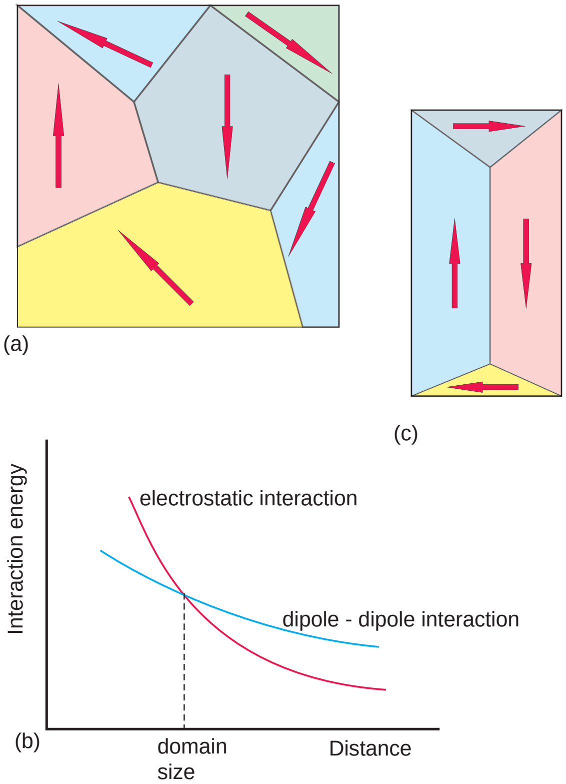


Figure 12.15 Ferromagnetic domains: (a) Weiss domains, schematic; (b) domain size reflects the balance between dipole–dipole and electrostatic interactions; (c) domain closure.

magnetisation (Figure 12.15a). In an apparently non-magnetic sample the alignment directions in neighbouring domains are different, so that the external net magnetic field is very small. (Note that domains are distinct from the grain structure of a metal and occur in single crystals as well as polycrystalline ferromagnetic solids.)

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 (Section 2.1.3). On the other hand, the interaction between the magnetic dipoles is a *long-range* weak bonding force (Section 3.1) 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 12.15b). The domains form as a balance between these two competing effects. The balance is achieved for domains of 1–100 micrometres 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 12.15c).

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 12.16).

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.

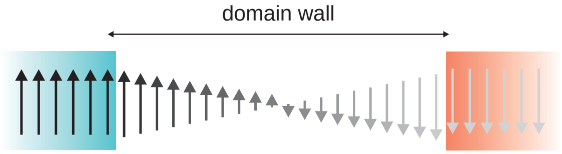


Figure 12.16 Schematic representation of a Bloch wall between two magnetic domains. On the far left, the magnetic dipoles point upwards. Moving to the right, they spiral through 180° until, at the far right, they point downwards.

12.3.4 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 reorient 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 versus magnetic field strength (Figure 12.17). As **H** increases, dipoles will start to gain sufficient energy to overcome this energy barrier and will be able to 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 (Figure 12.17). 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 (Figure 12.17). 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 12.17). 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 c*oercivity*, 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 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.

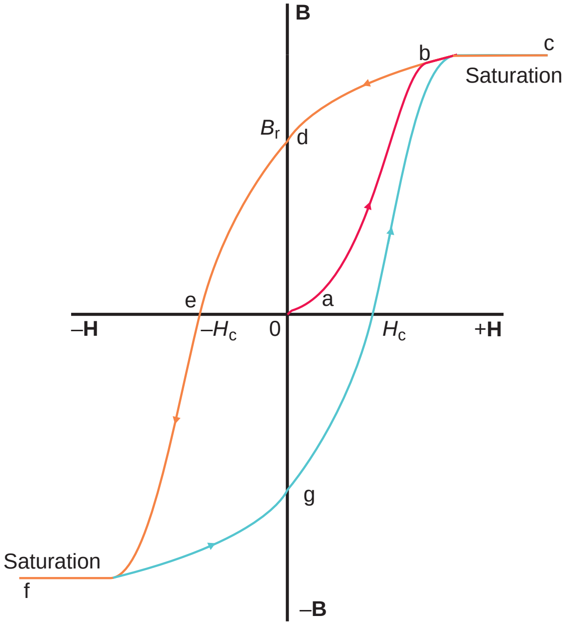


Figure 12.17 The **B**–**H** (hysteresis) loop of a ferromagnetic solid.

12.3.5 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 12.18). 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

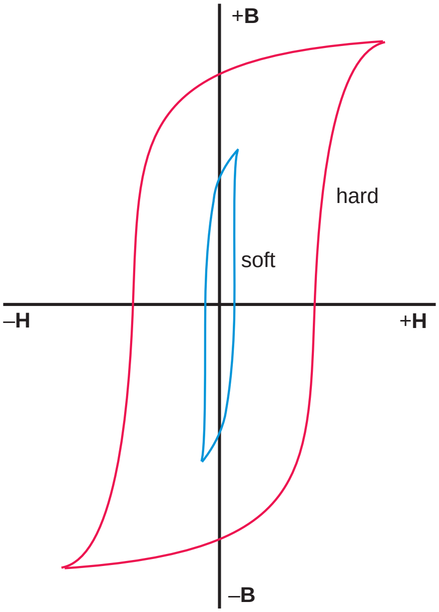


Figure 12.18 Schematic **B–H** loops for soft and hard magnetic solids.

with the cubic spinel structure (Section 12.5.1) 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 not only have to have carefully chosen compositions, but also need to be carefully fabricated in order to reproduce the desired magnetic performance.

Hard magnetic materials have rather rectangular broad hysteresis curves (Figure 12.18). These materials are used in permanent magnets, with applications from door catches to electric motors. The hexagonal ferrites (Section 12.5.3) 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 are stored by magnetising small volumes of material, often 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.

12.4 Antiferromagnetic materials and superexchange

Most antiferromagnetic materials are compounds in which transition metal ions are separated by a nonmetal, typified by oxides such as NiO, the cubic ferrites AFe_2O_4 (Section 12.5.1) and oxide perovskites such as $LaMnO_3$ and $LaCoO_3$ (Section 12.7.2). These 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. 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 (Figure 12.19). 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 on either side of an oxygen ion. In this case, as shown, the

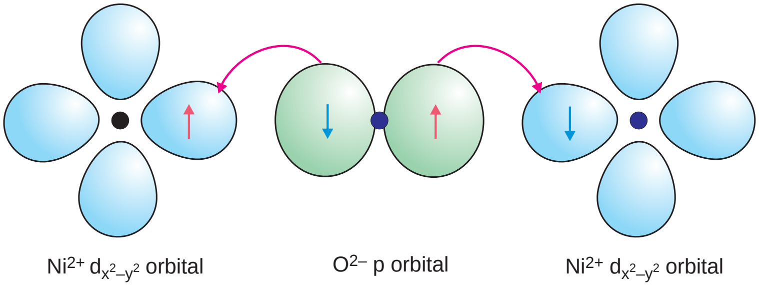

Ni^{2+} $d_{x^2-y^2}$ orbital
O^{2-} p orbital
Ni^{2+} $d_{x^2-y^2}$ orbital

Figure 12.19 Superexchange (electron coupling) between Ni^{2+} and O^{2-} (schematic) leading to antiferromagnetic alignment of spins on cations.

arrangement leads to an antiferromagnetic alignment of the unpaired d electrons on the Ni^{2+} ions (Figure 12.19). 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.

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 increases (Table 12.3). 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° angular 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.

12.5 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 12.3j). The observed magnetism is the due to the difference between the two sets of magnetic moments. As with antiferromagnetic compounds, they can only be understood in terms of both the crystal structures and the chemical bonding in the compounds.

12.5.1 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. The materials 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 5.3.9). 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 12.2 apply. Assuming all the excess A^{2+} moments are completely aligned:

$$\mathbf{M}_s = N\, m_{\text{ion}} \mu_B$$

where N is the number of unpaired dipoles per unit volume and $m_{\text{ion}}\ \mu_B$ 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, $A^{2+}B_2^{3+}O_4$, 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:

$$(1-x)NiFe_2O_4 + xZnFe_2O_4 \rightarrow Zn_xNi_{1-x}Fe_2O_4$$

The rather straightforward formula hides an intriguing situation. This is made clear if we write out the formula as

$$\begin{aligned}&xZn^{2+}\left[Fe^{3+}\downarrow\ Fe^{3+}\uparrow\right]O_4 + (1-x)Fe^{3+}\uparrow\\&\left[Fe^{3+}\downarrow\ Ni^{2+}\uparrow\right]O_4 \rightarrow Zn_x^{2+}Fe_{1-x}^{3+}\uparrow\\&\left[Fe^{3+}\downarrow\ Ni_{1-x}^{2+}\uparrow\ Fe_x^{3+}\uparrow\right]O_4\end{aligned}$$

When x is zero 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*. The distribution of cations in spinels is rarely perfectly normal or inverse, and the distribution tends to vary with temperature. 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.

The saturation magnetisation, $\mathbf{M}_s$, of these spinels can be estimated by assuming the spin-only magnetic moments listed in Table 12.2 apply. In these more complex cases, though, it is necessary to determine which spins cancel and which are aligned.

12.5.2 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 8-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.

12.5.3 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 12.20). There are ten of these layers in the unit cell, which contains two $BaFe_{12}O_{19}$ formula units. Using the normal description of close-packed layers (Section 5.4.1), the relative positions are:

$$\mathrm{A\,B'\,A\,B\,C\,A\,C'\,A\,C\,B}$$

where the symbols B′ and C′ represent BaO_3 layers, and the other symbols represent oxygen-only (O_4) layers. These ten 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 (Figure 12.20). (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.

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.

12.5.4 Double exchange

A number of ferrimagnetic ferrites are good electrical conductors and the same is true of the cobaltites and manganites (Section 12.7.2). Appreciable electrical conductivity implies that electrons are free 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 (Section 12.3.2). 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 below.) 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 12.21). 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

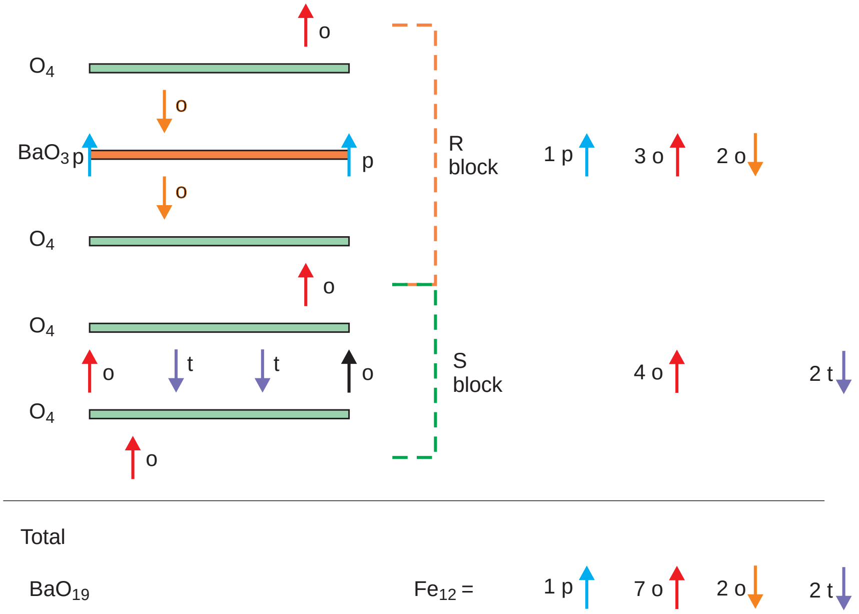


Figure 12.20 Schematic representation of the structure and spin arrangement of $BaFe_{12}O_{19}$ with the **c**-axis (c bold) vertical. The Fe ions are in octahedral (o), tetrahedral (t) and trigonal prismatic (p) sites. Note that the (o) sites on the RS boundaries count double as one site obscures a second and that the (p) sites count as 1/2 as these are on the unit cell edge.

another takes place, (quite unlike the situation in superexchange), and ferromagnetic alignment always results.

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 12.7.2).

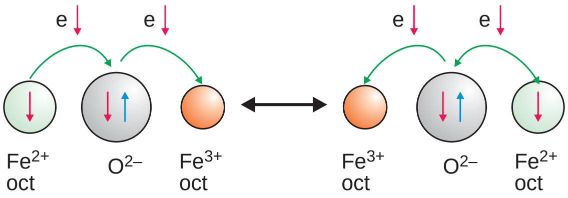


Figure 12.21 Double exchange (electron transfer) between Fe^{2+} and Fe^{3+} (schematic) is possible only with parallel alignment of spins on cations and always leads to a ferromagnetic alignment of dipoles.

To return to magnetite, double exchange is found not to operate between cations in the 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.

12.6 Nanostructures

12.6.1 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 11.3.11.)

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 forms the stored data. The data are 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.

12.6.2 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 (Figure 12.15b). 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 12.22a). 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 12.22b). 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 12.22c). The implications of these scale effects are of importance in attempts to increase the density of magnetic recording media.

12.6.3 Superlattices

Just as with ferroelectrics (Section 11.3.11), superlattices consisting of alternating layers, a few unit cells thick, of two or three perovskite-structure

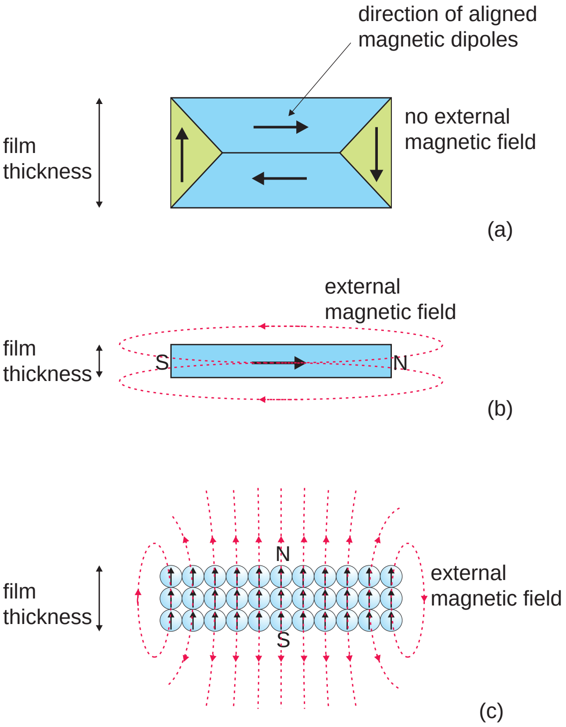


Figure 12.22 (a) Domain closure in a bulk film prevents magnetic flux from escaping. (b) Films less than a domain wide have magnetic dipoles aligned so the flux escapes longitudinally. (c) An atomically thin film has elementary dipoles aligned perpendicular to the film, allowing flux to escape normal to the film.

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 Section 11.3.11 and Figure 11.29).

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 12.23). $LaNiO_3$ contains the unusual valence state Ni^{3+}, a $3d^7$ ion. $LaMnO_3$ contains Mn^{3+}, a $3d^4$ ion with the electron

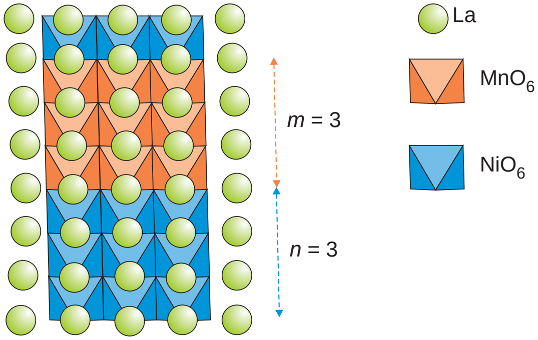


Figure 12.23 Part of an $m = n = 3$ magnetic $LaMnO_3/LaNiO_3$ superlattice in (111) orientation.

distribution $t_{2g}^3\ e_g^1$ (Figure 12.9). 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.

12.6.4 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 is responsible for the low-spin versus high-spin magnetic properties of these ions and 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 12.24a). Molecules in which this state of affairs can occur are known as *spin-crossover* (SCO) complexes or compounds.

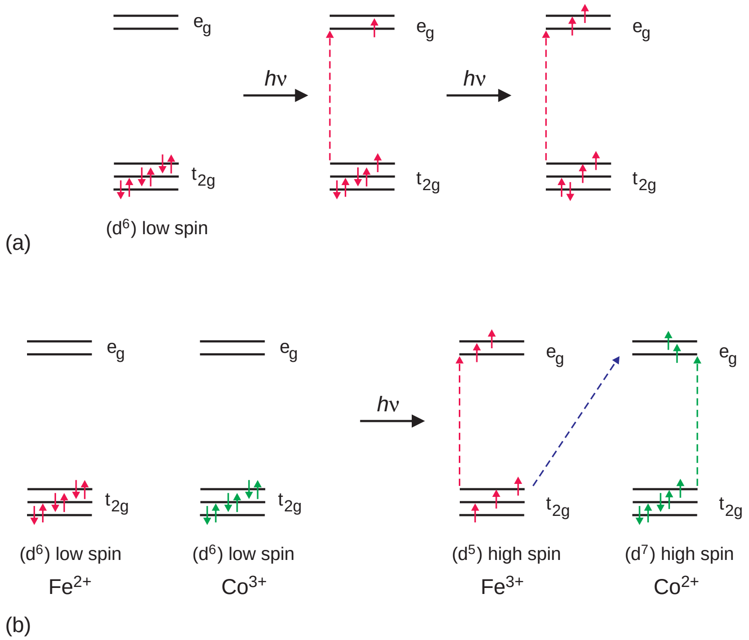


Figure 12.24 Photoinduced spin crossover in transition metal ions: (a) incident photons can successively promote electrons from the t_{2g} ground state to the higher-energy e_g state; (b) spin crossover involving electron excitation and transfer in $KFeCo(CN)_6$.

(Spin crossover can also be induced by heat or pressure, but light-induced transformations are most important.)

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 neither associated strongly with the high-spin nor low-spin 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:

$$\mathrm{Fe}^{2+}\left(t_{2g}^{6}\mathrm{LS}\right)\!-\!\mathrm{CN}\,-\!\mathrm{Co}^{3+}\left(t_{2g}^{6}\mathrm{LS}\right)\rightarrow$$
$$\mathrm{Fe}^{3+}\left(t_{2g}^{3}e_g^{2}\mathrm{HS}\right)\!-\!\mathrm{CN}\,-\!\mathrm{Co}^{2+}\left(t_{2g}^{5}e_g^{2}\mathrm{HS}\right)$$

In the initial state, all transition metal ions are in the low-spin state, giving an effectively non-magnetic solid. Irradiation transfers an electron from Fe to Co and also promotes electrons to create all high-spin ions with a considerable magnetic moment (Figure 12.24b). 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 low-spin to high-spin electron transfer mediated by light. Although the energy gap between the high-spin and low-spin 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.

12.7 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 a low-spin state cation is transformed into a high-spin state. In this section, point defects that considerably alter the magnetic properties of the host material are described.

12.7.1 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+}. The material most studied, (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, for example, 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 ferromagnetic or antiferromagnetic structures.

The magnetic properties of these doped semiconductors are, however, complicated and do not reside only in the magnetic impurities. 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 to 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 12.25a). 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:

$$r = \varepsilon_r \left(\frac{m_e}{m_e^*} \right)$$

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 12.25b).

The complex role of defects in these materials is underlined by noting that many undoped oxide

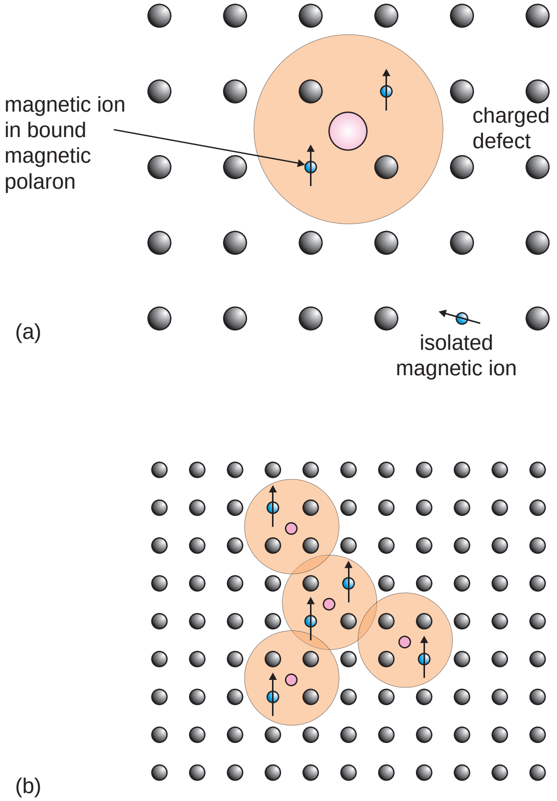


Figure 12.25 Schematic depiction of a bound magnetic polaron: (a) one bound magnetic polaron located on a charged defect; (b) overlapping bound magnetic polarons leading to ferromagnetic alignment of dipoles.

films, including TiO_2, HfO_2, ZnO, CeO_2, SnO_2, In_2O_3 and Al_2O_3, 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 gives rise to the ferromagnetic ordering observed.

12.7.2 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 important 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, for example, $LaBO_{3+\delta}$, 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 low-spin (LS) state, with an electron configuration t_{2g}^6 (Figure 12.26a). 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 12.26b), or $t_{2g}^4 e_g^2$, the high-spin (HS) state (Figure 12.26c). There is evidence that these spin states are ordered, especially when there are equal numbers of the two present. In the temperature range from 350–650 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.

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

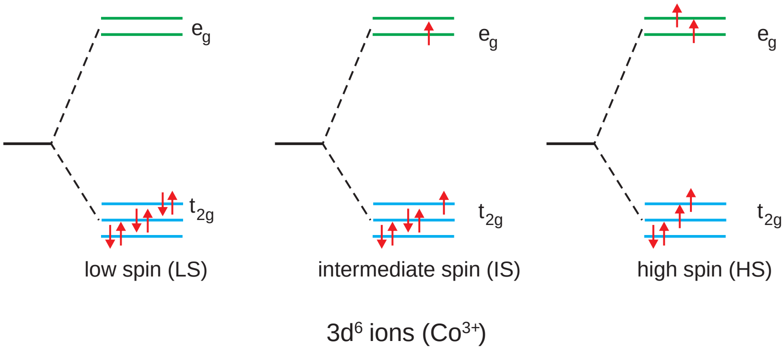


Figure 12.26 Electron configurations possible for Co^{3+} ($3d^6$) cations in an octahedral crystal field: (a) low-spin (LS); (b) intermediate spin (IS); (c) high-spin (HS).

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.

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+} or 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 semiconductor terminology, Chapter 13, this is called acceptor doping or hole doping.) 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 $t_{2g}^3e_g^2$. 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 a high-spin 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 a high-spin 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 occurs. 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 12.27).

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

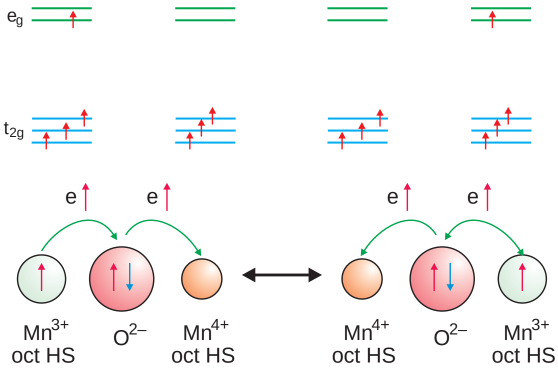


Figure 12.27 Double exchange between Mn^{3+} and Mn^{4+} high-spin (HS) ions in ferromagnetic $La_{1-x}Ca_xMnO_3$ (schematic).

as the change in resistance when a magnetic field is applied to that in zero field:

$$\text{MR ratio} = \frac{R_H - R_0}{R_0} = \frac{\Delta R}{R_0}$$

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 magnetoresistive 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 high-spin $t_{2g}^3e_g^1$ state, as are the Mn^{4+} ions, with an electron configuration $t_{2g}^3e_g^0$. The Mn^{2+} ions can adopt either a low-spin $t_{2g}^5e_g^0$

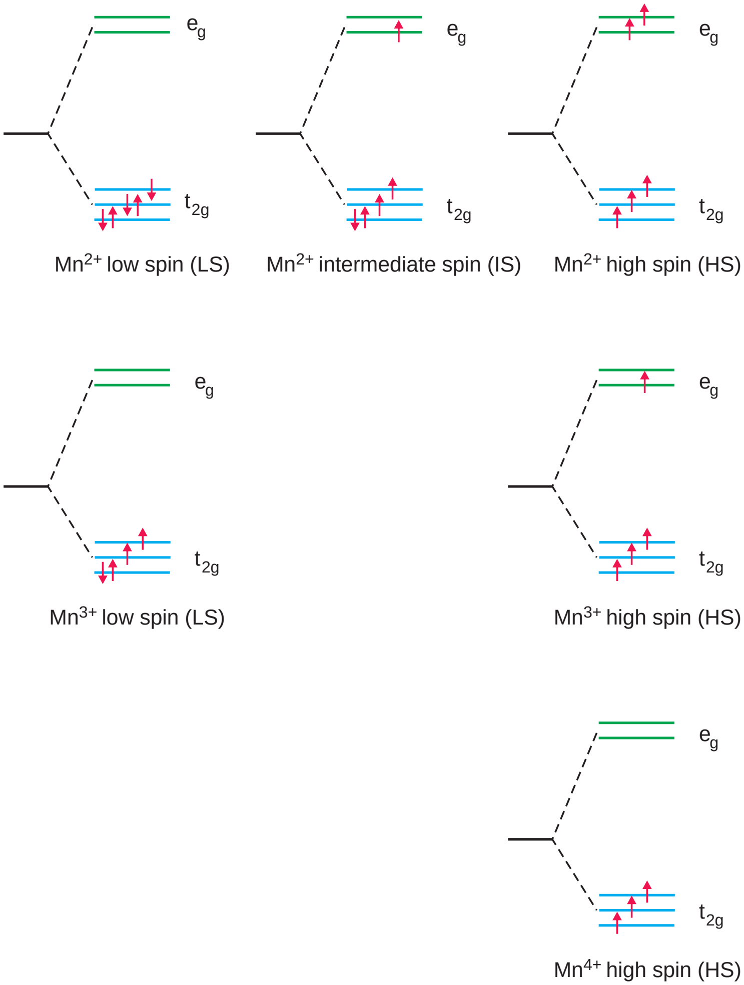


Figure 12.28 Electron configurations possible for Mn cations in an octahedral crystal field: low-spin (LS), intermediate spin (IS), high-spin (HS).

configuration, an intermediate spin $t_{2g}^4 e_g^1$ configuration, or a high-spin $t_{2g}^3 e_g^2$ configuration (Figure 12.28). The distribution between these states is temperature-sensitive, and at room temperature the intermediate spin 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 $Pr_{0.7}Ca_{0.2}Sr_{0.1}Mn_{0.98}Mg_{0.02}O_3$.

Neither the magnetic defect structures of these phases nor the mechanism of the CMR effect are yet 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:

Spaldin, N. (2003) *Magnetic Materials*. Cambridge University Press, Cambridge.

Epstein, A.J. (2003) Organic-based magnets: opportunities in photoinduced magnetism, spintronics, fractal magnetism and beyond. *Materials Research Society Bulletin*, **28**: 492.

Nanoparticles and defects:

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**: 278–82.

Sharrock, M.P. (1990) Particulate recording media. *Materials Research Society Bulletin*, **XV** (March): 53.

Judy, J.H. (1990) Thin film recording media. *Materials Research Society Bulletin*, **XV** (March): 63.

A starting point for the detection of magnetic fields by animals:

Eder, S.H.K., *et al.* (2012) Magnetic characterisation of isolated candidate vertebrate magnetoreceptor cell. *PNAS*, **109**: 12022–7. This paper details magnetic receptor cells in trout.

Winklhofer, M. (2012) An avian magnetometer. *Science*, **336**: 991–2.

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**: 239–99.

Magnetic superlattices:

Gilbert, M., *et al.* (2012) Exchange-bias in $LaNiO_3$–$LaMnO_3$ superlattices. *Nature Materials*, **11**: 195–8.

Yang, H.Y. *et al.* (2012) Emergent phenomena at oxide interfaces. *Nature Materials*, **11**: 103–13.

A starting point for studies on photomagnetism:

Dong, D.-P., *et al.* (2012) Photoswitchable dynamic magnetic relaxation in well-isolated $\{Fe_2Co\}$ double-zigzag chains. *Angew. Chem. Int. Ed.*, **51**: 5119–23.

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**: 2089–93.

Ohkoshi, S., *et al.* (2011) Light induced spin-crossover magnet. *Nature Chemistry*, **3**: 564–9.

Magnetites and colossal magnetoresistance (CMR):

Coey, J.M.D., Viret, M. and von Molnár, S. (1998) Mixed-valence manganites. *Adv. Physics*, **48**: 167–293.

Raveau, B., Maignan, A., Martin, C. and Hervieu, M. (1998) Colossal magnetoresistance manganite perovskites: relations between crystal chemistry and properties. *Chem. Mater.*, **10**: 2641–52.

Problems and exercises

Quick quiz

1 A material that is slightly attracted to a magnetic field is:
(a) Diamagnetic.
(b) Paramagnetic.
(c) Ferromagnetic.

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.

3 The Curie Law describes the magnetic behaviour of:
(a) Paramagnetic solids.
(b) Diamagnetic solids.
(c) Ferrimagnetic solids.

4 Above the Curie temperature, ferromagnetic materials become:
(a) Diamagnetic.
(b) Paramagnetic.
(c) Antiferromagnetic.

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.

6 The Néel temperature is the temperature at which an antiferromagnetic material becomes:
(a) Ferromagnetic.
(b) Ferrimagnetic.
(c) Paramagnetic.

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.

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.

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 The interaction of d-orbitals and oxygen atoms that leads to antiferromagnetic ordering is called:
(a) Exchange interaction.
(b) Superexchange.
(c) Double exchange.

11 The interaction between magnetic atoms in a ferrimagnetic material is called:
(a) Double exchange.
(b) Superexchange.
(c) Exchange interaction.

12 Ferrites with the spinel structure are:
(a) Paramagnetic materials.
(b) Ferromagnetic materials.
(c) Ferrimagnetic materials.

13 The general formula of hexagonal ferrites is:
(a) $A^{2+}Fe_{12}O_{19}$.
(b) $A^{4+}Fe_{12}O_{19}$.
(c) $A^{3+}Fe_{12}O_{19}$.

14 Hexagonal ferrites are:
(a) Paramagnetic materials.
(b) Ferromagnetic materials.
(c) Ferrimagnetic materials.

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.

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.

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.

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

12.1 Determine the ground state values of *S*, *J*, *L* and **m** for the f^3 ion Nd^{3+}.

12.2 Determine the ground state values of *S*, *J*, *L* and **m** for the f^7 ion Gd^{3+}.

12.3 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.

12.4 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.

12.5 Why don't the lanthanoid ions possess high-spin and low-spin states?

12.6 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; and (b) using the measured magnetic moment per nickel atom of $0.58\,\mu_B$. The metal has an A1 structure with a cubic unit cell parameter of 0.3524 nm.

12.7 Iron has a saturation magnetisation of $1.72 \times 10^6\,A\,m^{-1}$. What is the measured magnetic moment, in Bohr magnetons, of an iron atom? Iron has the A2 structure, with a cubic unit cell parameter of 0.2867 nm.

12.8 The saturation magnetisation of cobalt is $1.446 \times 10^6\,A\,m^{-1}$.

(a) Calculate the number of magnetic dipoles per unit volume in cobalt, knowing the effective magnetic moment per atom is $1.72\,\mu_B$.

(b) If it is assumed that there is one atom per primitive cubic unit cell, determine the length of the unit cell edge.

12.9 The magnetic moment of Fe^{3+} ions in the species $[Fe(H_2O)_6]^{3+}$ is $5.3\,\mu_B$. What is the likely electron configuration of the Fe^{3+} ion?

12.10 The magnetic moment of Fe^{3+} ions in the species $[Fe(CN)_6]^{3-}$ is $2.3\,\mu_B$. 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?

12.11 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 $[CoF_6]^{3-}$.

12.12 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 $[Fe(NH_3)_6]^{2+}$.

12.13 The crystal field splitting of diamagnetic $[Co(NH_3)_6]^{3+}$, containing Co^{3+} ions, is $4.57 \times 10^{-19}\,J$. What wavelength light would produce photoinduced paramagnetism in this molecule?

12.14 The crystal field splitting of diamagnetic $[Fe(CN)_6]^{4-}$, containing Fe^{2+} ions, is

6.54×10^{-19} J. What wavelength light would produce photoinduced paramagnetism in this molecule?

12.15 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?

12.16 Calculate (a) the paramagnetic energy level splitting for a Pr^{3+} ion in a magnetic flux density of 0.5 T; and (b) the corresponding wavelength of radiation for a transition between these energy levels.

12.17 Calculate (a) the paramagnetic energy level splitting for a Ho^{3+} ion in a magnetic flux density of 0.75 T; and (b) the corresponding wavelength of radiation for a transition between these energy levels.

12.18 Calculate: (a) 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; (b) the corresponding wavelength of radiation for a transition between these energy levels.

12.19 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) the corresponding wavelength of radiation for a transition between these energy levels.

12.20 Calculate the mass susceptibility of $NiSO_4.7H_2O$ 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.

12.21 Calculate the mass susceptibility of $CuSO_4.5H_2O$ 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.

12.22 Calculate the volume susceptibility of $MnSO_4.4H_2O$ 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 spinonly formula is adequate.

12.23 (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 12.2). The Curie law requires that $x \ll 1$.

(b) Estimate the temperature at which x is 0.01.

12.24 (a) Calculate the value of x in the Brillouin function $B_J(x)$, where:

$$x = \frac{g_J \mu_B JB}{k_B T}$$

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 12.2).

(b) Repeat the calculation assuming that the orbital angular momentum is not quenched and $L = 2$, $J = 3/2$. The Curie law requires that $x \ll 1$.

(c) Estimate the temperature at which x is 0.005.

12.25 Estimate: (a) the saturation magnetisation; (b) 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.

12.26 Estimate: (a) the saturation magnetisation; (b) 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.

12.27 Estimate: (a) the saturation magnetisation; (b) the magnetic inductance for the hexagonal ferrite ferroxdur, $BaFe_{12}O_{19}$. The hexagonal unit cell has parameters $a = 0.58778$ nm, $c = 2.1236$ nm. There are two formula units in the unit cell. Assume that the orbital angular momentum is quenched.

12.28 Estimate: (a) the saturation magnetisation; (b) the magnetic inductance for the hexagonal ferrite $SrFe_{12}O_{19}$. The hexagonal unit cell has parameters $a = 0.58836$ nm, $c = 2.30376$ nm. There are two formula units in the unit cell. Assume that the orbital angular momentum is quenched.

12.29 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 5.3.9).

13

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.

13.1 Metals

13.1.1 Metals, semiconductors and insulators

One of the defining physical properties of a metal is its electrical conductivity, defined via Ohm's Law:

$$V = IR$$

where V is the voltage applied to either end of 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:

$$R = \rho\left(\frac{L}{A}\right)$$

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:

$$\sigma = \frac{1}{\rho}$$

Electrical conductivity in a metal is due to electrons that are free to move, that is, to gain energy, under the influence of an applied voltage. Metallic bonding (Section 2.3) 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

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

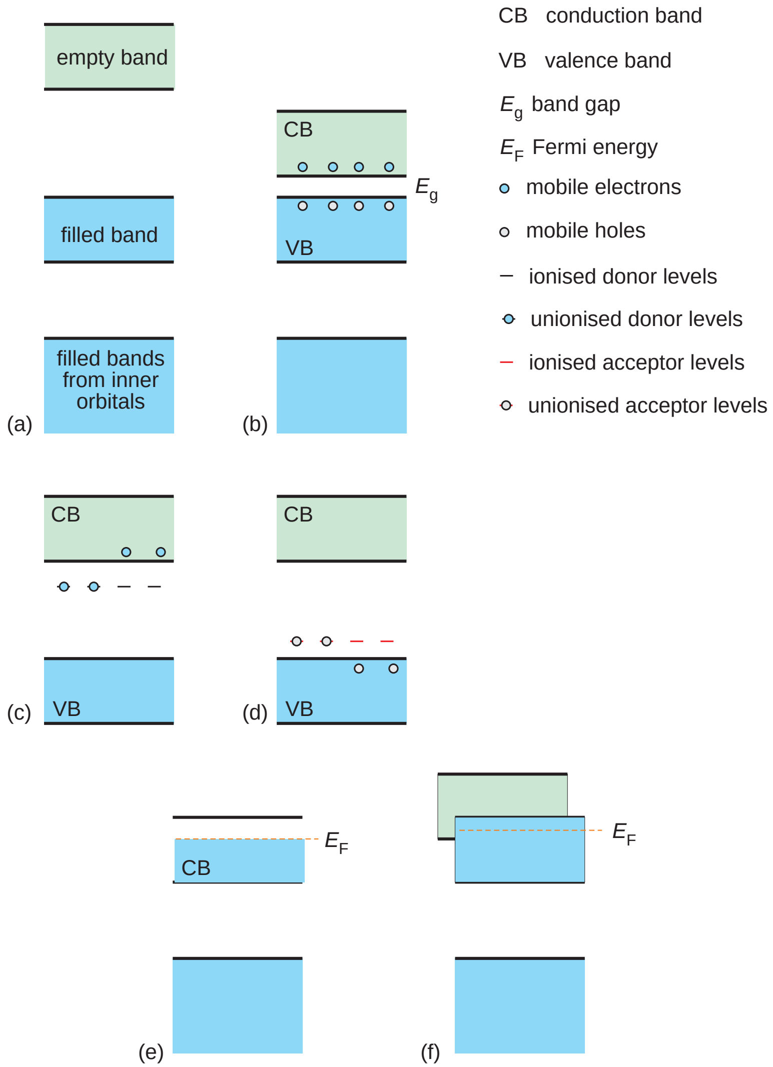


Figure 13.1 Energy band representations of materials: (a) insulators; (b) intrinsic semiconductors; (c) n-type extrinsic semiconductors; (d) p-type extrinsic semiconductors; (e) metals; (f) semimetals. The innermost filled energy bands are omitted in (c) and (d).

an *insulator* (Figure 13.1a). 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*.

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, the electronic properties change. Such materials are called *intrinsic semiconductors* (Figure 13.1b). 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 band is called the *valence band* and the almost-empty band is called the *conduction band*. 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 the 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 13.1c), or as *acceptors*, receiving electrons from the valence band, and thus contributing holes to the valence band, to form p-type semiconductors (Figure 13.1d). 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, small voltages will produce significant conductivity, and the solid is a *metal* (Figure 13.1e). 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 13.1f).

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 partly 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.

13.1.2 Electron drift in an electric field

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 (Section 2.3.2) is related to its wave vector **k** by a parabolic curve (Figure 13.2a). The wave vector can also be expressed in terms of the momentum or the velocity of the electron:

$$\mathbf{k} = \frac{2\pi}{\lambda} = \left(\frac{2\pi}{h}\right)\mathbf{p} = \left(\frac{2\pi m}{h}\right)\mathbf{v}$$

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 13.2b). When an electric field, **E**, is

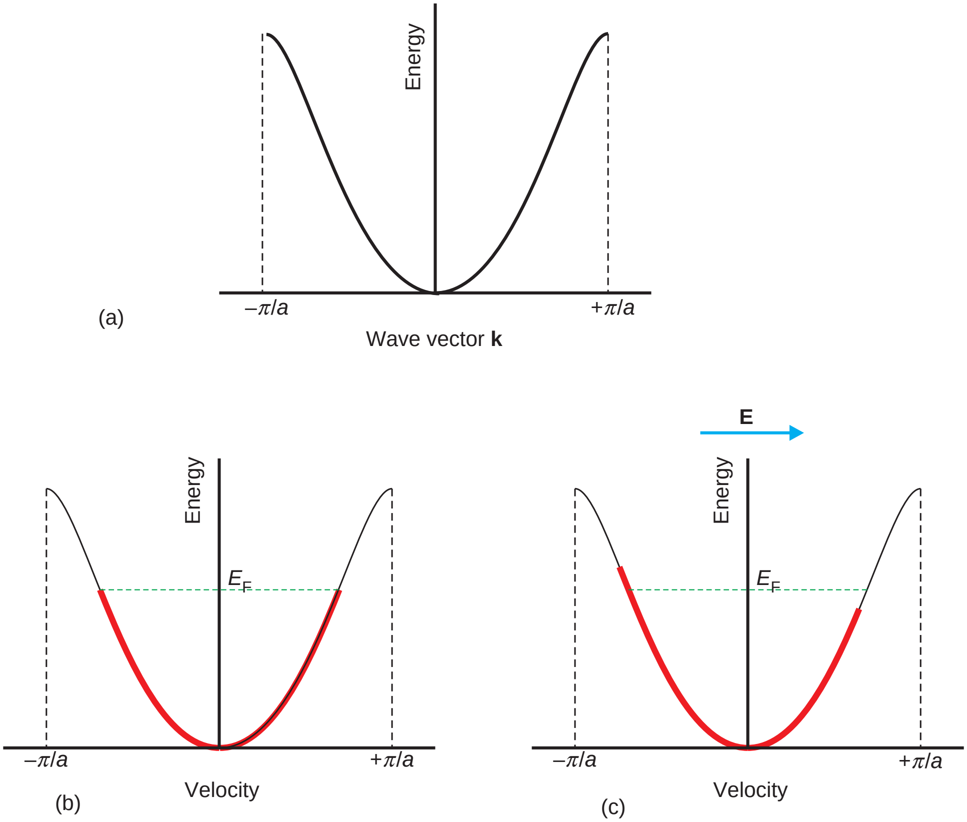


Figure 13.2 A free electron in a one-dimensional metal: (a) energy versus wave vector at equilibrium; (b) energy versus electron velocity at equilibrium; (c) energy versus electron velocity under an electric field **E**.

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 13.2c). The effect is limited by collisions with atoms, impurities and defects, and ultimately a steady current is reached.

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 13.3a). 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 13.3b). 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 $-x$.

13.1.3 Electronic conductivity

The magnitude of the current, I, that results from electron drift is defined as the amount of charge that

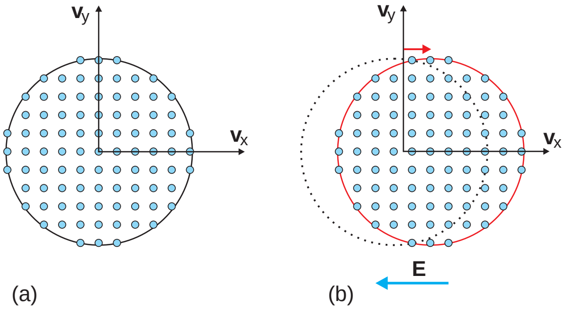


Figure 13.3 The circular Fermi surface of a two-dimensional metal marks the boundary up to which velocity states (dots) of free electrons are occupied: (a) in the absence of an electric field; (b) in an electric field E (solid circle compared with dotted circle).

passes through a unit cross-sectional area of the conductor per second:

$$I = n\,e\,A\,\mathrm{v}$$

where n is the number of mobile electrons per unit volume, A is the cross-sectional area of the conductor and v is the drift velocity. The force exerted on an electron by an electric field, E, is given by $-\mathrm{e}E$, and the acceleration, a, imposed on an electron is then given by:

$$a = \frac{e\,E}{m_{\mathrm{e}}^{*}} \qquad (13.1)$$

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, $\mathrm{m_e}$, i.e. 0.067 $\mathrm{m_e}$.)

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 after the voltage is turned off, where τ is the time between successive collisions, called the *relaxation time*. The *mean free path* of the electron, which is the length of the path between successive collisions, is given by:

$$\Lambda = \tau\,\mathrm{v_F}$$

where $\mathrm{v_F}$ is the electron velocity at the Fermi surface. In addition, the drift velocity of the electrons is:

$$\mathrm{v} = a\,\tau$$

Substituting for the acceleration from equation (13.1):

$$\mathrm{v} = \frac{e\,E\,\tau}{m_{\mathrm{e}}^{*}} \qquad (13.2)$$

The total current flowing is therefore:

$$I = \frac{n\,e^2 A\,E\,\tau}{m_{\mathrm{e}}^{*}}$$

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:

$$I = \left(\frac{n\,e^2\,\tau}{m_{\mathrm{e}}^{*}}\right)\left(\frac{AV}{L}\right)$$

Ohm's law can be written:

$$I = \frac{AV}{L\,\rho}$$

so that:

$$\frac{1}{\rho} = \sigma = \frac{n\,e^2\,\tau}{m_{\mathrm{e}}^{*}} \qquad (13.3)$$

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:

$$\mathrm{v} = \mu \, E$$

Comparing this with equation (13.2) makes it apparent that:

$$\mu = \frac{e\,\tau}{m_{\mathrm{e}}^{*}}$$

This is sometimes called the *drift mobility*, to distinguish it from mobility measured via the Hall effect (Section 13.2.3). The conductivity and the mobility are then related by substituting equation (13.3) into (13.2) to give:

$$\sigma = n\, e\, \mu_{\mathrm{e}}$$

13.1.4 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 13.4), except for the superconductors (Section 13.4). 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.

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:

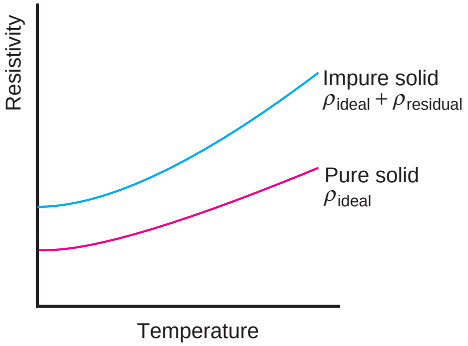


Figure 13.4 The variation of the resistivity of a metal with temperature (schematic).

$$\tau_{\mathrm{th}} = \frac{\Lambda_{\mathrm{th}}}{\mathrm{v}_{\mathrm{F}}}$$

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:

$$\rho = \frac{m_{\mathrm{e}}^{*}}{e^{2} n \tau_{\mathrm{th}}} + \frac{m_{\mathrm{e}}^{*}}{e^{2} n \tau_{\mathrm{def}}} + \frac{m_{\mathrm{e}}^{*}}{e^{2} n \tau_{\mathrm{imp}}}$$

Taking into account these features, the total resistivity can be written as:

$$\rho = \rho_{\mathrm{th}} + \rho_{\mathrm{d}} + \rho_{\mathrm{i}}$$

This is known as *Mattiesen's rule*, sometimes written as:

$$\rho = \rho_{\mathrm{ideal}} + \rho_{\mathrm{residual}}$$

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

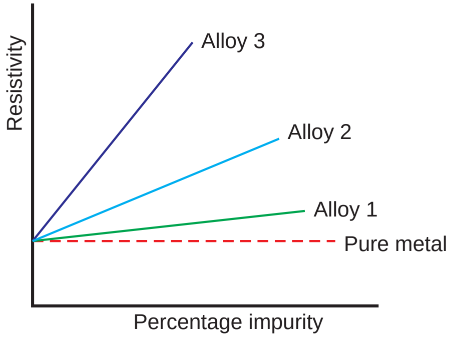


Figure 13.5 The variation of the resistivity of alloys with concentration of the alloying elements (schematic).

metal because the elements added to form the alloy have the same effect as impurities (Figure 13.5). 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.

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.

13.2 Semiconductors

13.2.1 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 13.6). 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, not the magnitude of the conductivity. (Remember that for metals, conductivity falls with increasing temperature.)

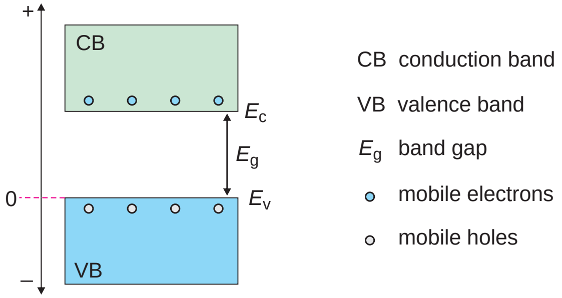


Figure 13.6 The simplest (flat-band) energy band description of an intrinsic semiconductor.

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 (Section 2.3.2). 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:

$$n = N_c \exp\left(\frac{-(E_g - E_F)}{k_B T}\right)$$

where E_F is the Fermi energy and N_c is the effective density of states function in the conduction band, given by:

$$N_c = 2\left(\frac{2\pi\, m_e^* k_B T}{h^2}\right)^{3/2}$$

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:

$$p = N_v \exp\left(\frac{-E_F}{k_B T}\right)$$

where N_v is the effective density of states of states function in the valence band, given by:

$$N_v = 2\left(\frac{2\pi\, m_h^* k_B T}{h^2}\right)^{3/2}$$

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:

$$N_c \exp\left(\frac{-(E_g - E_F)}{k_B T}\right) = N_v \exp\left(\frac{-E_F}{k_B T}\right)$$

Taking logarithms and rearranging allows the Fermi energy to be written as:

$$\begin{aligned} E_F &= {}^1/_2\, E_g + {}^1/_2 k_B T \ln\left(\frac{N_v}{N_c}\right) \\ &= {}^1/_2 E_g + {}^3/_4 k_B T \ln\left(\frac{m_h^*}{m_e^*}\right) \end{aligned}$$

In the case when the effective masses are identical:

$$E_F = {}^1/_2 E_g$$

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:

$$\begin{aligned} n_i &= N_c \exp\left(\frac{-E_g}{2k_B T}\right) \\ p_i &= N_v \exp\left(\frac{-E_g}{2k_B T}\right) \\ n_i p_i &= N_c N_v \exp\left(\frac{-E_g}{k_B T}\right) \end{aligned}$$

As $n_i = p_i$:

$$\begin{aligned} n_i = p_i &= \sqrt{N_c N_v} \exp\left(\frac{-E_g}{2k_B T}\right) \\ &= 2\left(\frac{2\pi\, k_B T}{h^2}\right)^{3/2} (m_e^* m_h^*)^{3/4} \exp\left(\frac{-E_g}{2k_B T}\right) \end{aligned}$$

Inserting values for the constants, and separating the effective mass of electrons and holes, gives:

$$n_i = p_i = 4.826 \times 10^{21} \left(\frac{m_e^* m_h^*}{m_e^2}\right)^{3/4} T^{3/2} \exp\left(\frac{-E_g}{2k_B T}\right) \qquad (\text{units m}^{-3})$$

Writing the total conductivity, σ, as:

$$\begin{aligned} \sigma &= n\, e\, \mu_e + p\, e\, \mu_h \\ &= 4.826 \times 10^{21} \left(\frac{m_e^* m_h^*}{m_e^2}\right)^{3/4} T^{3/2}\, e(\mu_e + \mu_h) \exp\left(\frac{-E_g}{2k_B T}\right) \\ &= 773.1 \left(\frac{m_e^* m_h^*}{m_e^2}\right)^{3/4} T^{3/2}\, (\mu_e + \mu_h) \exp\left(\frac{-E_g}{2k_B T}\right) \end{aligned}$$

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.

13.2.2 Band gap measurement

The total conductivity, σ, can be expressed in the form:

$$\sigma = \sigma_0 \exp\left(\frac{-E_g}{2k_B T}\right)$$

where σ_0 is a constant, E_g is the band gap, k_B is Boltzmann's constant and T the absolute temperature. Taking logarithms:

$$\ln \sigma = \ln \sigma_0 - \frac{E_g}{2k_B T}$$

The gradient of a plot of conductivity versus $1/T$ will therefore yield a value for the *thermal* band gap (Figure 13.7).

An alternative method for 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 from 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 13.1). 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. 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

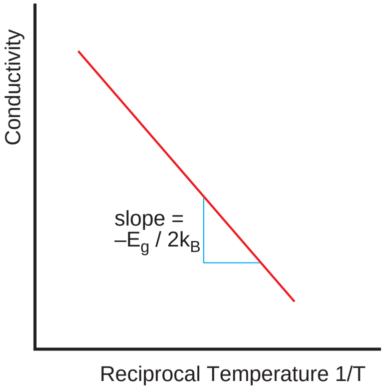


Figure 13.7 The variation of resistivity versus reciprocal temperature for an intrinsic semiconductor.

Table 13.1 Approximate values for the band gap of some semiconductors

Compound	Formula	Band gap/eV*
Elements		
Diamond	C	5.47
Silicon	Si	1.12
Germanium	Ge	0.66
Grey tin	Sn	0.08
III–V semiconductors		
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

*The band gap is normally given in electron-volts in most compilations. 1 eV is equal to 1.60219×10^{-19} J.

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.

13.2.3 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 (Section 5.3.6). 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 13.8a). 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.

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:

$$E = \frac{-m_e e^4}{8\varepsilon_0^2 h^2}$$

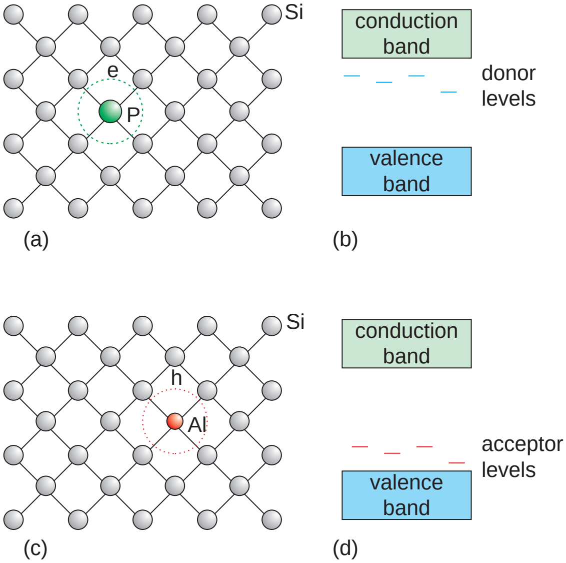


Figure 13.8 Impurities in a crystal of an extrinsic semiconductor: (a) donor (P) atoms; (b) donor energy levels below the conduction band; (c) acceptor (Al) atoms; (d) acceptor energy levels above the valence band.

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:

$$E(\mathrm{P}) = \frac{-m_e^* e^4}{8\varepsilon_0^2\, \varepsilon_r^2\, h^2} = \frac{E\, m_e^*}{m_e\, \varepsilon_r^2}$$

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 13.8b).

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 13.8c). Provided 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 13.8d).

The energy to liberate both donor electrons and acceptor holes is about 8×10^{-21} J (0.05 eV). These values are comparable to 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*.

13.2.4 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:

$$K = n\,p = n_\mathrm{i}^2$$

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.

The 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 so that 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 is at the mid-point (Figure 13.9a). 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 to maintain the balance (Figure 13.9b). In a p-type semiconductor the number of holes is much higher than the number of electrons and the Fermi energy moves towards the p-type side to maintain the balance (Figure 13.9c).

The number of holes created by the addition of acceptors is equal to $N_\mathrm{a}{}^-$, 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 $N_\mathrm{a}{}^-$. In this case, at ordinary temperatures, the Fermi energy moves so as to lie approximately halfway between the top of the valence band and the acceptor energy levels (Figure 13.10a) Similarly, the number of electrons created by the addition

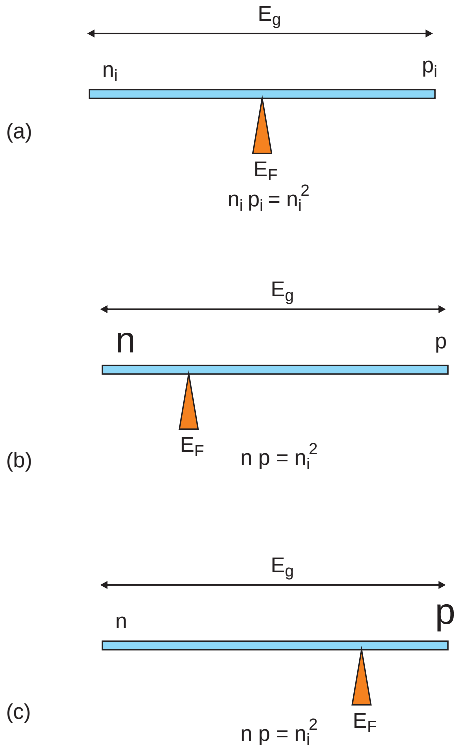


Figure 13.9 The position of the Fermi energy in semiconductors varies with dopant concentration so as to always maintain the relationship *np* as constant: (a) an intrinsic semiconductor; (b) an n-type semiconductor; (c) a p-type semiconductor.

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 now moves so as to lie approximately halfway between the bottom of the conduction band and the donor energy levels (Figure 13.10b).

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:

$$n + N_a = p + N_d$$

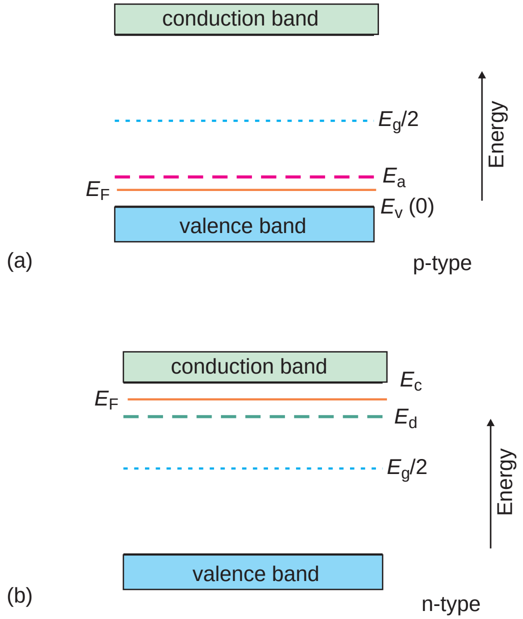


Figure 13.10 The position of the Fermi energy in (a) a p-type semiconductor, and (b) an n-type semiconductor, at low temperatures.

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:

$$n\ (\text{n-type}) \sim N_d$$
$$p\ (\text{n-type}) \sim \frac{n_i^2}{N_d}$$

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:

$$p\ (\text{p-type}) \sim N_a$$
$$n\ (\text{p-type}) \sim \frac{n_i^2}{N_a}$$

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 13.11). 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.

13.2.5 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 (Figure 13.12). 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:

$$\rho = 2\pi\, s\left(\frac{V}{I}\right)$$

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 13.12a), or $d \gg s \gg t$ (Figure 13.12b), the expression for the resistivity is:

$$\rho = \frac{\pi\, t}{\ln 2}\left(\frac{V}{I}\right) = 4.54t\frac{V}{I}$$

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

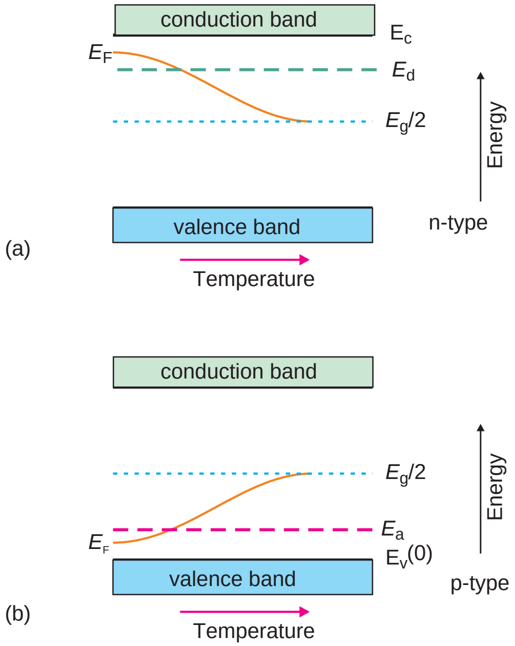


Figure 13.11 The variation of the position of the Fermi energy of (a) an n-type and (b) a p-type semiconductor with temperature.

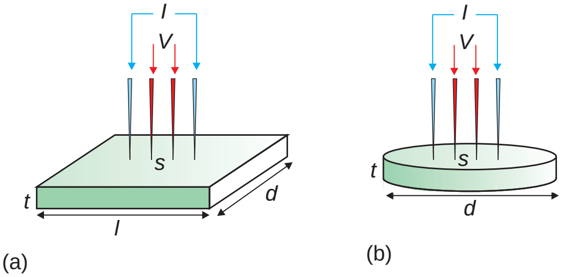


Figure 13.12 Arrangement of the four-point probe method of measurement of resistance of semiconductors, schematic: (a) a rectangular specimen; (b) a disc.

ρ, and the dimensions of the sample. For a rectangular specimen:

$$R_s = \frac{\rho}{t}; \quad \rho = \frac{Rwt}{l}$$

hence

$$R_s = \frac{Rw}{l}; \quad \rho = R_s\, t$$

The sheet resistance is quoted as ohms per square, $\Omega/\square$.

The carrier type can be found in two ways, via the *Hall effect* or the *Seebeck effect* (Section 15.4). 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 13.13a). 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 I 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 counterintuitively, 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.

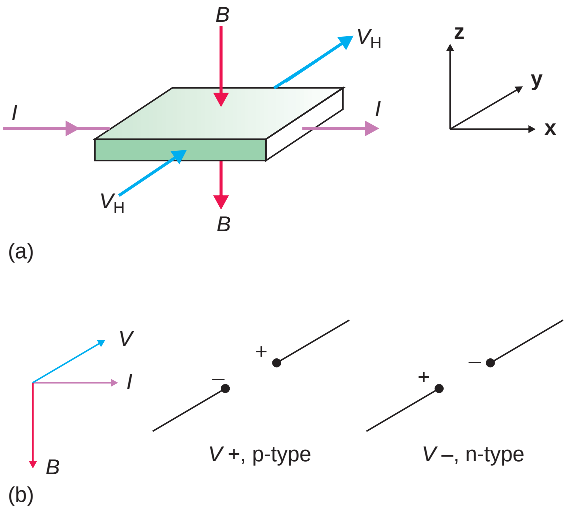


Figure 13.13 Measurement of the Hall effect, schematic: (a) axes used; (b) the sign of the Hall voltage with respect to the direction of the current and magnetic induction.

The relationship between the current, magnetic induction and electric field is:

$$\mathbf{E}_y = \pm R_H\, \mathbf{J}_x\, \mathbf{B}_z \tag{13.4}$$

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 13.13a), 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 13.13b).

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 $\mathbf{J}_x$ into equation (13.4):

$$R_{\mathrm{H}} = \frac{-1}{n\,e}$$

where n is the number of electrons per unit volume, each with a charge of $-e$. For a flow of positive holes:

$$R_{\mathrm{H}} = \frac{1}{p\,e}$$

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, $\mathrm{R_H}$, it is possible to determine the number of charge carriers and their mobility, as:

$$\begin{aligned}\sigma_{\mathrm{e}} &= n\,e\,\mu_{\mathrm{e}}\\ -R_{\mathrm{H}}\,\sigma_{\mathrm{e}} &= \mu_{\mathrm{e}}\end{aligned}$$

or:

$$\begin{aligned}\sigma_{\mathrm{h}} &= p\,e\,\mu_{\mathrm{h}}\\ +R_{\mathrm{H}}\,\sigma_{\mathrm{h}} &= \mu_{\mathrm{h}}\end{aligned}$$

If both electrons and holes are present:

$$\begin{aligned}\sigma\,(\text{total}) &= n\,e\,\mu_{\mathrm{e}} + p\,e\,\mu_{\mathrm{h}}\\ &= |e|\,(n\,\mu_{\mathrm{e}} + p\,\mu_{\mathrm{h}})\end{aligned}$$

and:

$$R_{\mathrm{H}} = \frac{(n\,\mu_{\mathrm{e}}^2 + p\,\mu_{\mathrm{h}}^2)/(n\,\mu_{\mathrm{e}} + p\,\mu_{\mathrm{h}})^2}{|e|}$$

The mobility derived from a Hall measurement is often called the *Hall mobility* to distinguish it from the drift mobility (Section 13.1.3).

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 (Section 2.3.2). In some metals, such as Er and Ho, the Hall coefficient varies from negative to positive as a function of crystallographic direction. The explanation of 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 used here.

13.2.6 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 to 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 13.14a,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 13.14c). To achieve this, the energy levels have been shifted vertically with respect to each other to give a distorted band structure in the

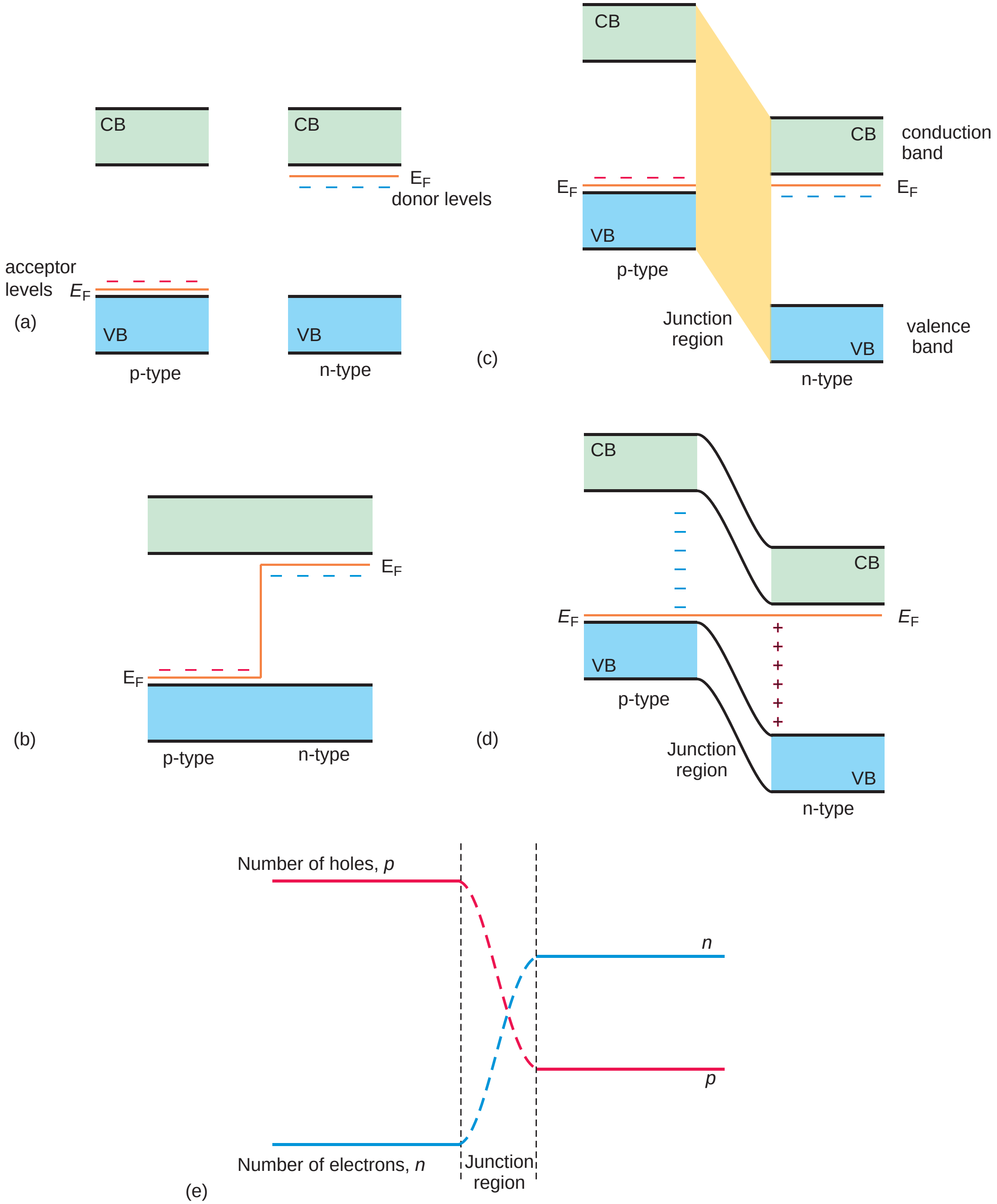


Figure 13.14 The p-n junction, schematic: (a) energy bands of separated p-type and n-type materials; (b) energy bands for juxtaposed p-type and n-type materials; (c) energy bands at equilibrium; (d) distorted energy bands in the junction region at equilibrium; (e) electron and hole numbers across the junction region at equilibrium.

junction region (Figure 13.14d). This means that the populations of electrons and holes change dramatically as the junction is traversed (Figure 13.14e).

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 13.15a). 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 to cross it. Current flow now virtually ceases (Figure 13.15c).

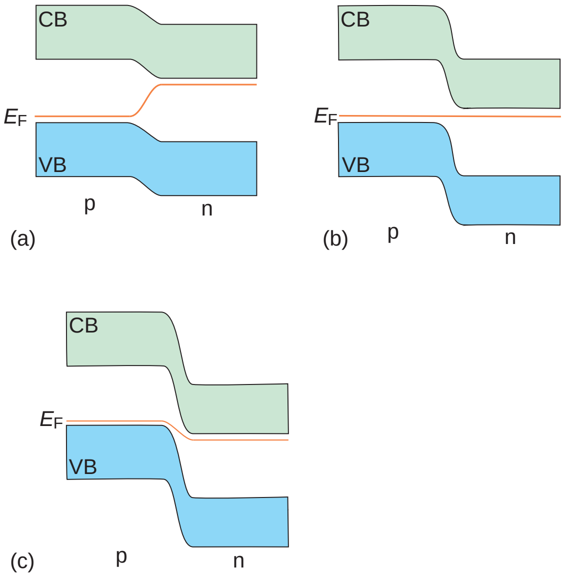


Figure 13.15 Band structure across a p-n junction: (a) under forward bias; (b) no bias; (c) under reverse bias.

The change of current with applied voltage is given by the *Shockley equation*. In a simplified form this is:

$$I = I_0\left[\exp\left(\frac{eV}{k_BT}\right) - 1\right]$$

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 13.16).

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

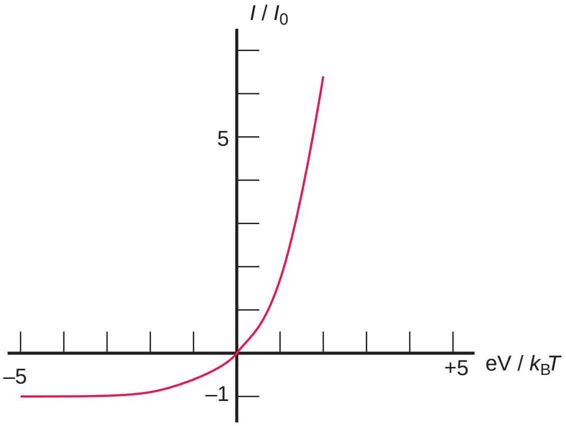


Figure 13.16 The ideal current–voltage curve for a p-n junction.

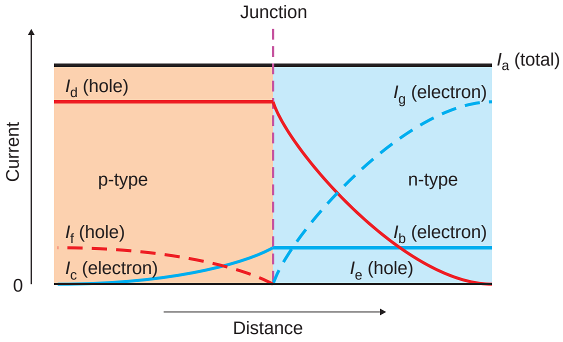


Figure 13.17 The currents flowing across a p-n junction under forward bias.

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 moved 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 13.17):

$$I_a = I_b + I_e + I_g = I_c + I_d + I_f$$

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 the p-type region (decaying); I_d is the hole current in the p-type region (constant); I_e is the injected hole current in the n-type region (decaying); I_f is the declining hole current in the p-type region to balance and annihilate I_c; and I_g is the declining electron current in the n-type region to balance and annihilate I_e. This is quite different to 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.

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.

13.3 Metal–insulator transitions

13.3.1 *Metals and insulators*

The simple theory of metals indicates that metallic conduction arises when the highest energy band is only part-filled with electrons. 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 versus insulating states 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 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 explain 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 with insulating CoO and NiO.

A second suggestion might be that the antiferromagnetic ordering in, for example, NiO (Section 12.4) 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_{2g} 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_{2g} band, filling it, but only two electrons to the e_g band which is capable of holding four, resulting in a half-filled band.

These difficulties were tackled during the middle years of the 20th century and gave rise to a number of modifications to simple band theory. The mathematics of these modifications are rather complex, and just one example – the effects of electron repulsion – will be described.

13.3.2 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 13.18a). 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 13.18b). The difference between these two energy terms, U, is given by:

$$U = I - A$$

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 sub-bands*, are separated by an energy gap U, called the *Mott-Hubbard splitting* (Figure 13.18c). The lower sub-band, 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 between atoms decreases. This results in increased interaction between the atoms, leading to broader bands.

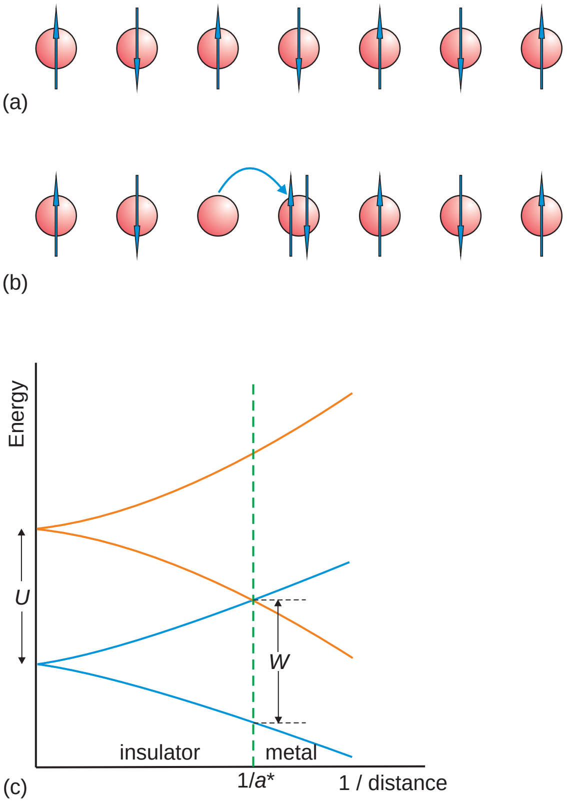


Figure 13.18 Weakly interacting atoms: (a) ground state with one localised electron per atom; (b) after electron transfer; (c) resulting Hubbard sub-bands plotted against the inverse distance between the atoms, $1/a$. A metal–insulator transition occurs at $1/a^*$.

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 the energy to transfer an electron from one atom to its neighbour becomes zero (Figure 13.18c). 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 sub-bands 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 sub-bands. The lower, ground-state band corresponding to ($3d^n$) cations, although only partly filled, is an insulating band, because it requires 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 8 electrons in the sub-band, has a low degree of d-d interaction and the sub-band 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+} (Figure 2.1). 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 sub-bands has occurred and the oxides are metallic.

13.3.3 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, since raising the temperature above 67 °C restores the rutile structure and metallic properties are regained.

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. A typical example of such a pressure-induced transformation is shown by the perovskite-related Mott insulator phase Ca_2RuO_4, which adopts the K_2NiF_4 ($\approx La_2CuO_4$, Figure 13.34) structure. In this phase, Ru^{4+} ($4d^4$) ions occupy oxygen octahedra, but the width, W, 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 W as the d orbitals are forced into greater overlap. In this respect it has been suggested that even solid hydrogen could become metallic at sufficiently high pressures, leading to speculation that metallic hydrogen might be present on the planet Jupiter.

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. The system $W_xV_{1-x}O_2$ has been extensively studied because of this. At room temperature, thin films of pure VO_2 are fairly transparent. Above 67 °C thin films show characteristic metallic reflectivity. This change has a value for the fabrication of 'smart' windows which can reflect sunlight when the day is hot yet allow it through when the day is cool. Unfortunately the transition temperature is too high for ordinary use. Doping, especially with tungsten, causes the transition temperature to fall. 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 $W^{6+}_{x}V^{4+}_{1-3x}V^{3+}_{2x}O_2$. 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.% of 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.

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 12.7.2). 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+} ($La^{3+}_{1-x}A^{2+}_{x}Mn^{3+}_{1-x}Mn^{4+}_{x}O_3$) while one A^+ gives rise to two Mn^{4+} ($La^{3+}_{1-x}A^{+}_{x}Mn^{3+}_{1-2x}Mn^{4+}_{2x}O_3$). Electron transfer can now take place between Mn^{3+} and Mn^{4+} via double exchange (Section 12.5.4) but only if the spins on the two Mn ions are parallel – that is, provided that the metallic state is ferromagnetic.

13.3.4 Transparent conducting oxides

A semiconductor can only absorb light if the band gap is lower in energy than the energy of the incident photons (Section 13.2.2). 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. Whilst there is still uncertainty about the nature of the defects that form on incorporation of SnO_2 into In_2O_3, the following broad-brush picture describes the state of affairs that is believed to occur. The Sn^{4+} ions mainly occupy In^{3+} sites forming substitutional defects. Each Sn^{4+} ion that replaces an In^{3+} must be accompanied by an additional oxygen to maintain charge neutrality, which is accommodated as interstitial O^{2-} ions. (There is some evidence to suppose that the tin and oxygen interstitial defects may aggregate into a defect complex rather than remain isolated, but this does not change the overall argument.) The number of oxygen interstitials formed is equal to half the number of Sn^{4+} ions introduced. However, these interstitial oxygen defects are in equilibrium with the ambient (surrounding) oxygen gas during thin film preparation. The reaction is:

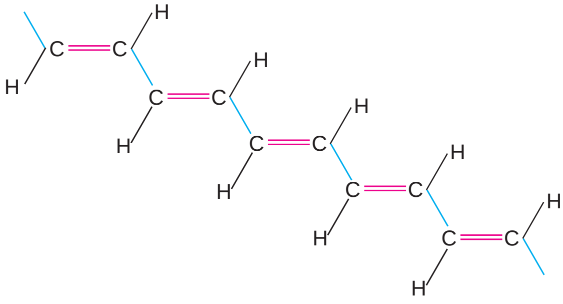
$$O_i^{2\prime} \leftrightarrow {}^1/_2O_2 + 2e'$$

This means that if the films are prepared at higher ambient oxygen pressures, interstitial oxygen defects are preferred. If the films are prepared at lower pressures the oxygen interstitials combine to form oxygen molecules that join the gas phase and leave behind two electrons. These electrons enter the conduction band to induce *n*-type conductivity. This accounts for the fact that the most 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. As the dopant concentration rises, the number of electrons in the oxide increases. At a dopant concentration of about $2 \times 10^{19}\,cm^{-3}$ the material is so heavily doped that the behaviour is that of a degenerate semiconductor (Section 13.1.1) 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.

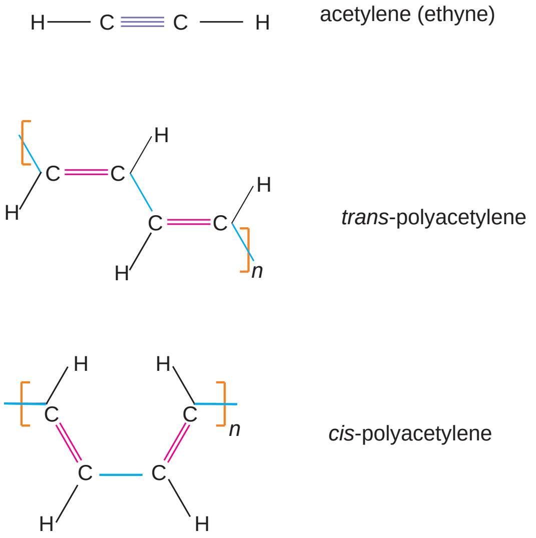
A number of other *n*-type transparent oxide conductors have been found, including tin oxide doped with F and zinc oxide doped with Al_2O_3. Unfortunately, a matching transparent *p*-type oxide conductor has not yet been found, although delafossite structure oxides $CuM^{3+}O_2$, including $CuGaO_2$, $CuInO_2$, $CuScO_2$, 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.

13.4 Conducting polymers

Ordinary polymers are good insulators, and they are widely used in this capacity as insulating coverings 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 (Scheme 13.1). The framework of these molecules is composed of sp^2-hybrid σ bonds, at 120° to each other (Section 2.2.1.3). 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.



Scheme 13.1 Part of a conjugated hydrocarbon molecule (schematic), in which carbon atoms are linked alternately by single and double bonds: C, carbon; H, hydrogen.



Scheme 13.2 The structures of polyacetylenes, schematic: (a) acetylene (ethyne); (b) repeating unit of the *trans*- polymer; (c) repeating unit of the *cis*- polymer.

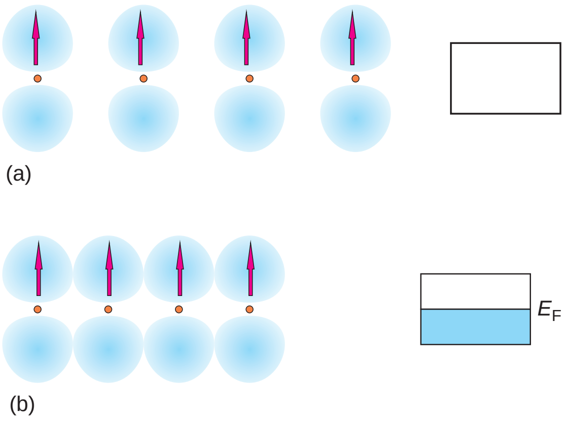


Figure 13.19 Conjugated polymers: (a) a chain of isolated half-occupied p orbitals would lead to an insulator; (b) a chain of overlapping half-occupied p orbitals would lead to a half-filled conduction band and metallic conductivity.

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 (Scheme 13.2). 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 $10^{-3}\,\mathrm{S\,m^{-1}}$), and the *cis* form with a conductivity similar to that of water, (approximately $0.1\,\mathrm{S\,cm^{-1}}$).

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 (Figure 13.19a). 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 p-band. Each atom would contribute just one electron, which would produce a half-full band, and the material should show metallic conductivity (Figure 13.19b).

The real situation corresponds to neither of these alternatives. A linear chain of equispaced atoms in a metal is found to be energetically unstable. Instead, the spacing between the atoms will adjust itself 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 13.20). 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 ($= 2.88 \times 10^{-19}$ J) in the *trans*- form and about 2 eV ($= 3.20 \times 10^{-19}$ J) in the *cis*- form of the polymer, accounting for the observed electronic properties.

Polyacetylene is transformed into a metallic conductor by doping. This involves oxidation or

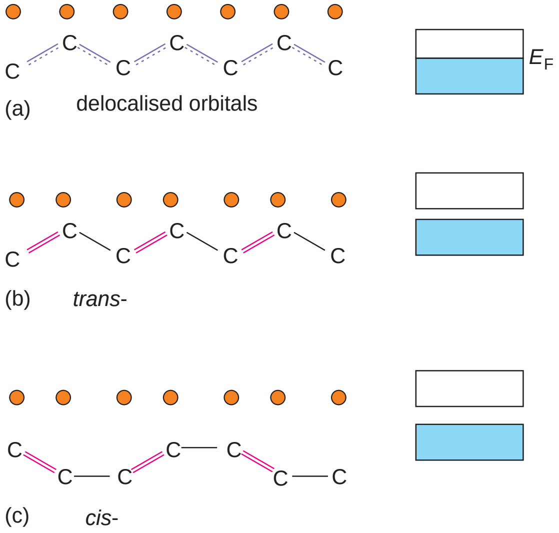


Figure 13.20 Distortion of a polymer chain: (a) delocalised orbitals along a polymer chain lead to equally spaced atoms and a half-filled energy band; (b) Peierls distortion in *trans*-polyacetylene; (c) Peierls distortion in *cis*-polyacetylene; both leading to alternating short and long bonds and a band structure similar to that of an intrinsic semiconductor.

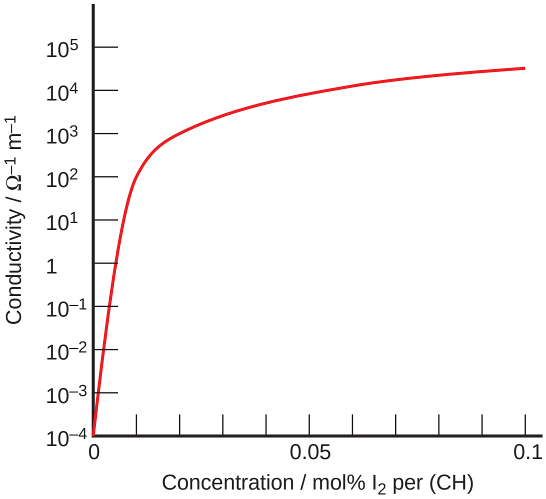


Figure 13.21 Variation of the conductivity of polyacetylene with iodine dopant concentration. The conductivity changes from a value typical of an insulator to that associated with a metal.

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:

$$[\mathrm{CH}]_n + \frac{3x}{2}\mathrm{I}_2 \rightarrow \left[(\mathrm{CH})_n^{x+} x(\mathrm{I}_3^-)\right]$$

The iodine enters the polymer between the molecular chains. Doping with alkali metals, (lithium, sodium, etc), reduces the polymer. 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:

$$[\mathrm{CH}]_n + x\,\mathrm{Li} \rightarrow [(\mathrm{CH})_n^{x-} x\,\mathrm{Li}^+]$$

The conductivities of the doped polymers, of the order of $10^8\,\mathrm{S\,m^{-1}}$, are similar to those of copper or silver and vary with dopant concentration (Figure 13.21). Notice that much greater concentrations of dopant are needed (at.%) than those used in the traditional semiconductors such as silicon (parts per million).

The conductivity and other electrical properties of both the semiconducting and the metallic regions 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 13.22), whereas in an ordinary metal the reverse is true. Moreover, the resistivity seems to show Arrhenius-like temperature-dependence:

$$\sigma = \sigma_0 \exp\left(\frac{-E_\mathrm{a}}{k_\mathrm{B}T}\right)$$

where E_a is the activation energy, k_B is Boltzmann's constant, T the temperature, and σ_0 is a constant (Figure 13.22). This suggests that the dopant does not simply add mobile carriers to a conduction or valence band, as with silicon.

The explanation of this is bound up with the structure of the polymer. Polyacetylene is composed

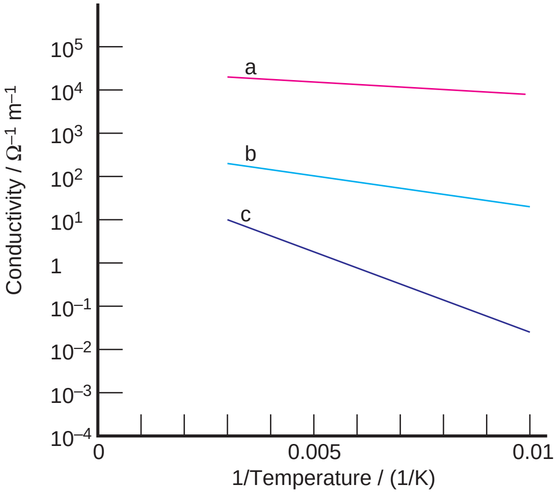


Figure 13.22 Variation of the conductivity of polyacetylene doped with iodine versus reciprocal temperature. Dopant concentration is approximately (a) 0.15 mol%, (b) 0.01 mol%, (c) 0.005 mol% of I_2 per CH unit.

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 13.23). 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.

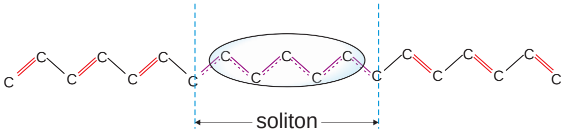


Figure 13.23 Schematic diagram of a soliton in *trans*-polyactylene. The shaded ellipse may represent either a hole (creating a positively charged soliton), a single electron (creating a neutral soliton), or two electrons (creating a negative soliton).

This gives rise to two effects. Firstly, at the soliton, the alternating bond lengths required for the Peierls transition are suppressed. Ultimately, when enough solitons are present, the Peierls transition 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.

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 (Section 7.12) 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, polyactylene is widely used as an electrode material in lightweight batteries.

13.5 Nanostructures and quantum confinement of electrons

Quantum wells, quantum wires and quantum dots (Section 3.2.6) have unique electronic and optical properties because of the way in which electrons are

[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 per hour.

localised, or *confined*. In bulk solids, electrons are located on atom cores in ionic solids, in localised bonds in normal covalent solids, delocalised over molecular orbitals, as in 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.

13.5.1 Quantum wells

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 13.24a). 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 of the composite material (Figure 13.24b). Similarly, the holes in the thin layer of semiconductor are trapped at the 'hill' in the valence band of the composite material. The properties of single quantum wells are enhanced when a number of these

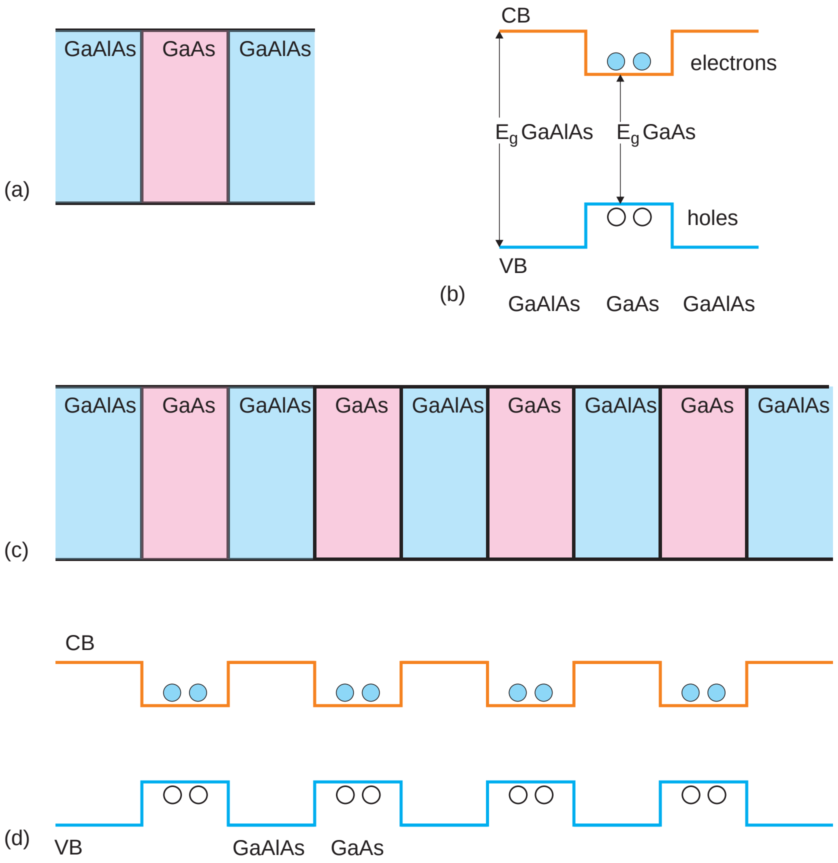


Figure 13.24 Quantum wells: (a) a single quantum well in gallium arsenide, GaAs, formed from a thin layer of gallium aluminium arsenide, GaAlAs; (b) schematic energy band structure of the quantum well; (c) a multiple quantum well superlattice in gallium arsenide; (d) schematic energy band structure of the superlattice.

features are combined to form a *multiple quantum well* or *superlattice* (Figure 13.24c,d).

Quantum confinement becomes important at a dimension Δx, of:

$$\Delta x \approx \sqrt{\left(\frac{h^2}{m_e^* k_B T}\right)}$$

for electrons, and for holes:

$$\Delta x \approx \sqrt{\left(\frac{h^2}{m_h^* k_B T}\right)}$$

where h is Planck's constant, m_e^* is the effective mass of the electron in the semiconductor, m_h^* is the effective mass of the hole in the semiconductor, k_B is Boltzmann's constant and T is 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 (Sections 2.3.2, 2.2.3). 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 will apply. Thus, the energy, E, of a free electron in a rectangular parallelepiped with edges a, b and c is given by:

$$E(n_x, n_y, n_z) = \frac{h^2}{8m_e}\left(\frac{n_x^2}{a^2} + \frac{n_y^2}{b^2} + \frac{n_z^2}{c^2}\right) \qquad (13.5)$$

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 equation (13.5) as:

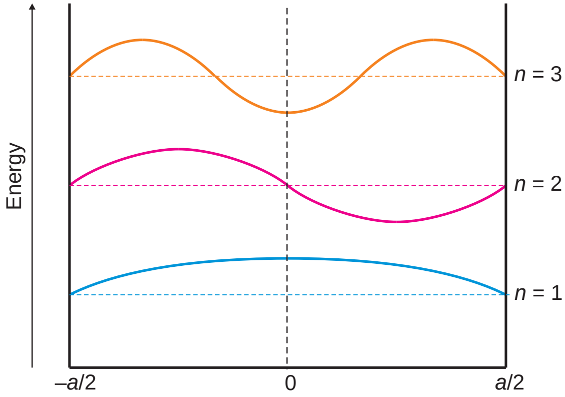


Figure 13.25 The first three energy levels for an electron trapped in a one-dimensional quantum well are the same as an electron trapped on a line.

$$E(n_x, n_y, n_z) = \left(\frac{h^2}{8m_e^*}\right)\left(\frac{n_x^2}{a^2}\right) + \left(\frac{h^2}{8m_e^*}\right)\left(\frac{n_y^2}{b^2} + \frac{n_z^2}{c^2}\right) \qquad (13.6)$$

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 equation (13.6). This introduces a new set of energy levels, associated with electron waves trapped in the well (Figure 13.25). The electron energy level in the lowest, $n = 1$, state is raised by $h^2/8m_e^* a^2$ compared to 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*.

13.5.2 Quantum wires and quantum dots

The above 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 equation (13.5), where a and b are small and c is large. In the case of the quantum dot, equation (13.5) is retained, but the third dimension, c, is also small.

The optical properties of these structures are described in Section 14.10.

13.6 Superconductivity

13.6.1 Superconductors

In 1911, H. Kamerlingh Ohnes found that mercury lost all electrical resistance when cooled to the temperature of liquid helium (4.2 K), and reached the *superconducting* state. Subsequently a large number of materials, including metallic elements, alloys, organic compounds, sulphides, oxides and nitrides have been found to exhibit superconductivity. In this state all electrical resistance is lost and electrical current, once started, will flow forever without diminishing. Superconductivity is a quantum mechanical phenomenon, one of the few that are apparent in the macroscopic world, and its features are puzzling when viewed from the standpoint of classical physics.

The temperature at which materials become superconducting is called the superconducting *transition temperature*, T_c (Figure 13.26). 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 led to an explosion of research that produced large numbers of new oxide superconductors. The current record for T_c, 135 K, is held by $Hg_{0.8}Tl_{0.2}Ba_2Ca_2Cu_3O_{8.33}$ (Figure 13.27).

Recently the variety of materials showing superconductivity has 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

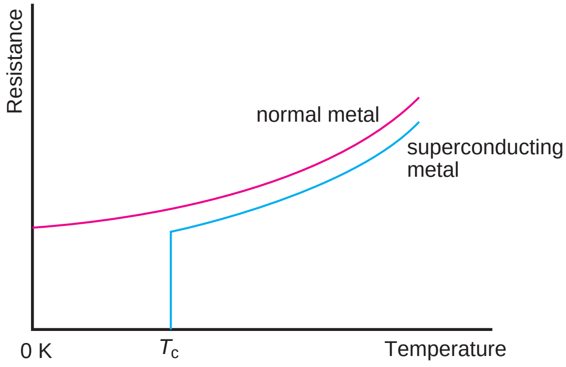


Figure 13.26 The variation of the resistance of a normal metal compared with a superconducting metal, as the temperature approaches 0 K.

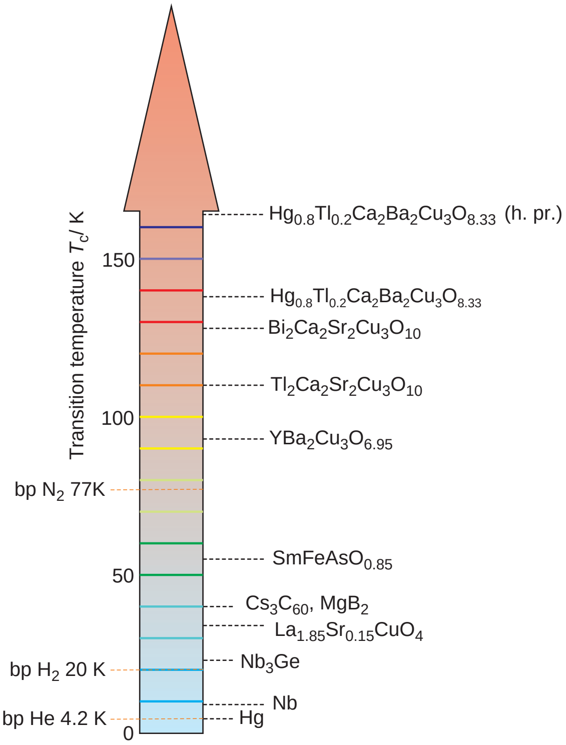


Figure 13.27 Superconducting transition temperatures.

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.

13.6.2 The effect of magnetic fields

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 13.28). 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. 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 13.29a). The relationship between H_c and the critical temperature, T_c, is given by an approximate equation:

$$H_c \approx H_c(0)\left[1 - \left(\frac{T}{T_c}\right)^2\right] \qquad (13.7)$$

where $H_c(0)$ is the critical magnetic field[2] at 0 K, and T is the absolute temperature.

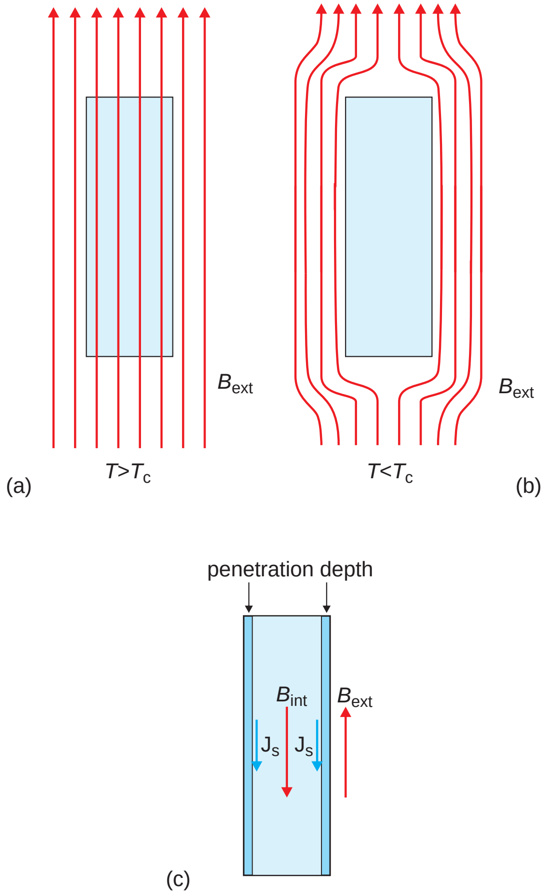


Figure 13.28 The Meissner effect: (a) a normal metal or a superconducting metal above the transition temperature, T_c, allows the penetration of magnetic flux B into the bulk; (b) a superconducting metal below the transition temperature expels the magnetic flux; (c) a surface current, J_s, is induced in the superconductor that creates an internal magnetic field that cancels the external flux.

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 starts to penetrate the material to produce threads that are no longer superconducting (Figure 13.30). 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

[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}.

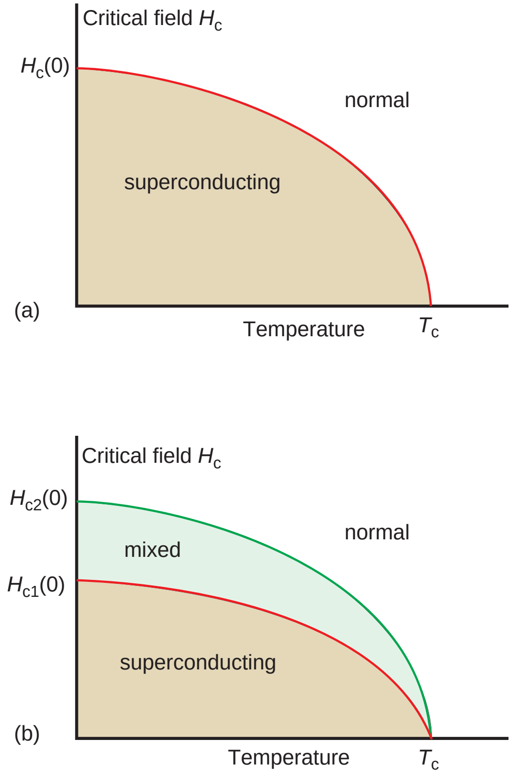


Figure 13.29 The variation of the superconducting properties with the external magnetic field: (a) a Type I superconductor ($H_c(0)$ is the critical field at 0 K); (b) a Type II superconductor ($H_{c2}(0)$ and $H_{c1}(0)$ are the upper and lower critical fields at 0 K).

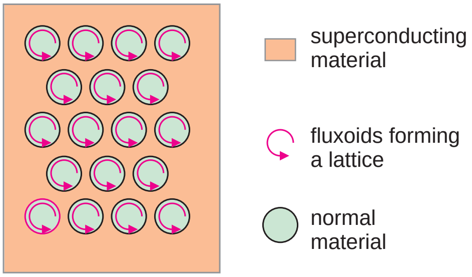


Figure 13.30 A Type II superconductor in the mixed state. The solid contains small threads of normal material, fluxoids, which penetrate the superconducting bulk and repel each other by virtue of the surface currents, to form a fluxon lattice.

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 13.29b). The temperature dependence given in equation (13.7) also holds for type II superconductors if H_c is replaced by $H_{c2}(0)$.

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_o, 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.

13.6.3 The effect of current

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 13.31). As either temperature or magnetic field increase, 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.

13.6.4 The nature 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

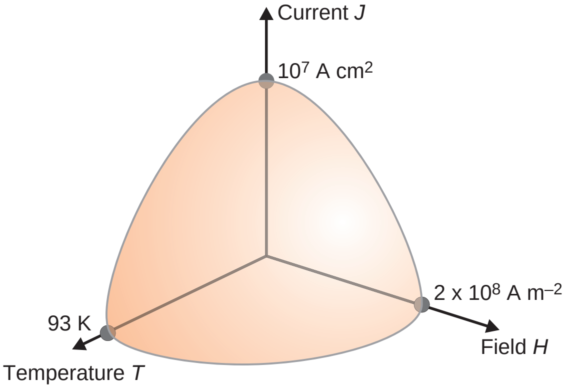


Figure 13.31 The phase space delimiting superconductivity is constrained by the magnetic field, the temperature and the current in the solid. The values of the critical field, temperature and current are for the high-temperature superconductor $YBa_2Cu_3O_{7-x}$.

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 *Bardeen-Cooper-Schrieffer* (BCS) theory. In this, 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 (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 13.32). This leads to 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. 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.

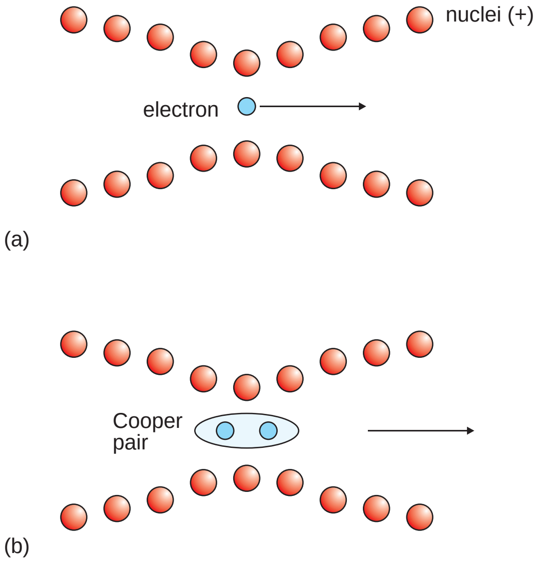


Figure 13.32 Cooper pairs: (a) an electron passing through a solid attracts the positively charged atomic nuclei slightly, creating a slightly distorted region of enhanced positive charge; (b) at low temperatures, another electron can be attracted into this region to form a Cooper pair, which behaves as a single particle.

The formal description relates the distortions to phonons. When an electron in a Cooper pair passes through the crystal, the atom cores are attracted 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 to single electrons. For example, they share the same wave function, and are able to pass through the crystal unimpeded.

In the newly discovered high-temperature superconductors (described below) it has been shown 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 theory has yet been suggested.

13.6.5 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 a 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 13.33a), the maximum current, i_c, that can flow across the device is a 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:

$$i_{\text{cmax}} = 2I_{\text{J}}\cos\left(\frac{\pi\,\Phi}{\Phi_0}\right)$$

where I_{J} is a constant depending upon the junction geometry, Φ is the enclosed magnetic flux and Φ_0 is the flux quantum.

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, $i_{\text{c}(n+{}^1\!/_2)}$. Because of this, the voltage recorded, V_n, $V_{n+{}^1\!/_2}$, or an intermediate value, depends upon the magnetic flux penetrating the loop (Figure 13.33b), and as the magnetic flux changes, the voltage varies in a sinusoidal fashion (Figure 13.33c). A 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, a 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.

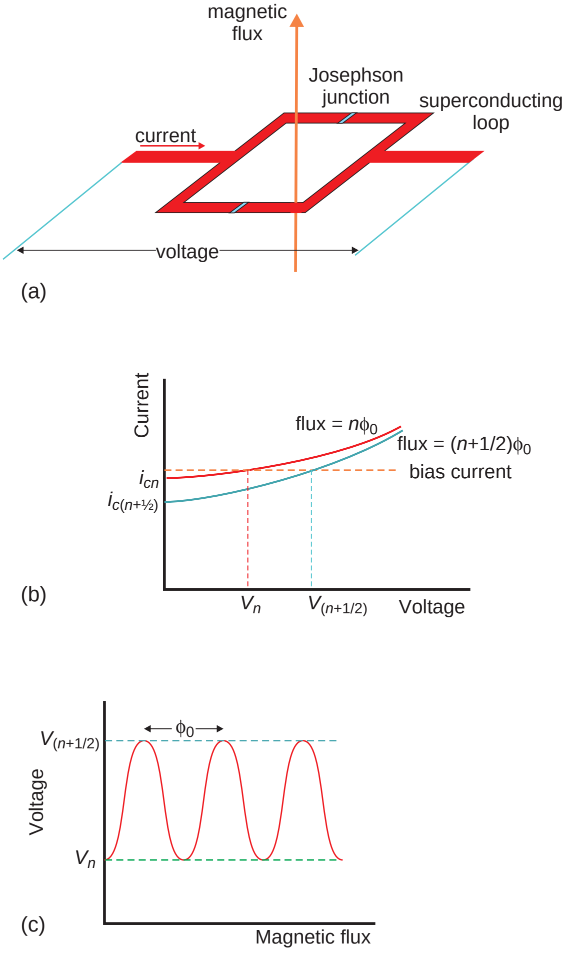


Figure 13.33 A dc SQUID: (a) the circuit consists of a loop of superconducting material containing two Josephson junctions, one in each arm; (b) with a bias current above the critical current, i_c, the voltage depends upon the number of flux quanta that penetrate the loop; (c) in a varying magnetic field, the voltage cycles sinusoidally, with a period equal to the flux quantum.

13.6.6 Cuprate high-temperature superconductors

High-temperature cuprate superconductors are copper oxides that maintain the superconducting state

Table 13.2 Some high-temperature cuprate superconducting oxides

Compound*	T_c/K
$La_{1.84}Sr_{0.16}CuO_4$	38
$Nd_{1.85}Ce_{0.15}CuO_4$	20
$YBa_2Cu_3O_{6.95}$	93
Bi_2O_2/Tl_2O_2layers	
$Bi_2Sr_2CuO_6$	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	
$TlBa_2CuO_5$	
$TlBa_2CaCu_2O_7$	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

*The formulae are representative and do not always show the exact oxygen stoichiometry for the optimum T_c values given.

to temperatures above that of liquid nitrogen (Table 13.2). Crystallographically the phases are all related to the perovskite structure type, ABO_3, where A is a large cation and B is Cu. The structures of the superconductors are built up of slices of perovskite type 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 non-stoichiometric solids, which all exhibit considerable degrees of oxygen composition variation.

13.6.6.1 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 be thought of as sheets of perovskite-type one CuO_6 octahedron in thickness, interleaved by slabs of the halite type, or as CuO_2 and LaO layers stacked in the sequence ... CuO_2, LaO, LaO, CuO_2 ... (Figures 13.34 and 13.36a). 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.

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 most studied phase is $La_{2-x}Sr_xCu^{2+}O_4$. 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 Sr^{2+} concentration passes approximately 0.07, it becomes a *hole superconductor*. The Cu^{3+} population rises, peaking when x is approximately 0.16, to give a maximum superconducting transition temperature of 38 K at a composition of approximately $La_{1.84}Sr_{0.16}CuO_4$.

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

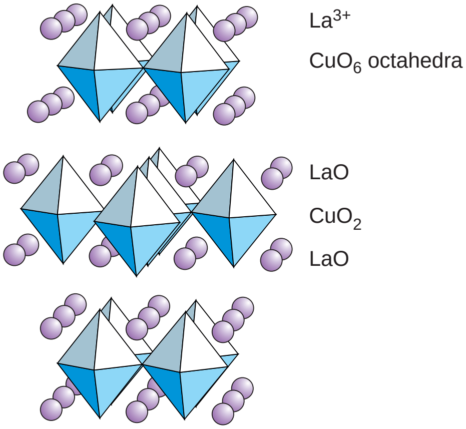


Figure 13.34 The crystal structure of La_2CuO_4 represented as a stacking of distorted corner-linked CuO_6 octahedra and layers of La^{3+} ions; the **c**-axis is vertical.

$La_{2-x}Sr_xCu^{2+}O_{4-x/2}$. 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 (Figure 13.35). Superconductivity occurs within the phase region known as the *superconducting dome*. 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.

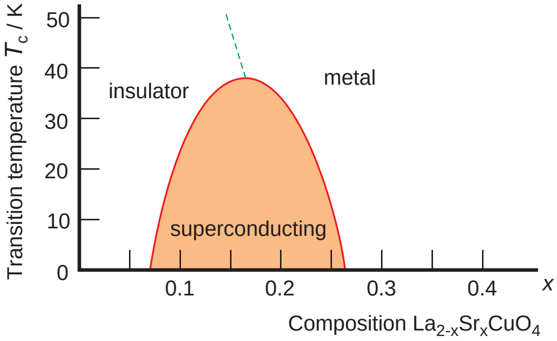


Figure 13.35 The phase region supporting superconductivity, the superconducting dome, for $La_{2-x}Sr_xCuO_4$. Outside this region the electronic properties are complex and approximated here as insulator and metal.

13.6.6.2 Neodymium cuprate, Nd_2CuO_4

The structure of Nd_2CuO_4 is very similar to La_2CuO_4, the principal difference lying in the disposition of the oxygen atoms, as the cations in the two structures are in almost identical positions (Figure 13.36b). 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 type and the structure may be described as an intergrowth of perovskite and fluorite structures. The Cu atoms lie at the centres of square coordination 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 $Nd_{1-x}Ce_xCu^{2+}_{1-x}Cu^+_xO_4$. The Cu^+ ion is equivalent to a Cu^{2+} ion together with a trapped

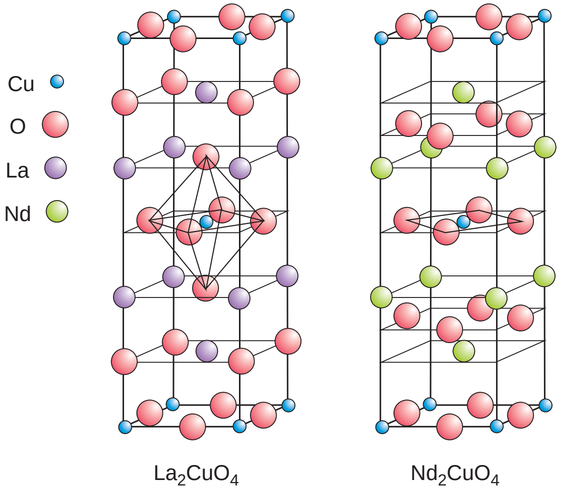


Figure 13.36 Comparison of the structures of La_2CuO_4 and Nd_2CuO_4.

electron, and the compounds are called *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 $Nd_{1.85}Ce_{0.15}CuO_4$.

13.6.6.3 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 unit cells stacked one on top of the other (Figure 13.37a). 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 coordination, 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.

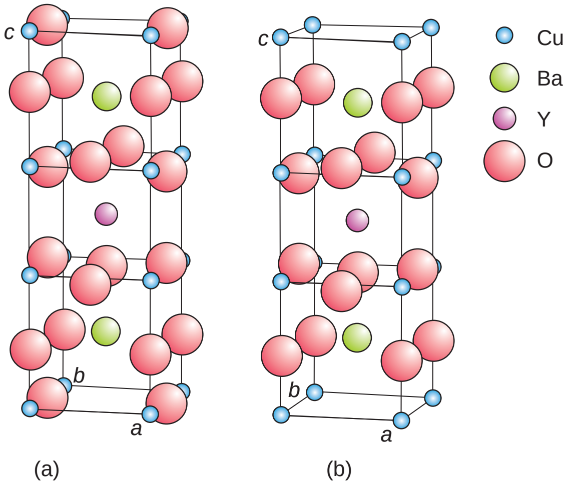


Figure 13.37 The crystal structures of (a) $YBa_2Cu_3O_7$; (b) $YBa_2Cu_3O_6$.

$YBa_2Cu_3O_7$ is a 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 13.38). 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 13.37b).

The oxygen atoms are not lost at random during reduction, but come solely from the CuO_4 square

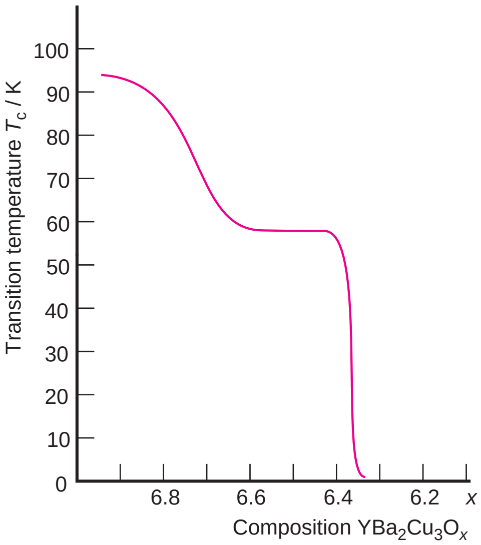


Figure 13.38 The variation of the superconducting transition temperature, T_c, with oxygen content, x, for $YBa_2Cu_3O_x$.

planar units and lie along the **b**-axis. This has the effect of converting the copper coordination 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.

13.6.6.4 Perovskite-related structures and series

The highest transition temperatures found in the cuprate superconductors are shown by 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 13.3).

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 13.39). 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.

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 an alkaline earth, typically Sr or Ba, M is a metal, typically Bi, Pb, Tl or 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 13.40).

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_2-CuO_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 13.41). 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 13.41a). 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 13.41b). To make the relationship clearer, the octahedra are completed in faint outline in Figure 13.41c. In the phases $Ba_2Ca_2Sr_2Cu_3O_{10}$ and $Tl_2Ca_2Sr_2Cu_3O_{10}$ the three CuO_6 octahedral

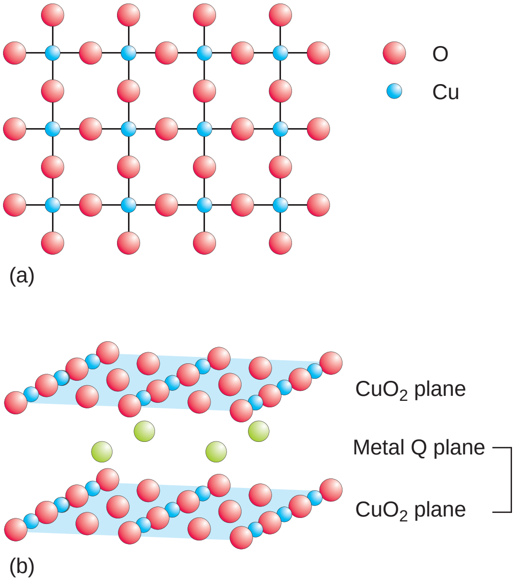


Figure 13.39 Superconducting sheets found in cuprate superconductors: (a) a CuO_2 layer; (b) a CuO_2 (Q–CuO_2) layer.

Table 13.3 Homologous series of cuprate superconductors

Charge reservoir*	Charge reservoir formula	Superconducting slab formula	Idealised series formula*	Examples**
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	$CuO_2\ (YCuO_2)_{n-1}$	$Ba_2Y_{n-1}Cu_{n+1}O_{2n+\delta}$	$CuBa_2YCu_2O_{6+\delta} = YBa_2Cu_3O_{6+\delta}$ ($n = 2$)
		$CuO_2\ (CaCuO_2)_{n-1}$	$CuBa_2Ca_{n-1}Cu_nO_{2n+2}$	$CuBa_2Ca_2Cu_3O_{8+\delta}$ ($n = 3$)
				$CuBa_2Ca_3Cu_4O_{10+\delta}$ ($n = 4$)
	BaO-Hg-BaO	$CuO_2\ (CaCuO_2)_{n-1}$	$HgBa_2Ca_{n-1}Cu_nO_{2n+2}$	$HgBa_2CuO_{4+\delta}$ ($n = 1$)
				$HgBa_2CaCu_2O_{6+\delta}$ ($n = 2$)
				$HgBa_2Ca_2Cu_3O_{8+\delta}$ ($n = 3$)
AO-MO-AO	BaO-TlO-BaO	$CuO_2\ (CaCuO_2)_{n-1}$	$TlBa_2Ca_{n-1}Cu_nO_{2n+3}$	$TlBa_2CuO_5$ ($n = 1$)
				$TlBa_2CaCu_2O_{7+\delta}$ ($n = 2$)
				$TlBa_2Ca_2Cu_3O_{9+\delta}$ ($n = 3$)
	BaO-BiO-BaO	$CuO_2\ (CaCuO_2)_{n-1}$	$BiBa_2Ca_{n-1}Cu_nO_{2n+3}$	$BiBa_2CuO_{5+\delta}$ ($n = 1$)
				$BiBa_2CaCu_2O_{7+\delta}$ ($n = 2$)
				$BiBa_2Ca_2Cu_3O_{9+\delta}$ ($n = 3$)
	SrO-GaO-SrO	$CuO_2\ (Y,CaCuO_2)_{n-1}$	$GaSr_2(Y,Ca)_{n-1}Cu_nO_{2n+3}$	$GaSr_2(Y,Ca)Cu_2O_{7+\delta}$ ($n = 2$)
				$GaSr_2(Y,Ca)_2Cu_3O_{9+\delta}$ ($n = 3$)
AO-MO-MO-AO	SrO-BiO-BiO-SrO	$CuO_2\ (CaCuO_2)_{n-1}$	$Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4}$	$Bi_2Sr_2CuO_{6+\delta}$ ($n = 1$)
				$Bi_2Sr_2CaCu_2O_{8+\delta}$ ($n = 2$)
				$Bi_2Sr_2Ca_2Cu_3O_{10+\delta}$ ($n = 3$)
	BaO-TlO-TlO-BaO	$CuO_2\ (CaCuO_2)_{n-1}$	$Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+4}$	$Tl_2Ba_2CuO_{6+\delta}$ ($n = 1$)
				$Tl_2Ba_2CaCu_2O_{8+\delta}$ ($n = 2$)
				$Tl_2Ba_2Ca_2Cu_3O_{10+\delta}$ ($n = 3$)
AO-MO-M′-MO-AO	SrO-PbO-Cu-PbO-SrO	$CuO_2\ (Y,CaCuO_2)_{n-1}$	$Pb_2Sr_2(Y,Ca)_{n-1}Cu_{n+1}O_8$	$Pb_2Sr_2Y_{0.5}Ca_{0.5}Cu_3O_{8+\delta}$ ($n = 2$)

*The charge reservoir layer is non-stoichiometric. This is not indicated in the idealised formulae given.

**The composition is variable, due to the non-stoichiometric nature of the charge reservoir slabs; δ indicates this and may be either positive or negative.

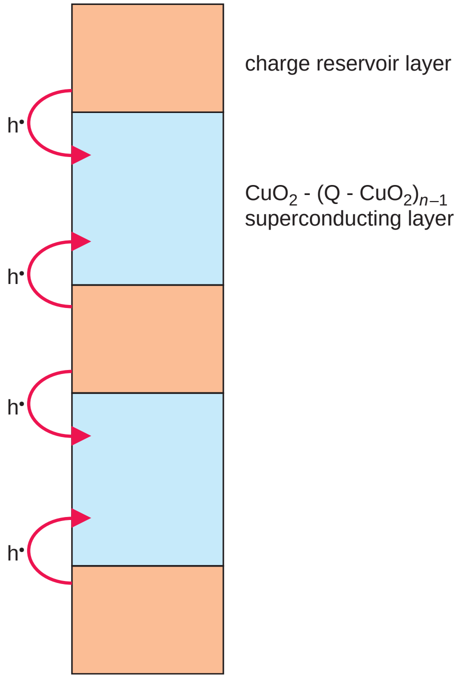


Figure 13.40 Doping into the charge reservoir layers of a cuprate superconductor results in transfer of holes into the superconducting layers.

perovskite layers have been replaced by two sheets of square pyramids and the middle layer by a sheet of CuO_4 squares (Figure 13.41d). The octahedra are completed in faint outline in Figure 13.41e.

Homologous series of superconductors are derived from one type of charge reservoir slab interleaved with $n = 1, 2, 3, 4 \ldots$ CuO_2–(Q–CuO_2)$_{n-1}$ slabs (Table 13.3). 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.

13.6.6.5 Hole doping from charge reservoir layers

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 (*A*O *A*O), 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+} aceptor 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

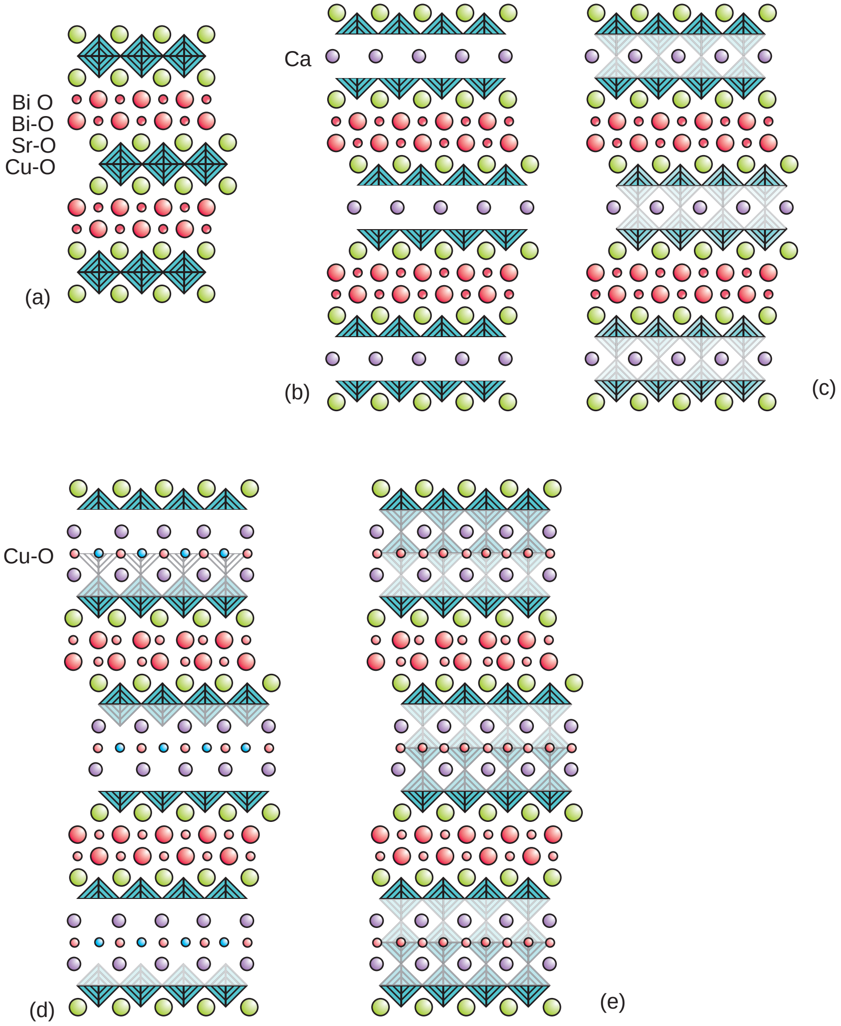


Figure 13.41 Idealised structures of some cuprate superconductors: (a) $Bi_2Sr_2CuO_6$ ($=Tl_2Ba_2CuO_6$); (b) $Bi_2Sr_2CaCu_2O_8$ ($=Tl_2Ba_2CaCu_2O_8$); (c) as (b), but with the nominal CuO_6 octahedra completed in faint outline; (d) $Bi_2Sr_2Ca_2Cu_3O_{10}$ ($= Tl_2Ba_2Ca_2Cu_3O_{10}$); (e) as (d), but with the nominal CuO_6 octahedra completed in faint outline.

oxygen interstitials to the charge reservoir to give a composition $YBa_2Cu_3O_{6+\delta}$:

a. Charge reservoir:
   Formula: ($BaO\ CuO_\delta\ BaO$)
   Charge: $(+5-4-2\delta)=+1-2\delta$

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 $2\delta\,h^\bullet$. 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.

Further reading

Metals:

Cottrell, A. (1988) *Introduction to the Modern Theory of Metals*. Institute of Metals, London.

Cox, P.A. (1987) *The Electronic Structure and Chemistry of Solids*, Oxford University Press, Oxford.

Mott, N.F. (1974) *Metal–Insulator transitions*. Taylor & Francis, London.

Pearson, W.B. (1972) *The Crystal Chemistry and Physics of Metals and Alloys*, especially Chapter 5. Wiley–Interscience.

The band theory definition of a semiconductor is due to A.H. Wilson:

Wilson, A.H. (1931) *Proc. Roy. Soc. Lond.*, **A133**: 458.

Conductivity of (mainly) inorganic solids due to defects is covered in:

Tilley, R.J.D. (2008) *Defects in Solids*, especially Chapters 7 and 8. John Wiley & Sons, Ltd., Hoboken.

Polymers:

Epstein, A.P. (1997) Electrically conducting polymers: science and technology. *Materials Res. Soc. Bulletin*, **22** (June): 6.

Heeger, A.J. (2001) Semiconducting and metallic polymers: the fourth generation of polymeric materials. *Materials Res. Soc. Bulletin*, **22** (November): 900.

Superconductivity:

Waldron, J.R. (1996) *Superconductivity of Metals and Cuprates*. Institute of Physics, Bristol.

The following articles in Scientific American give a good overview of the early years of high-temperature superconductivity:

Hazen, R.M. (1988) Perovskites. *Scientific American*, **258** (June): 52.

Wolsky, A.M., Giese, R.F. and Daniels, E.J. (1989) The new superconductors: prospects for applications. *Scientific American*, **260** (February): 44.

Cava, R.J. (1990) Superconductors beyond 1–2–3. *Scientific American*, **263** (August): 24.

Clarke, J. (1994) SQUIDs. *Scientific American*, **271** (August): 36.

Ouboter, R. de B. (1997) H. K. Ohnes's discovery of superconductivity. *Scientific American*, **276** (March): 84.

These recent reviews give up-to-date information:

Mann, A. (2011) *Nature*, **475**: 280–82.

Various authors (2011) *Science*, **332**: 189–204.

Problems and exercises

Quick quiz

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.

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.

3 In an intrinsic semiconductor the current is carried by:
(a) Electrons.
(b) Holes.
(c) Electrons and holes.

4 Donors make a semiconductor:
(a) p-type.
(b) n-type.
(c) Degenerate.

5 Acceptors make a semiconductor:
(a) p-type.
(b) n-type.
(c) Degenerate.

6 A semimetal has:
(a) A partly filled uppermost band.
(b) Overlapping uppermost bands.
(c) A small band gap between the two uppermost bands.

7 The conductivity of a metal:
(a) Increases as the temperature increases.
(b) Falls as the temperature increases.
(c) Is insensitive to temperature.

8 The conductivity of a semiconductor:
(a) Increases as the temperature increases.
(b) Falls as the temperature increases.
(c) Is insensitive to temperature.

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.

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 Atoms to the right of silicon and germanium in the periodic table act as:
(a) Donors.
(b) Acceptors.
(c) Neither.

12 Atoms to the left of silicon and germanium in the periodic table make the material:
(a) n-type.
(b) p-type.
(c) Neither.

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.

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.

15 Conducting polymers contain:
(a) Metal atoms in the structure.
(b) Conjugated double bonds
(c) Conjugated triple bonds.

16 Doping polyacetylene with sodium makes it:
(a) A metallic conductor.
(b) A p-type semiconductor.
(c) An n-type semiconductor.

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.

18 Electrons in a quantum wire are strongly confined:
(a) In one dimension.
(b) In two dimensions.
(c) In three dimensions.

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.

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.

21 Superconductivity in conventional (low-temperature) superconductors is due to:
(a) Pairs of electrons.
(b) Pairs of holes.
(c) Electron–hole pairs.

22 Ceramic superconductors mostly have structures closely related to:

(a) Spinel.

(b) Perovskite.

(c) Halite.

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.

24 A SQUID measures:

(a) Resistivity.

(b) Superconductivity.

(c) Magnetic fields.

Calculations and questions

13.1 The electrical resistivity of gold at 273 K is $2.05 \times 10^{-8}\,\Omega$ m. Gold adopts the A1 structure with a lattice parameter of 0.4078 nm. The velocity of electrons at the Fermi surface is $1.40 \times 10^{6}\,\mathrm{m\,s^{-1}}$. 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.

13.2 The electrical resistivity of silver at 273 K is $1.47 \times 10^{-8}\,\Omega$ m. Silver adopts the A1 structure with a lattice parameter of 0.4086 nm. The velocity of electrons at the Fermi surface is $1.39 \times 10^{6}\,\mathrm{m\,s^{-1}}$. 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 to the interatomic spacing of silver atoms in the crystal.

13.3 The electrical resistivity of rubidium at 273 K is $11.5 \times 10^{-8}\,\Omega$ m. Rubidium adopts the A2 structure with a lattice parameter of 0.5705 nm. The velocity of electrons at the Fermi surface is $8.1 \times 10^{7}\,\mathrm{m\,s^{-1}}$. 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.

13.4 The electrical resistivity of magnesium at 273 K is $4.05 \times 10^{-8}\,\Omega$ m. Magnesium adopts the A3 structure with lattice parameters of $a = 0.3209$ nm, $c = 0.5211$ nm. The velocity of electrons at the Fermi surface is $1.58 \times 10^{6}\,\mathrm{m\,s^{-1}}$. 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 to the interatomic spacing of rubidium atoms in the crystal.

13.5 The electrical resistivity of liquid mercury as a function of temperature is given in the table. The velocity of electrons at the Fermi surface is $1.52 \times 10^{6}\,\mathrm{m\,s^{-1}}$. Determine how the mean free path varies with temperature. The density of liquid mercury is 13 456 $\mathrm{kg\,m^{-3}}$. Assume that this does not vary with temperature and that each mercury atom contributes two mobile electrons to the liquid.

$\rho/(10^{-8}\,\Omega\,\mathrm{m})$	94.1	103.5	128.0	214.0	630.0
T/°C	0	100	300	700	1200

13.6 The resistivity of cadmium metal crystals at room temperature is $7.79 \times 10^{-8}\,\Omega$ m parallel to the **c**-axis and $6.54 \times 10^{-8}\,\Omega$ m parallel to the **a**-axis. Cadmium adopts the A3 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.

13.7 The resistivity of zinc metal crystals at room temperature is $6.05 \times 10^{-8}\,\Omega$ m parallel to the **c**-axis and $5.83 \times 10^{-8}\,\Omega$ m parallel to the **a**-axis. Zinc adopts the A3 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.

13.8 The resistivity of a sample of brass containing 70 wt.% Cu and 30 wt.% Zn is $6.3 \times 10^{-8}\,\Omega$ m at 0 °C, compared with that of pure copper, which is $1.54 \times 10^{-8}\,\Omega$ m at the same temperature. Determine the residual resistivity at this temperature.

13.9 The resistivity of a sample of bronze containing 90 wt.% Cu and 10 wt.% Sn is $1.36 \times 10^{-7}\,\Omega$ m at 0 °C, compared with that of pure copper, which is $1.54 \times 10^{-8}\,\Omega$ m at the same temperature. Determine the residual resistivity at this temperature.

13.10 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/ $10^{-8}\,\Omega$ m	6.16	28.1	70	107.3	42.9
Wt.% Ni	100	95	90	77.3	67.1

13.11 The resistivity at 0 °C for a number of aluminium alloys is given in the 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/ $10^{-8}\,\Omega$ m	2.42	3.5	3.95	3.5	3.95
Wt.% Al	100	93	89	87	85

13.12 The resistivity at 0 °C for a number of copper alloys is given in the 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 silver	Constanin
Resistivity/ $10^{-8}\,\Omega$ m	1.54	19.8	41.5	6.3	40	49
Wt.% Cu	100	90	84	70	62	60

13.13 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.

13.14 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_e. E_g is 1.42 eV.

13.15 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.

13.16 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.

13.17 (a) Determine the band gap of silicon from the conductivity data given in the table. (b) 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}

13.18 (a) Determine the band gap of germanium from the conductivity data given in the Table. (b) 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	0.09	0.05	0.1	0.6	1.0

13.19 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.

13.20 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.

13.21 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.

13.22 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.

13.23 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?

13.24 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?

13.25 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?

13.26 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 13.15, determine the conductivity of the sample.

13.27 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.

13.28 (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.

13.29 A single crystal of germanium doped with antimony to make it p-type is used in a Hall experiment. The crystal dimensions are $20 \times 10 \times 1$ 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

13.30 Using Figure 13.22, estimate the activation energy for the conductivity of polyacetylene doped with iodine for the three concentrations shown.

13.31 The oxides TiO_2 and Ti_2O_3 both show composition ranges.

(a) 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?

(b) 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?

13.32 The oxide $LaCoO_3$ is an insulator. (a) 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. (b) What are the charges on the cations present? Is the doped material a p-type or n-type semiconductor?

13.33 The compound Mg_2TiO_4 is an inverse spinel and the compound $MgTi_2O_4$ is a normal spinel (Section 5.3.9). 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 $Mg_{2-x}Ti_{1+x}O_4$. What are the charges on the ions in this phase?

(c) How are the ions distributed in the spinel structure?

(d) Will the material be an electron or hole conductor?

13.34 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?

13.35 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?

13.36 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 $0.1\,m_e$.

13.37 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 $0.067\,m_e$.

13.38 Determine the energy of the $n = 1$ energy level in a fragment of n-type gallium nitride 20 × 393 × 1700 nm. The effective mass of electrons in this material is $0.19\,m_e$.

13.39 The critical field of the superconductor $PbMo_5S_6$ is given as 60 T. What is the critical field in $A\,m^{-1}$?

13.40 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.

13.41 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.

13.42 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?

13.43 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?

14

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. When light falls onto a solid it is absorbed and/or scattered. Scattering generally refers to interactions in which there is little or no energy loss, and scattering in its various forms is most easily discussed by treating light as a wave. Absorption, on the other hand, generally involves considerable energy exchange with the absorbing centres. In these instances, light is best regarded as a stream of particles called photons. The chapter opens with a description of these two aspects of light.

14.1 Light

14.1.1 Light waves

Light is the form of energy detected by the eye, and at ordinary scales can be treated as a *wave*. Light waves are part of the *electromagnetic spectrum*, ranging continuously from very long radio waves, with wavelengths of Gm, to high-energy cosmic rays, with wavelengths of the order of fm (Figure 14.1). For convenience, different frequency ranges are given broad-brush names, for example, radio waves and microwaves, although the boundaries between regions are blurred.

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:

$$\mathscr{E}_{\mathrm{y}} = \mathscr{E}_0 \cos\left\lfloor \left(\frac{2\pi}{\lambda}\right)(x - vt) + \phi \right\rfloor$$

where $\mathscr{E}_{\mathrm{y}}$ 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 $\left\lfloor \left(\frac{2\pi}{\lambda}\right)(x - vt) + \phi \right\rfloor$ 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 14.2). The separation between two adjacent peaks or two adjacent

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

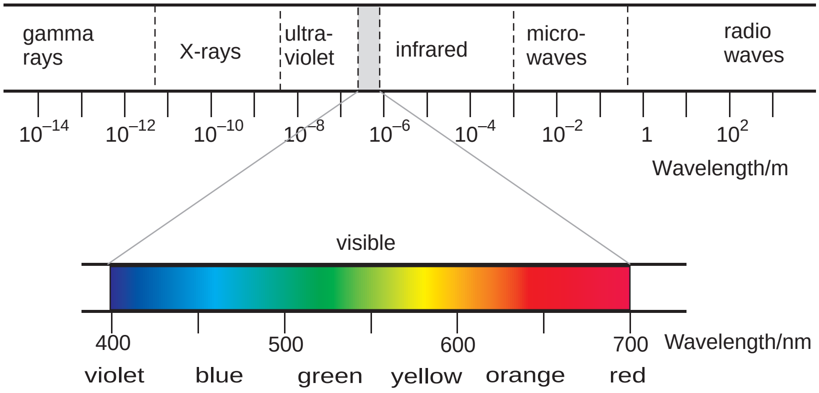


Figure 14.1 The electromagnetic spectrum; the visible region only occupies a small part of the whole, from approximately 400–700 nm.

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 is related to the (*temporal) frequency* ν, of the wave by the equation:

$$\mathrm{v} = \lambda \nu$$

The velocity of light in a vacuum, $2.99792 \times 10^8\ \mathrm{ms}^{-1}$, 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*.

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 14.1).

The visible spectrum is bounded by *infrared* at long wavelengths and *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* is

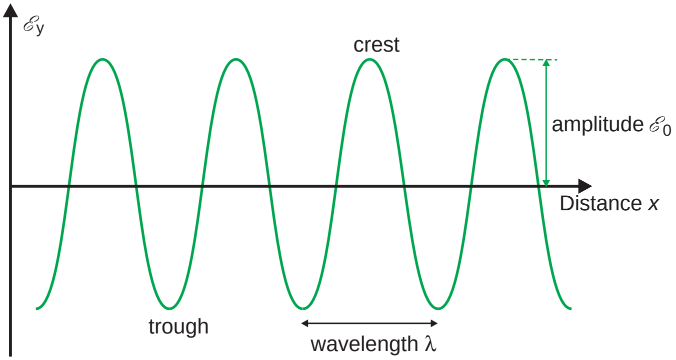


Figure 14.2 The electric field of a light wave can be represented as a sinusoidal wave of amplitude $\mathscr{E}_0$ and wavelength λ, equal to the separation between two crests or troughs.

Table 14.1 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

closest to the violet region and although invisible to humans, can be seen by many animals. *Ultraviolet B* 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} s 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 14.3). 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.

14.1.2 Photons

When light interacts with individual atoms or molecules it is best regarded as a stream of *photons* of energy E given by:

$$E = h\nu = \frac{hc}{\lambda}$$

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 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:

$$\nu\lambda = c$$

Isolated atoms and molecules have sharp energy levels. When a stream of photons encounters an atom or a molecule, the light will be absorbed only if the photon energy precisely matches the energy

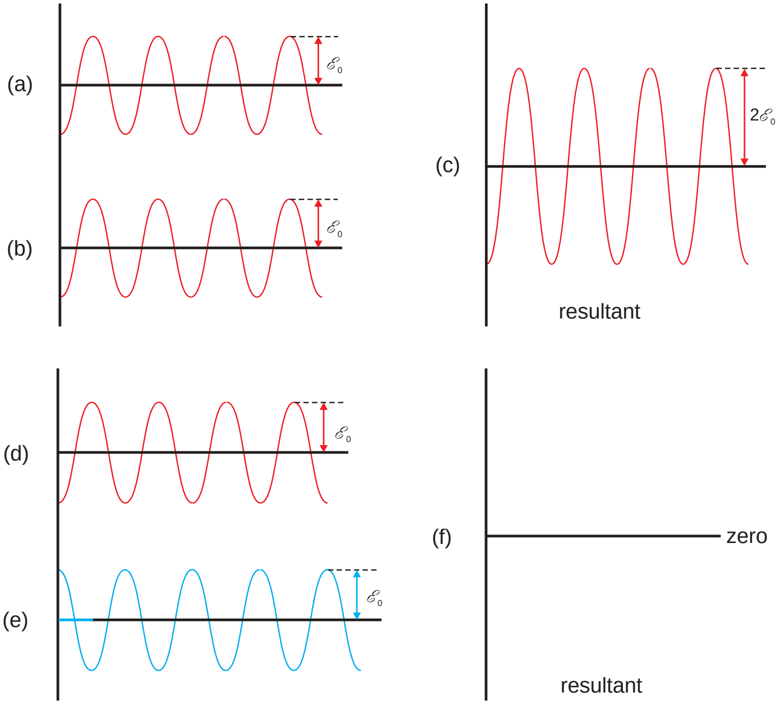


Figure 14.3 The interference of light waves: (a, b) two waves in step add to give a resultant with twice the amplitude of the original waves, (c); (d, e) two waves out of step add to zero, (f).

between the occupied energy level and one of the 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 14.4a). Under normal circumstances atoms or molecules 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.

When an atom or molecule is in a high-energy state, it may lose energy by dropping to a lower-energy state. At the same time a photon is emitted, which carries off the excess energy. The energy of the photon emitted E_e, is given by (Figure 14.4b):

$$E_e = h\nu = E_2 - E_1$$

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 14.1). Molecules also have energy levels due to vibration, giving rise to infrared photons (approximately 0.37 eV; 6×10^{-20} J), and rotation, giving rise to microwave photons (0.0037 eV; 6×10^{-22} J) (Figure 14.4c).

Although a photon may have the correct energy to be absorbed, the *probability* of the transition occurring is governed by quantum mechanical *selection*

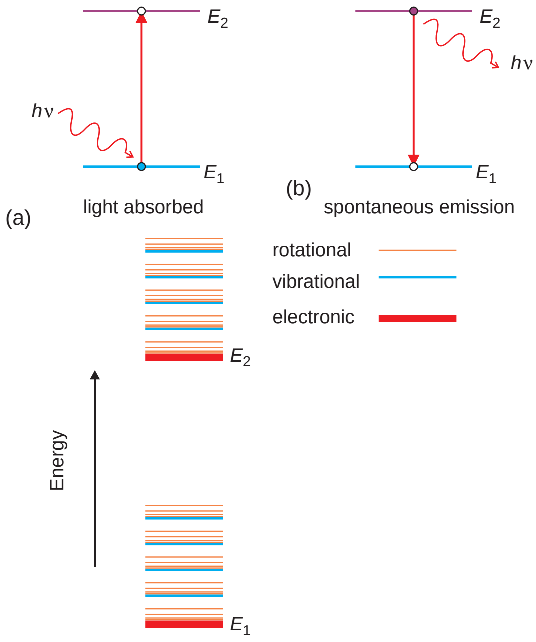


Figure 14.4 Absorption and emission of radiation: (a) light absorption occurs when a photon excites an atom or molecule from a lower electronic energy E_1 to a higher energy E_2; (b) atoms can emit an identical photon via spontaneous emission when dropping from E_2 to E_1; (c) each electronic energy level in a molecule has additional associated energy levels due to molecular vibration and rotation.

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.

14.2 Sources of light

14.2.1 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 14.1, 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 14.5). 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).

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

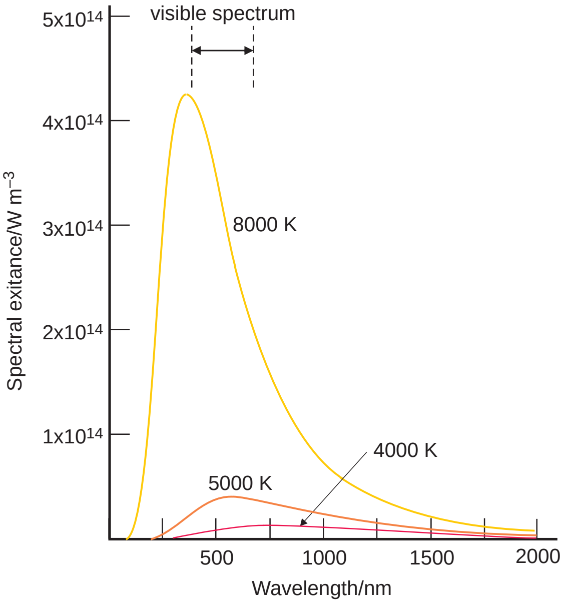


Figure 14.5 Black body emission spectra; the maxima move to shorter wavelengths as the temperature of the emitter increases.

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:

$$\varepsilon = \mathrm{h}\nu$$

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:

$$l_\lambda = \frac{2\pi hc^2}{\lambda^5\left[\exp\left(\dfrac{hc}{\lambda k_B T}\right) - 1\right]}$$

The form of this expression matched the experimental observations perfectly.

Until the end of the 20th 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*). It is clear (Figure 14.5) that the amount of light energy emitted by these sources is only a fraction of the total energy released, and these lamps were both hot (a fire risk) and inefficient. They have now largely been replaced by other forms of artificial lighting.

14.2.2 Luminescence and phosphors

The emission of radiation by solids at relatively low temperatures is called *luminescence*. For luminescence to occur, energy must be supplied to the solid in some form or another, and is divided into a number of subcategories that mirror this dependency (Table 14.2). 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*,

Table 14.2 Types of luminescence

Type	Source of energy
Photoluminescence	Photons, mainly ultraviolet
Triboluminescence	Mechanical bond-breaking, fracture or friction
Chemiluminescence	Chemical reactions
Cathodoluminescence	Electron bombardment
Thermoluminescence	Increase of temperature
Electroluminescence	Applied electric field
Bioluminescence	Life processes
Radioluminescence	Radioactive decay
Sonoluminescence	Ultrasonic waves

often by considerable lengths of time, because the light-emitting transitions are forbidden.

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 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 *sensitizer*, is needed as well. In this case the sensitizer absorbs the exciting photons and passes the energy to the activator.

14.2.2.1 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 has made the efficiency greater, and since the 1980s compact fluorescent light tubes have become commonplace.

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 Hg atoms are excited and emit copious ultraviolet radiation. This consists mainly of the wavelengths 185 nm, 254 nm and 365 nm, as well as some radiation in the visible range. Conversion of the ultraviolet radiation to visible is by way of a phosphor coated onto the inside of the tube (Figure 14.6). The various forms of fluorescent lamps now available differ in the type of phosphor coatings used.

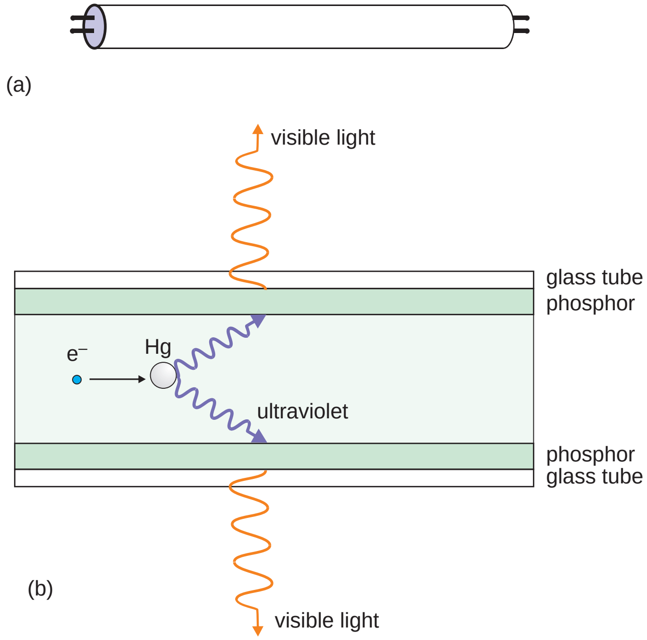


Figure 14.6 Fluorescent lamps: (a) schematic fluorescent tube lamp; (b) electrons (e^-) from the cathode collide with mercury (Hg) atoms in the tube which emit ultraviolet photons that are converted into visible light by a phosphor coating.

Halophosphate lamps use modified calcium fluorophosphate, $Ca_5(PO_4)_3F$, 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 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 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. These lamps are still available, and work continues on improving their performance.

Trichromatic (Colour 80) lamps improve the spectral balance compared with halophosphate lamps by using a phosphor mixture that emits equal amounts of the colours red, blue and green. The favoured red emitter in trichromatic lamps is Eu^{3+} as activator doped into Y_2O_3 matrix, Y_2O_3:Eu, with the Eu^{3+} ions occupying the Y^{3+} sites. The main transition leads to 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 fluorophosphate 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 parts of the spectrum.

The colour spectra of the fluorescent lamps described above, although satisfactory for many purposes, do not give an accurate impression of the colour of an object compared with 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.

14.2.3 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 13.2.6). 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 (Table 14.1, Figure 14.7).

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 13) is not adequate. In effect the bands must be described as a series of undulating surfaces.

There are two possibilities of importance for light absorption and emission. The first is that 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, **k**[1] (Figure 14.8a,b). In this case, 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.

In the second case, 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 14.8c,d). In this case, the photon that is to promote the electron from the top of the valance band to the conduction band must also interact with the lattice to pick up (or lose) sufficient momentum to make the transition possible. To complete this energy adjustment, the photon must interact with phonons (quanta of lattice vibrations). 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 $GaAs_xP_{1-x}$

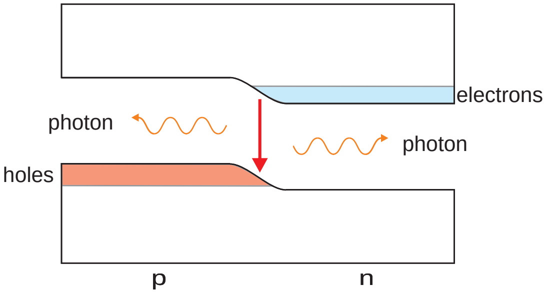


Figure 14.7 The principle of LED operation: under a forward bias, electrons and holes recombine in the junction region between p- and n-type semiconductors and emit radiation.

[1] The momentum of an electron is equal to $kh/2\pi$, and the **k** axis is often referred to as the momentum axis.

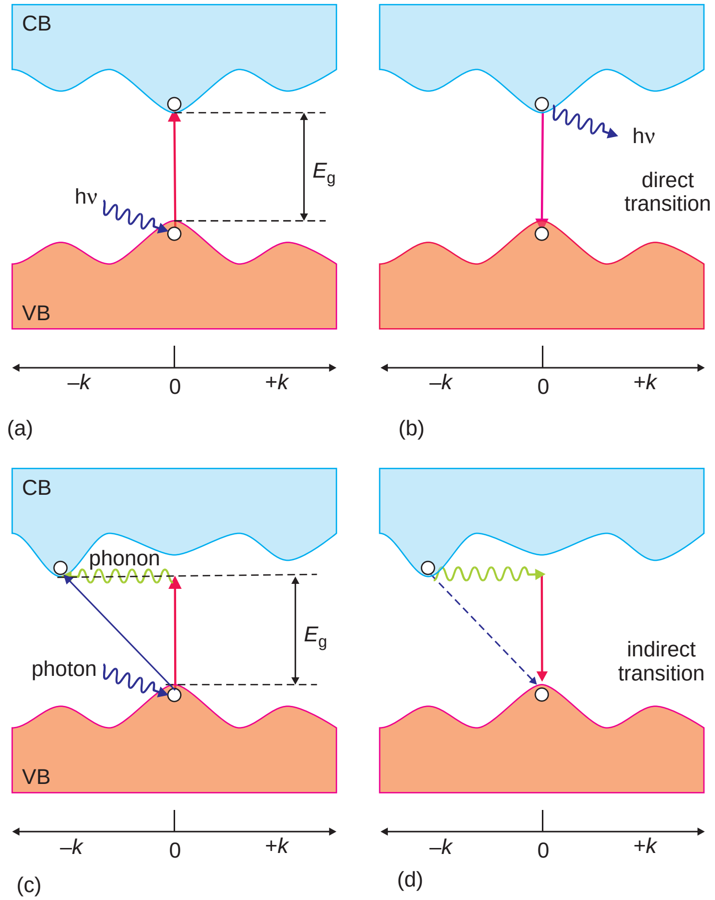


Figure 14.8 Direct and indirect band gap materials: (a, b) a direct band gap material can absorb and emit photons efficiently; (c, d) an indirect band gap material requires the interaction of a photon and a phonon in the process.

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 range. 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 were made with gallium nitride (GaN) and indium nitride (InN), which 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.

14.2.4 Solid-state lasers

When an atom in an excited state makes a transition to the ground state, energy 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 14.1.2). 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 the sources described above, which is produced by spontaneous emission and is 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 to those in the ground state, N_0, the *populations*, will be extremely small for energy levels that 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:

$$\frac{N_1}{N_0} = \exp\left[\frac{-(E_1 - E_0)}{k_B T}\right]$$

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, by spontaneous emission will be given by:

$$-\frac{dN_1}{dt} = A_{10}N_1$$

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 per second due to spontaneous emission 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:

$$-\frac{dN_0}{dt} = B_{01}\rho(\nu_{01})N_0$$

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:

$$-\frac{dN_1}{dt} = B_{10}\rho(\nu_{10})N_1$$

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:

$$\rho(\nu) = \frac{N_1A_{10}}{N_0B_{01} - N_1B_{10}}$$

At equilibrium, the Boltzmann distribution applies, thus:

$$\frac{N_1}{N_0} = \exp\left(\frac{-h\nu}{k_\mathrm{B}T}\right)$$

and by making this substitution:

$$\rho(\nu) = \frac{A_{10}}{\exp\left(\frac{-h\nu}{k_\mathrm{B}T}\right)B_{01} - B_{10}}$$

This expression represents the net radiation emitted or absorbed by the material. This should be equal to the radiation density in a black body, derived by Planck:

$$\rho(\nu) = \frac{8\pi h\nu^3}{c^3\left(\exp\left(\frac{-h\nu}{k_\mathrm{B}T}\right) - 1\right)}$$

leading to the conclusion that $B_{01} = B_{10}$, which will be replaced by the single symbol B, and:

$$\frac{A_{10}}{B} = \frac{8\pi h\nu^3}{c^3}$$

The *ratio* of the rate of spontaneous emission to stimulated emission under conditions of thermal equilibrium is given by:

$$R = \frac{A_{10}}{\rho(\nu)B} = \exp\left(\frac{h\nu}{k_\mathrm{B}T}\right) - 1$$

At 300 K, at visible wavelengths, R is much greater than 1. This shows that for light, stimulated emission will be negligible compared with 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 difficulties 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.

14.2.4.1 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:

$$2S + 1 = 2\left({}^1/_2 + {}^1/_2 + {}^1/_2\right) + 1 = 4$$

and the ground state term symbol is $^4\mathrm{F}$ (Section 1.3). The Cr^{3+} ion in ruby is in an octahedral site and the crystal field (Section 12.2.2) splits this term into three energy levels, which are, in ascending order, labelled $^4\mathrm{A}_{2\mathrm{g}}$, $^4\mathrm{T}_{2\mathrm{g}}$ and $^4\mathrm{T}_{1\mathrm{g}}$. 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

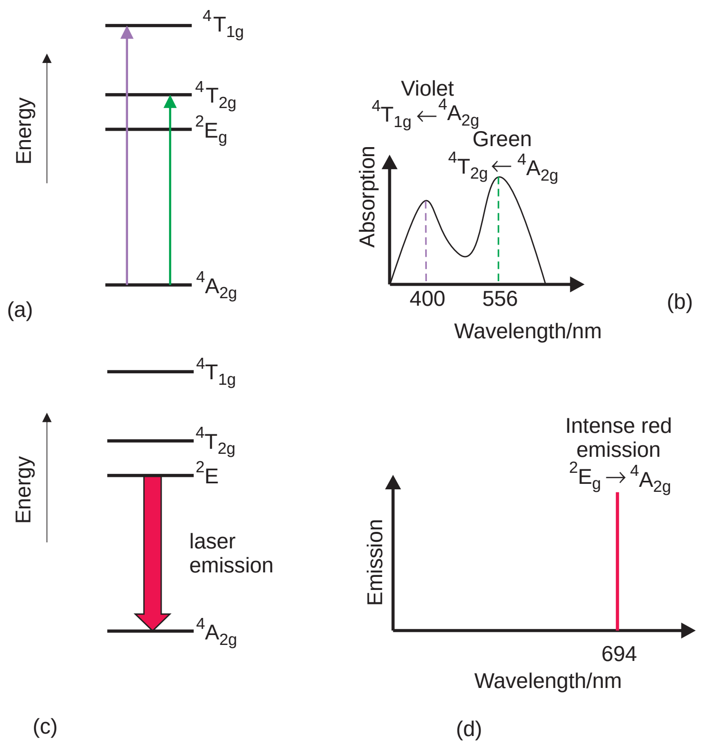


Figure 14.9 The energy levels involved in the ruby laser: (a) transitions from the ground state $^4A_{2g}$ to energy levels $^4T_{2g}$ and $^4T_{1g}$ produce the normal colour of ruby; (b) the absorption spectrum of ruby; (c) the transition from level ^2E_g to the ground state, $^4A_{2g}$, is responsible for the laser emission; (d) emission spectrum of laser.

spin-allowed transitions (Figure 14.9a,b):[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:

$$2S + 1 = 2\left(\tfrac{1}{2} + \tfrac{1}{2} - \tfrac{1}{2}\right) + 1 = 2$$

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 14.9a,b). A transition to or from $^4A_{2g}$, $^4T_{2g}$ and $^4T_{1g}$ to ^2E_g 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.

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

[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.

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 14.9c,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 lifetime (^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.

14.2.4.2 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 14.10). 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

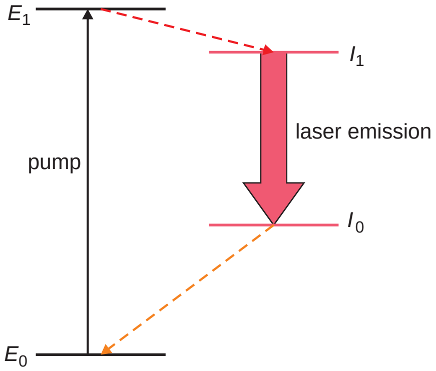


Figure 14.10 The principle transitions in a four-level laser. The laser transition occurs between two intermediate energy levels, I_1 and I_0. The pump transition is between the ground state, E_0 and energy level E_1.

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.

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 14.11). 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 14.8).

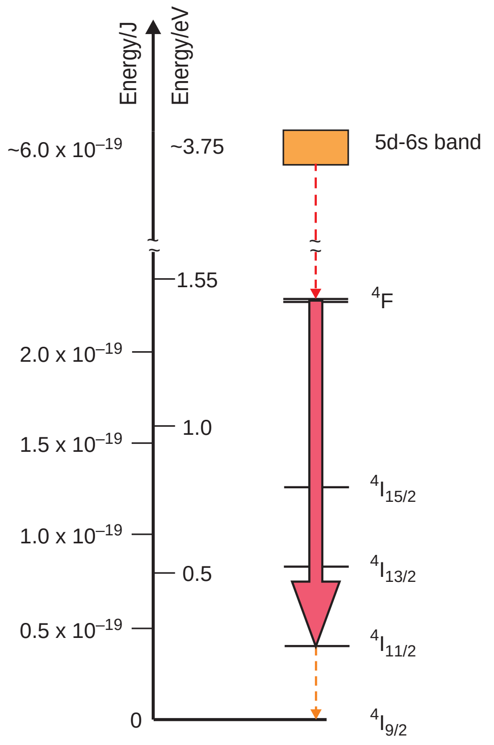


Figure 14.11 The energy levels of most importance in the neodymium laser. The pump transitions are from the ground state to a broad 5d–6s band. The main laser transition occurs between the ^4F and $^4I_{11/2}$ levels.

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.

14.2.4.3 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 occurs and light leaving the active layer is coherent laser light (Figure 14.12). 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 a number of carefully engineered layers with varying electronic characteristics.

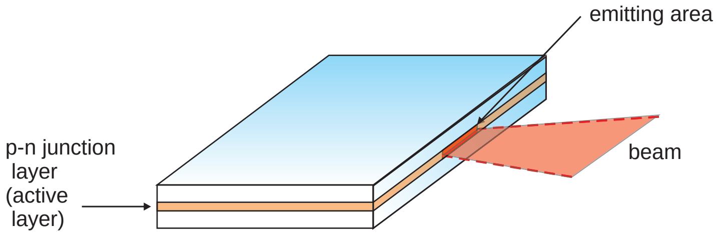


Figure 14.12 Schematic diagram of a semiconductor laser. The beam is emitted from a thin p-n junction active layer.

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.

14.3 Colour and appearance

14.3.1 Luminous solids

The colour of a solid depends upon the light that travels from the solid to the observer's eye. Luminous objects emit radiation directly and the colour of the object will correspond to the wavelength range recorded in the eye. Individual wavelengths are not perceived separately, and a mixture of wavelengths corresponding to, say, red and yellow, is seen as orange. The combination of different wavelengths of light is called *additive coloration*. Mixing just three different wavelengths of light in various proportions can reproduce the perceived colours of all light sources. The three wavelengths are called *additive primary colours* and the process of mixing lights to obtain other colours is called *additive mixing*. There is no fixed set of primary colours, and any three colours loosely designated as red, blue and yellow will suffice for the purpose. Mixing the three additive primary colours in equal proportions will produce *white light*.

14.3.2 Non-luminous solids

Non-luminous solids interact with light passively. When light of a particular wavelength falls onto such a solid it might be absorbed, by *absorption centres* in the material, in which case the energy of the light may end up in the solid structure as heat, and the solid is described as *opaque* to the absorbed wavelength. Other absorption centres may re-emit light as photoluminescence. Alternatively, the light may pass through the solid unhindered and the solid will be *transparent* to that wavelength. The light that leaves the material is the *transmitted* light. Finally, some of the incident light may be *scattered*. The commonest scattering process is *reflection*, which takes place at surfaces. Smooth surfaces reflect light uniformly, known as *specular reflection*. Rough surfaces will reflect incident light in all directions, called *diffuse* reflection. The appearance of the solid will depend upon which of these processes dominate (Figure 14.13).

The appearance of even a simple transparent solid will vary greatly due to the effects of scattering alone. For example, a transparent crystal will appear white when powdered because no light is absorbed but the numerous crystallite surfaces scatter light of all wavelengths equally. Transparent crystallites or voids within glass produce the same effect. If there are sufficient numbers, the glass will appear milky white. Opal glass is deliberately made to produce 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.

Transparent objects become coloured by the selective absorption of radiation. If white light passes through a material that absorbs blue and

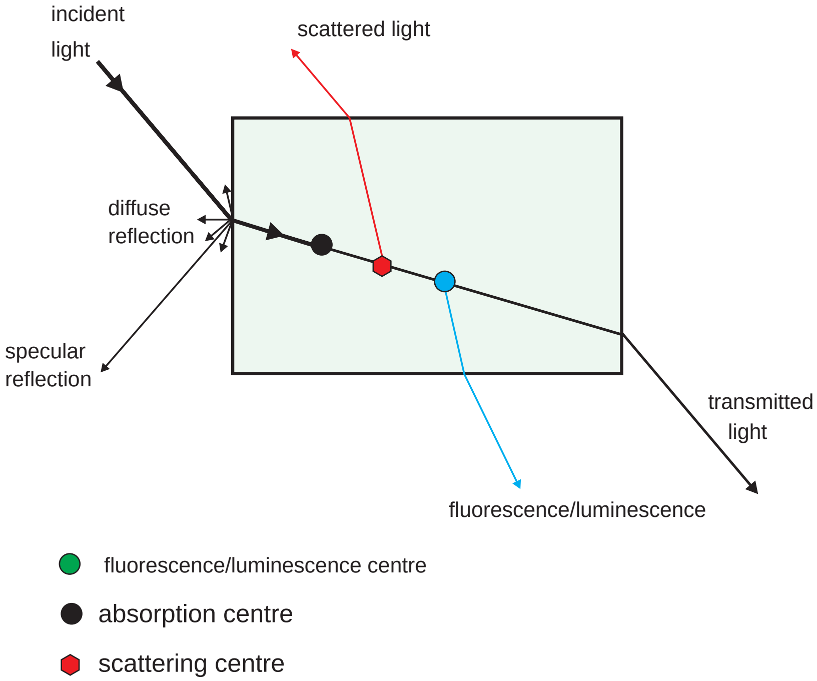


Figure 14.13 The interaction of light with a solid.

yellow, the object will transmit red light. This process is called *subtractive coloration*. There are three *subtractive primary colours*, which, when blended produce the subtractive colour spectrum. These are *cyan*, which is red-absorbing, *magenta*, which is green-absorbing and *yellow*, which is blue-absorbing. An object containing the appropriate amounts of the three subtractive primary colours will absorb all of the light falling onto its surface and appear *black* to the eye. Printing inks function by way of subtractive absorption.

Metals are a particular category of opaque solids. They 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. This prevents the light from penetrating much below the surface of the metal.

The appearance of a surface is not only bound up with the colour leaving it, but also with the texture of the surface. A smooth surface reflects light and looks shiny, even if coloured, because a certain amount of specular reflection takes place. A rough surface exhibits diffuse reflection. 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 whereas, apart from the notable exceptions of copper and gold, most smooth metal films have a silvery appearance. 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.

14.3.3 Attenuation

As a beam of light passes through a material it gradually loses intensity due to scattering or absorption, a process generally called *attenuation* (*extinction*). When attenuation takes place in a homogeneous solid, the amount of light transmitted by a plate of thickness x is given by:

$$I_x = I_0 \exp(-\alpha_e x) \tag{14.1}$$

where I_x is the irradiance leaving the plate,[3] I_0 is the incident irradiance, and α_e is the (*Napierian*) *linear attenuation coefficient* (*extinction coefficient*). Equation (14.1) 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*:

$$\log\left(\frac{I_x}{I_0}\right) = -\varepsilon\, c\, x$$

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 εcx 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:

$$\begin{aligned} I_0 &= I_r + I_t \\ 1 &= R + T \end{aligned}$$

In a pure liquid the Beer–Lambert law is often written in the form:

$$\log\left(\frac{I_x}{I_0}\right) = -ax$$

where $a = \varepsilon c$ is the *molar (decadic) attenuation (or absorption) coefficient*.

14.4 Refraction and dispersion

14.4.1 Refraction

Refraction is the apparent bending of a ray of light when it enters a transparent material such as water

[3] The older terminology to specify the amount of light, intensity, has now been replaced by more specific terms. Here the term irradiance, units W m^{-2}, is used to specify the radiant power incident upon a square metre of surface. The word intensity will be sometimes employed as an imprecise descriptive term for the amount of light in view of its historical use. The symbol I is used for irradiance instead of the recommended symbol E to avoid confusion with the use of E for energy.

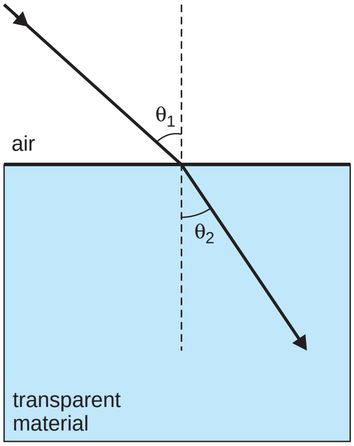


Figure 14.14 Refraction of a light beam on entering a transparent solid.

or glass (Figure 14.14). The magnitude of the deviation is given by the *index of refraction* or *refractive index*, *n*, where:

$$n = \frac{\sin\theta_1}{\sin\theta_2}$$

θ_1 being called the *angle of incidence* and θ_2 the *angle of refraction* (Table 14.3). 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:

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1}$$

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 many 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* upon 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:

$$n = \frac{\mathrm{c}}{\mathrm{v}}$$

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:

$$\nu\lambda = \mathrm{v}$$
$$n = \frac{c}{\mathrm{v}} = \frac{\lambda_\mathrm{v}}{\lambda_\mathrm{s}}$$

Table 14.3 Refractive indices

Substance	Refractive index*	Substance	Refractive index*
(Vacuum)	1.0 (definition)	Dry air, 1 atm. 15°C	1.00027
Water	1.3324	Na_3AlF_6 (cryolite)	1.338**
MgF_2	1.382**	Fused silica (SiO_2)	1.4601
KCl (sylvite)	1.490	Crown glass	1.522
Extra light flint glass***	1.543	NaCl (halite)	1.544
Flint glass***	1.607	MgO (periclase)	1.735
Dense flint glass***	1.746	Al_2O_3 (corundum)	1.765**
ZrO_2 (zirconia)	2.160**	C (diamond)	2.418
$CaTiO_3$ (perovskite)	2.740	TiO_2 (rutile)	2.755**

*A value appropriate to the yellow light emitted by sodium atoms, the sodium D-lines, with an average wavelength 589.3 nm, is given.
**The refractive index varies with direction; the average value is given.
***The flint glasses contain significant amounts of lead oxide, PbO, as follows: extra light flint, 24 wt.% PbO; flint, 44 wt.% PbO; dense flint, 62 wt.% PbO.

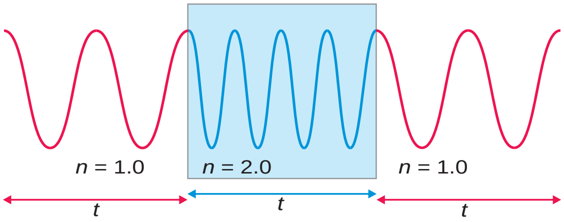


Figure 14.15 The effect of refractive index on the wavelength of light. The wavelength is compressed in materials with a high refractive index.

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 a vacuum (Figure 14.15).

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:

$$[d] = n_1t_1 + n_2t_2 + n_3t_3 + \ldots$$

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 14.16). If θ_c is exceeded then *no light escapes* and all behaves as if it were reflected from the surface. This effect is called *total internal reflection.*

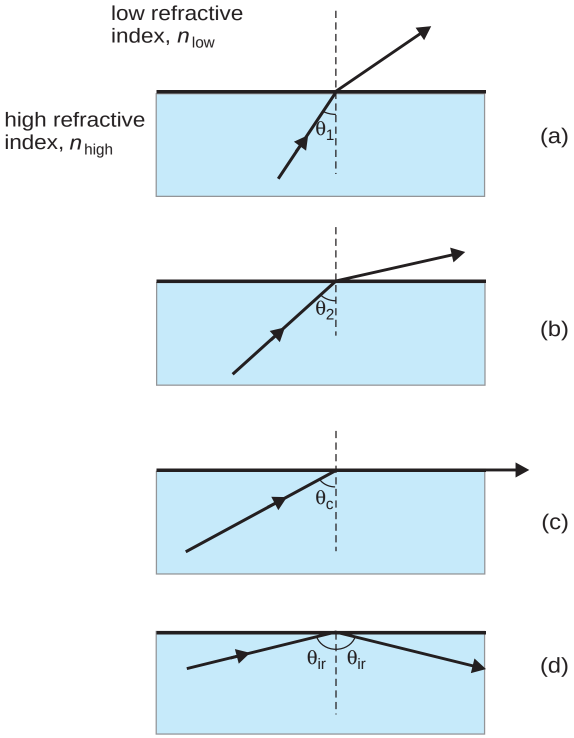


Figure 14.16 Total internal reflection of a beam of light in a solid with a higher refractive index than the surrounding medium.

The critical angle follows from the general relation given above when θ_i is equal to 90°:

$$\sin \theta_e = \frac{n_{\text{low}}}{n_{\text{high}}}$$

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 much 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.

14.4.2 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 11.1.4). 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:

$$n = 1 + \rho(p_1 k_{r1} + p_2 k_{r2} + p_3 k_{r3} \ldots)$$

or

$$n = 1 + \rho \sum p_i k_{ri}$$

where ρ is the density of the complex oxide. The factor k_r, called the *refractive coefficient*, is an empirically determined constant (Table 14.4). The amount of each oxide is taken into account by multiplying the refractive coefficient by its *weight fraction* in the compound, p.

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.

14.4.3 The refractive index of metals and semiconductors

The refractive index of metals and many semiconductors is related to the considerable numbers of free electrons present and is best expressed as a complex number:

$$N = n + \mathrm{i}k$$

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.

14.4.4 Dispersion

The refractive index of a solid varies with wavelength (Figure 14.17). 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 $\mathrm{d}n/\mathrm{d}\lambda$, which is the slope of the refractive index versus wavelength

Table 14.4 Refractive coefficients for some oxides*

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

[*] These refractive coefficients apply when the density of the material is measured in $\mathrm{g\,cm^{-3}}$. The use of density in $\mathrm{kg\,m^{-3}}$ requires that the coefficients above are each multiplied by 10^{-3}.

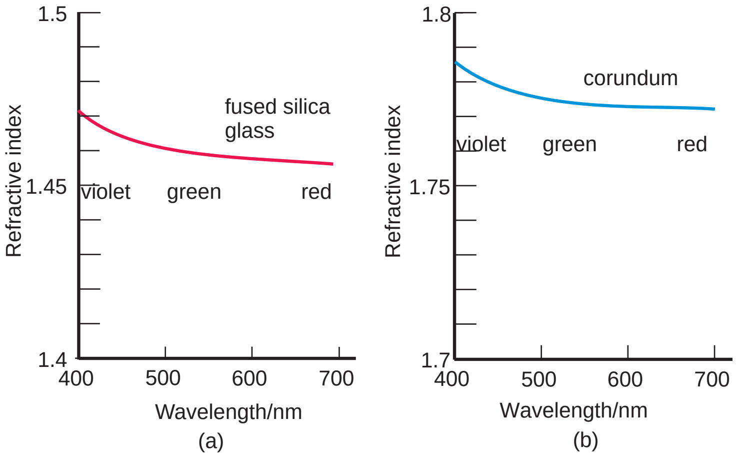


Figure 14.17 Dispersion: (a) fused silica glass; (b) corundum, Al_2O_3.

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 above no longer applies and the solid is said to exhibit *anomalous dispersion*.

14.5 Reflection

14.5.1 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:

$$r = \frac{n_0 - n_1}{n_0 + n_1}$$

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 14.18). 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

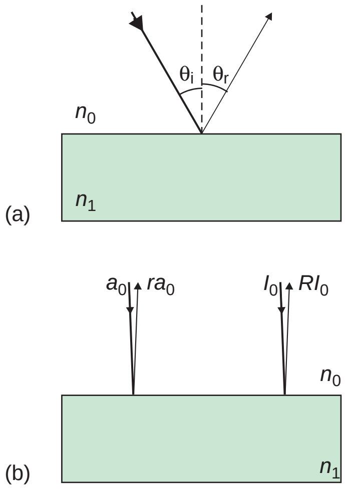


Figure 14.18 The reflection of light: (a) the angle of incidence, θ_i is equal to the angle of reflection, θ_r; (b) the coefficient of reflection at normal incidence, r, relates the amplitude reflected to the incident amplitude, while the reflectivity, R, relates the irradiance reflected to the incident irradiance.

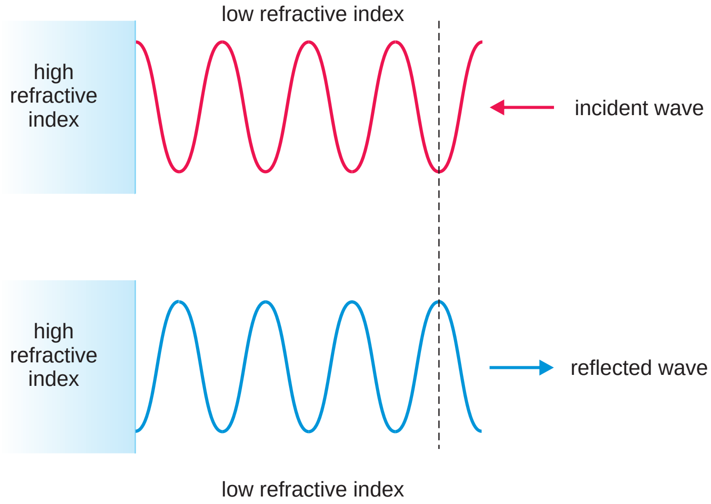


Figure 14.19 A light beam incident upon a plate with a higher refractive index than the surrounding medium suffers a phase change on reflection, so that an incident peak is reflected as a trough.

of a light wave, that a peak turns into a trough on reflection (Figure 14.19).

The eye detects *irradiance* changes rather than amplitude changes, and so it is more convenient to work with the *reflectivity* or *reflectance*, R:

$$R = r^2 = \left(\frac{n_0 - n_1}{n_0 + n_1}\right)^2$$

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:

$$R = \left(\frac{1 - n}{1 + n}\right)^2 = \left(\frac{n - 1}{n + 1}\right)^2$$

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, $N = n_1 + \mathrm{i}k$. In this case, the reflectivity of a surface at *normal incidence* is:

$$R = \frac{(n_1 - n_0)^2 + k^2}{(n_1 + n_0)^2 + k^2}$$

For the case of a surface in air this becomes:

$$R = \frac{(n_1 - 1)^2 + k^2}{(n_1 + 1)^2 + k^2}$$

14.5.2 Reflection from a single thin film

Monochromatic light travelling through air falling upon a homogeneous thin film of an insulator 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 (Figure 14.20). At each reflection, some of the light will escape to produce additional reflected and transmitted rays. Because of the difference in the paths taken by the repeatedly reflected rays, the waves will interfere with each other.

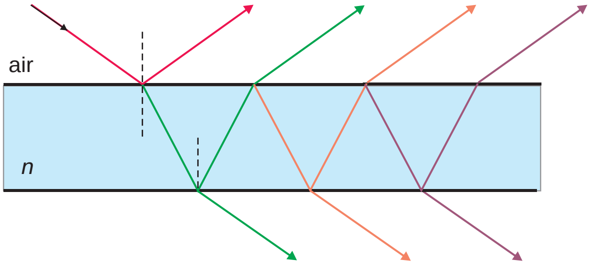


Figure 14.20 Multiple reflections at the upper and lower surfaces of a thin film.

As the reflectivity of single surfaces is rather small, the first reflected ray and the first transmitted ray are of most importance. The appearance of the film will largely depend upon the extent of the interference between these two waves. In the case of an observer looking down on a film in a perpendicular direction, some of the light incident on the surface and seen by the observer will have been reflected at the top surface (wave 1, Figure 14.21a). In addition, some light travels through the film and is reflected from the bottom surface before reaching the observer (wave 2, Figure 14.21a). In addition, because wave 1 is reflected at a surface of higher refractive index a wave peak will turn into a trough. Interference between the two waves will occur, which will cause the film to look either dark or bright. The optical path difference between waves 1 and 2, often called the *retardation*, will be [*p*]:

$$[p] = 2[d] = 2nt$$

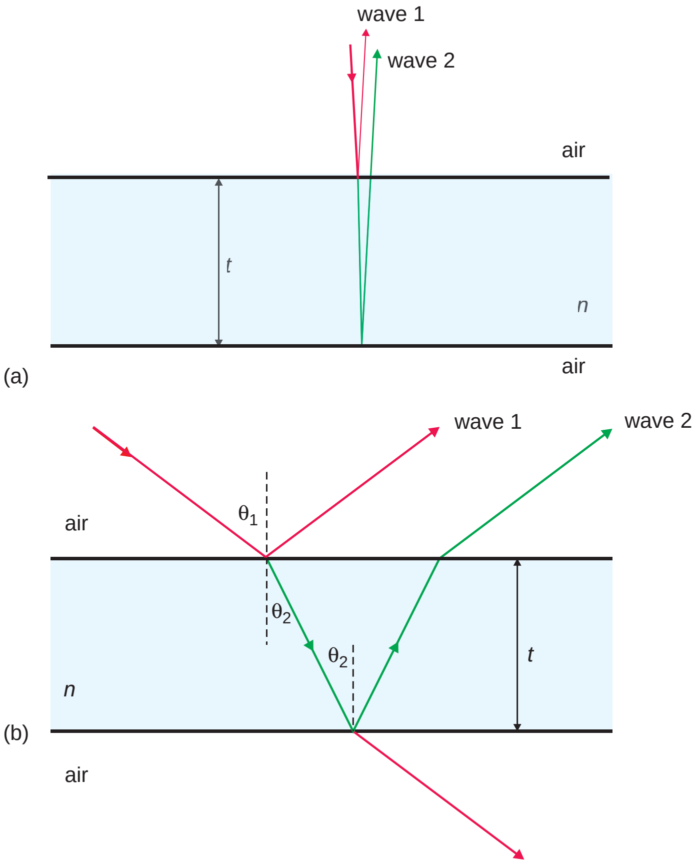


Figure 14.21 Reflection at a thin film in air: (a) perpendicular incidence; (b) incidence at an angle.

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.

$$\begin{aligned}[p] &= 2nt \\ &= m\lambda \quad (m = 1, 2, 3, \ldots) \quad \textit{minimum}\ \text{(dark)}\end{aligned}$$

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:

$$\begin{aligned}[p] &= 2nt \\ &= (m + {}^{1}/_{2})\lambda \quad (m = 1,2,3,\ldots) \quad \textit{maximum}\ \text{(bright)}\end{aligned}$$

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 an angle of refraction θ_2, (Figure 14.21b), 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):

$$[p] = 2nt \cos\theta_2 = m\lambda \qquad \textit{minimum}\ \text{(dark)}$$
$$[p] = 2nt \cos\theta_2 = (m + \tfrac{1}{2})\,\lambda \quad \textit{maximum}\ \text{(bright)}$$

14.5.3 The reflectivity of a single thin film in air

The reflectivity of a thin film in air will be different from that of a single surface as reflection and subsequent interference effects from the bottom surface also need to be considered. For light at normal incidence on a transparent solid, the reflectivity, R, of a homogeneous non-absorbing thin film on a substrate, illuminated by light of one wavelength perpendicular to the surface, is given by:

$$R = \frac{r_1^2 + 2r_1r_2\cos 2\delta + r_2^2}{1 + 2r_1r_2\cos 2\delta + r_1^2r_2^2}$$

where

$$r_1 = \frac{n_0 - n_f}{n_0 + n_f},\ r_2 = \frac{n_f - n_s}{n_f + n_s},\ \delta = \frac{2\pi[d]}{\lambda}$$

n_0 is the refractive index of the surrounding medium, n_f the refractive index of the film, n_s the refractive index of the substrate and $[d]$ is the optical thickness of the film. The reflectivity varies in a cyclic fashion, with zero for values of $[d]$ equal to 0, $\lambda/2$, λ, and so on, and a maximum for values of $[d]$ given by $\lambda/4$, $3\lambda/4$ (Figure 14.22). Because the value of n depends on wavelength, the reflectivity will also vary across the spectrum.

14.5.4 The colour of a single thin film in air

When a thin transparent film is viewed in white light, the same reflection and interference discussed above 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

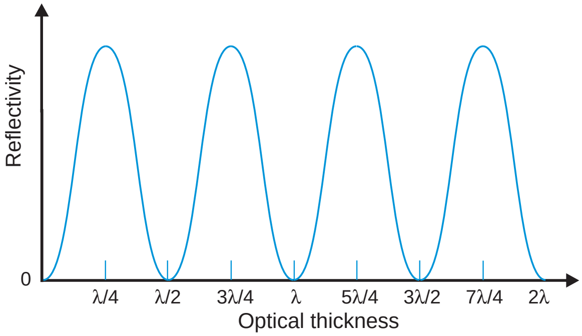


Figure 14.22 The sinusoidal variation of the reflectivity of a thin film.

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 with every 550 nm of retardation (Table 14.5, Figure 14.23).

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 14.5).

If the angle of viewing is not perpendicular to the film, the optical path difference is given by:

$$[p] = 2[d]\cos\theta_2 = 2nt\cos\theta_2$$

where θ_2 is the angle of *refraction*. This formula indicates that as the viewing angle moves away from perpendicular to 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.

14.5.5 The colour of a single thin film on a substrate

The behaviour of a single thin film on a substrate is similar to that discussed for the case 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 14.5.

Table 14.5 The colour of a thin film in white light

Path difference[*]/nm	Colour reflected[†]	Colour transmitted[‡]
	Start of 1st 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 1st order;*	*start of 2nd 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 2nd order;*	*start of 3rd order*

Beyond this point the film colour is less intense and ultimately becomes pale pink or pale green in reflection.

[*] The path difference p is often called the *retardation*.

[†] This colour is seen in *reflection* from a thin film in air when illuminated by white light at normal incidence.

[‡] 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.

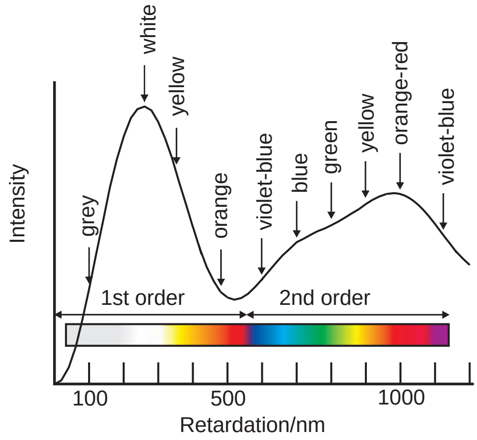


Figure 14.23 The approximate colour of a thin film when viewed from directly above as the film thickness varies.

14.5.6 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 (Section 14.5.2). For values of $[d]$ given by $\lambda/2$, λ, $3\lambda/2$, and so on:

$$R = \left(\frac{n_0 - n_s}{n_0 + n_s}\right)^2$$

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 (Section 14.5.1). Thus a layer of optical thickness $\lambda/2$, for example, can be considered to be *optically absent*.

For values of $[d]$ given by $\lambda/4$, $3\lambda/4$, and so on:

$$R = \left(\frac{n_f^2 - n_0 n_s}{n_f^2 + n_0 n_s}\right)^2$$

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 spectrum 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.

14.5.7 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 formula for the reflectance of such a mirror refers to the specific case in which all layers

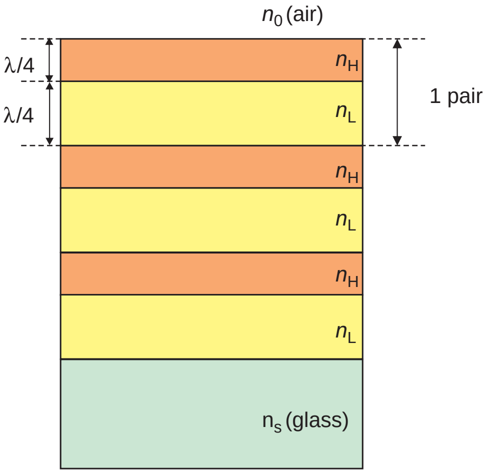


Figure 14.24 A quarter-wave stack of alternating layers of high and low refractive index solids, n_H and n_L, each of optical thickness one quarter of a wavelength.

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 14.24). 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:

$$R = \left(\frac{n_s f - n_0}{n_s f + n_0}\right)^2$$

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:

$$R = \left[\frac{n_s - (n_L/n_H)^{2N}}{n_s + (n_L/n_H)^{2N}}\right]^2$$

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 14.25)

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’.

14.6 Scattering

14.6.1 Rayleigh scattering

If a transparent medium 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 principal models, *Rayleigh scattering* and *Mie scattering*. Rayleigh scattering applies to spherical insulating particles with a diameter less than about a tenth of the wavelength of the incident light. When each photon in a beam of unpolarised light of irradiance I_0 is scattered only once, the irradiance of the scattered light I_s at a distance d from the scattering centre is:

$$I_s = I_0\left(\frac{9\pi^2 V^2}{2d^2\lambda^4}\right)\left(\frac{m^2-1}{m^2+2}\right)^2(1+\cos^2\theta)$$

where 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:

$$m = \frac{n_p}{n_m}$$

In this case, n_p is the refractive index of the particle and n_m the refractive index of the surrounding medium. For air, n_m is 1.0.

As much light is scattered backwards as forwards, and only half as much is scattered normal to the beam direction (Figure 14.26). 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. The blue appearance of the sky is the result of the preferential scattering of violet light combined

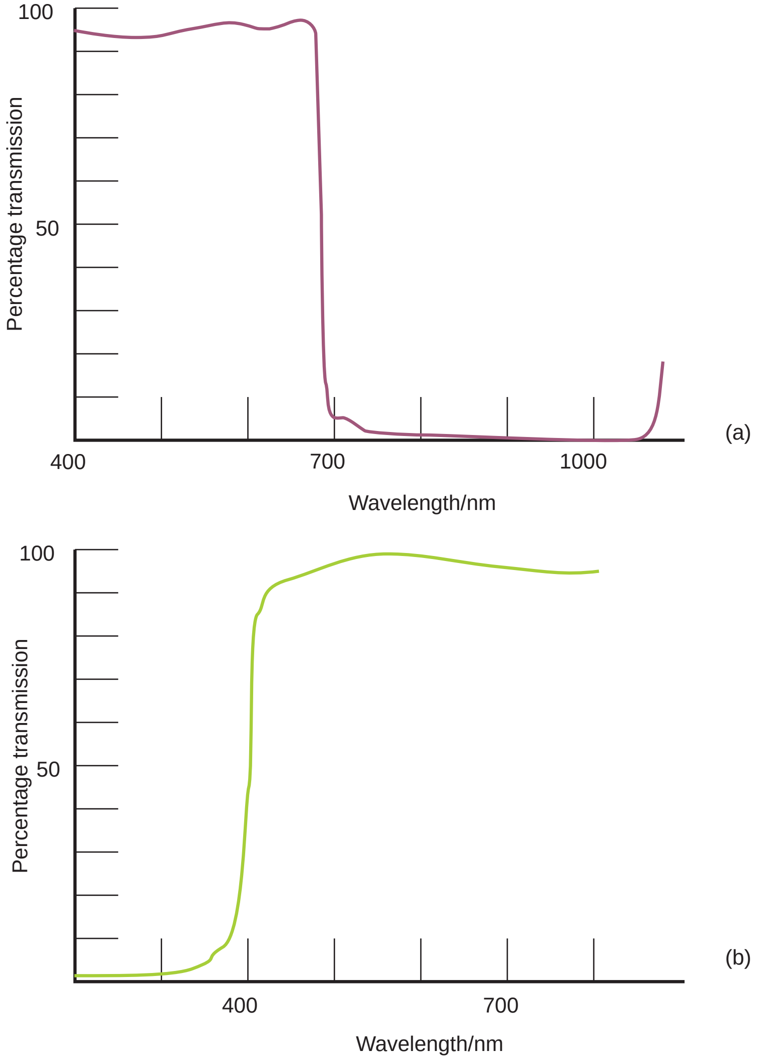


Figure 14.25 Interference filters: (a) a shortpass filter; (b) a longpass filter; (c) a bandpass filter.

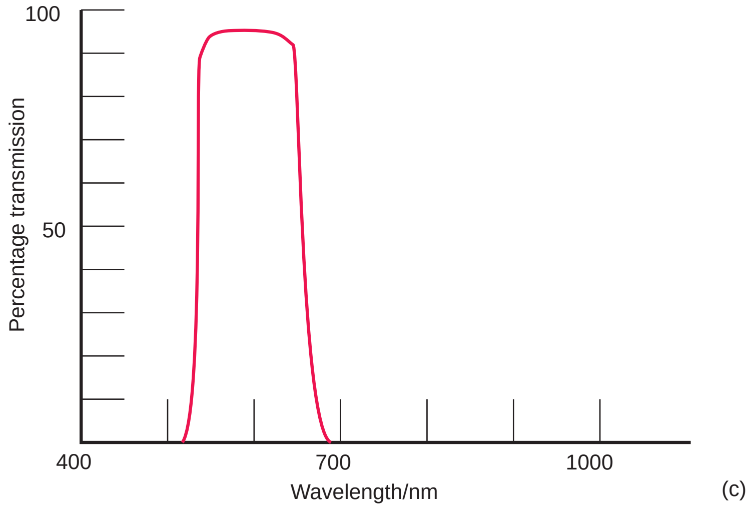


Figure 14.25 (*Continued*)

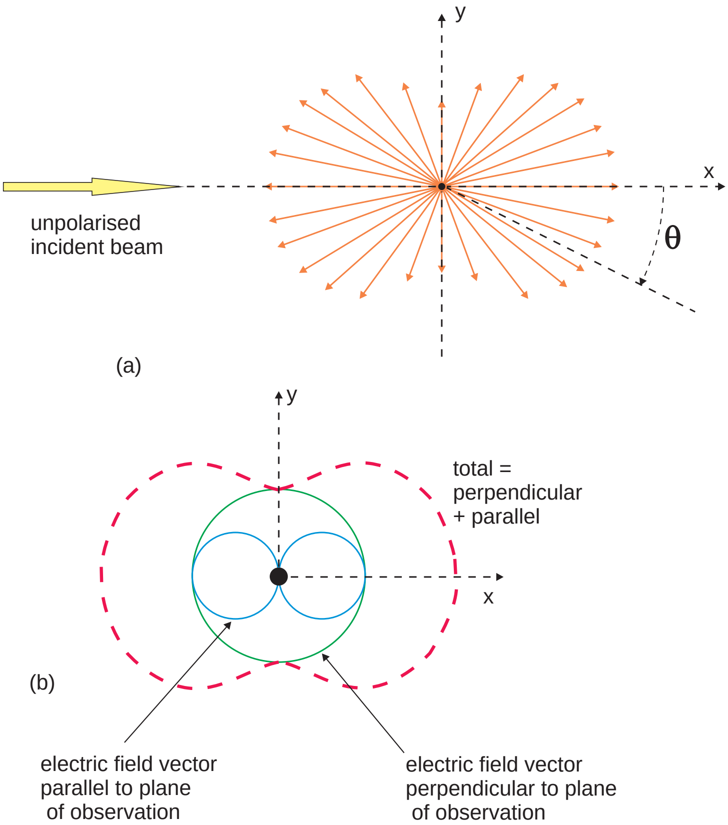


Figure 14.26 Rayleigh scattering: (a) the scattering pattern from a small spherical particle; (b) the pattern is the sum of radiation scattered with its electric field vector parallel and perpendicular to the plane of observation.

with the maximum sensitivity of the eye, which lies in the green-yellow. Similarly, the red skies visible at dawn and dusk, and the rare blue moon, are caused by scattering of light from small particles in the upper atmosphere.

14.6.2 Mie scattering

The term Mie scattering is reserved for scattering by particles that are larger than those for which Rayleigh scattering is valid, about a third the wavelength of light or more. It is also applicable to metallic as well as insulating bodies. The colour of ruby glass is due to Mie scattering from a dispersion of gold nanoparticles in the glass.

As the particle size increases from that appropriate to Rayleigh scattering, forward scattering begins to dominate over backward scattering (Figure 14.27). 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. 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 fogs and mists.

14.7 Diffraction

Diffraction occurs when waves interact with objects having a size similar to the wavelength of the radiation. 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 a set of bright and dark fringes, due to constructive and destructive interference, called a *diffraction pattern.*

14.7.1 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

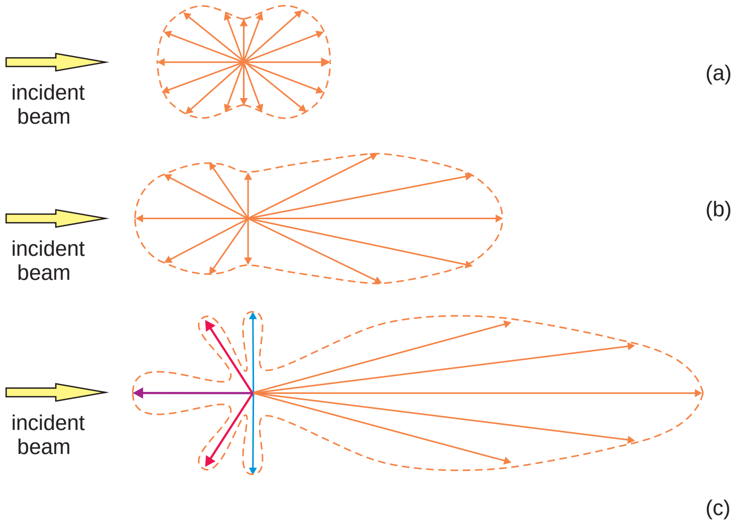


Figure 14.27 Scattering: as particle size increases, the scattering pattern changes from (a) Rayleigh scattering to (b, c) Mie scattering, in which the scattering pattern becomes asymmetrical and develops side lobes.

given by the expression:

$$I_x = I_0 \left[\frac{\sin x}{x}\right]^2$$

where

$$x = \frac{\pi w \sin\theta}{\lambda}$$

and 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:

$$\sin\theta_{\min} = \frac{m\lambda}{w}$$

where m takes values 1, 2, 3 . . . . For $\theta_{\min}$ 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 14.28). The positions of the maxima between these dark bands are not given by a simple formula, but are approximately midway between the minima.

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.

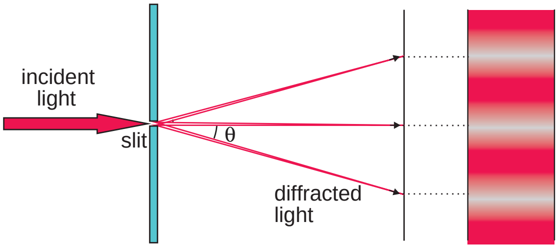


Figure 14.28 The diffraction pattern from a thin slit.

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 as described above.

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:

$$\sin\theta = \frac{m\lambda}{d}$$

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 computation of m requires rather sophisticated mathematics, the results of which show that 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 which 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.

14.7.2 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

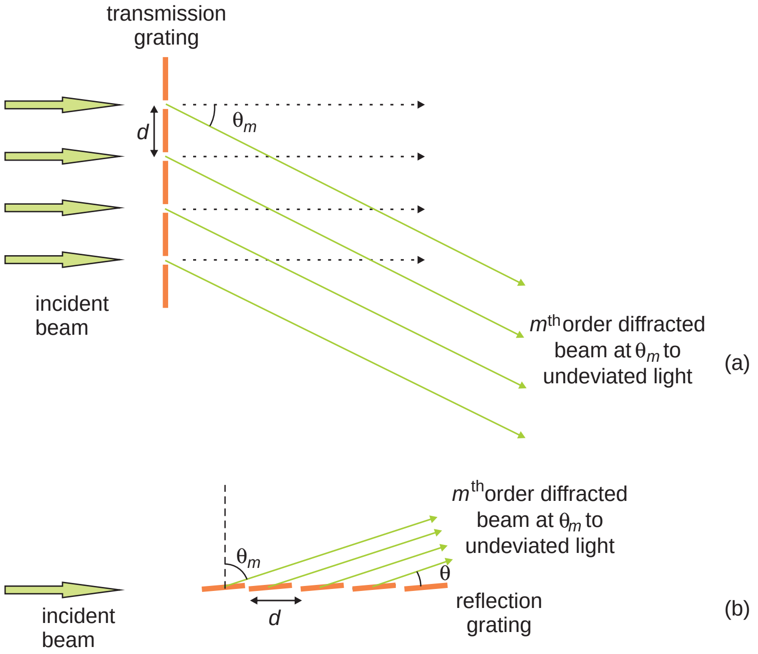

transmission grating
d
θ_m
incident beam
m^{th} order diffracted beam at θ_m to undeviated light
(a)
m^{th} order diffracted beam at θ_m to undeviated light
θ_m
θ
reflection grating
incident beam
d
(b)

Figure 14.29 Diffraction: (a) a transmission grating; (b) a reflection grating.

diffraction effects are observed in the light reflected from the patterned surface (Figure 14.29). The effectiveness of a grating is the same whether light is transmitted through it or reflected from it.

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:

$$\sin\theta_m = \frac{m\lambda}{d}$$

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:

$$1 - \sin\theta_m = \frac{m\lambda}{d}$$

$$1 - \cos\theta = \frac{m\lambda}{d}$$

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.

14.7.3 Diffraction from crystal-like structures

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

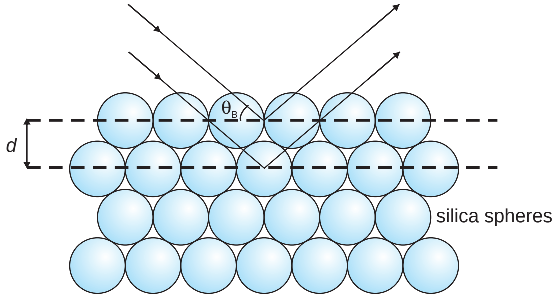


Figure 14.30 The diffraction of light by ordered arrays of silica spheres gives colour to precious opals.

between the beam and the scattering planes[4] (Section 5.2.2). The theory holds for any three-dimensional array no matter the size of the 'atoms'. Thus, an arrangement of particles or voids which 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.

The colour of precious opal arises in this way. The regions producing the colours are made up of an ordered packing of silica spheres that resemble crystallites embedded in amorphous silica (Figure 14.30). These interact with light when 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 above. The refractive index of the silica in opal, n_s, is about 1.45, giving:

$$m\lambda = 2n_s d \sin\theta_B \approx 2.9\, d \sin\theta_B$$

The longest wavelength diffracted back to the viewer by the opal is when light falls perpendicular to the diffracting layers:

$$\lambda_{max} = 2n_s d \approx 2.9\, d$$

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 (Section 5.4.1), the relationship between the sphere radius and the layer spacing is:

$$d = \frac{2\sqrt{2r}}{\sqrt{3}} = 1.633r$$

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.

14.7.4 Photonic crystals

Photonic crystals are artificial 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 a one-dimensional photonic crystal is a stack of transparent layers of alternating refractive indices similar to dielectric mirrors (Section 14.5.7). 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 14.31). When a beam of light is incident on such a grating, a single wavelength, λ, will be diffracted when Bragg's Law is obeyed:

$$\lambda = 2nd \sin\theta_B$$

where n is the refractive index of the material and d

[4] This equation is slightly different in form to equation (5.1). In this latter equation the order of the diffraction, m, is taken into account via the interplanar spacing of hkl planes. Both forms of the equation give the same result.

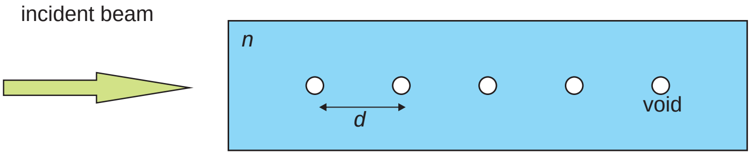


Figure 14.31 A one-dimensional photonic crystal can be thought of as a regularly-spaced linear array of diffracting centres such as small voids in a transparent solid.

the repeat spacing. For a beam normal to the voids, sin $\theta = 1$, hence:

$$\lambda = 2nd$$

and the beam will be diffracted back on itself.

This same idea was used in the description of electron waves in crystals (Section 2.3.3), the result of which was the creation of an energy band gap for the electrons. 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 a 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.

14.8 Fibre optics

14.8.1 Optical communications

The transmission of light along thin fibres of glass, plastic or other transparent materials is referred to as *fibre optics*. Data are carried by a series of pulses of light encoded so that information can be stored and retrieved. In this brief survey, the properties of the materials used in the fibre will be outlined.

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 (Section 14.4.1). 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 14.32). The cladding should not be confused with an outer plastic protective covering, which has no optical role to play.

14.8.2 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:

$$\text{loss} = -10\log_{10}\left(\frac{\text{power in}}{\text{power out}}\right)$$

$$= -10\log_{10}\left(\frac{P(x)}{P(0)}\right)$$

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

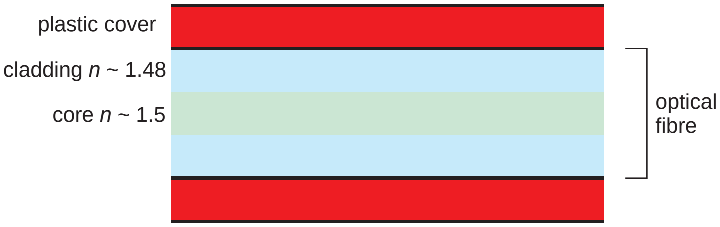


Figure 14.32 The structure of a silica glass optical fibre.

the decibel, dB. The attenuation is defined as the loss per kilometre, thus:

$$\text{attenuation} = \frac{-10}{x}\log_{10}\left(\frac{P(x)}{P(0)}\right)$$

For a material showing an attenuation of 1 dB km^{-1}, an input power of 10 watts would give an output power of 7.9 watts 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 by 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 14.6.1) 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 easily be 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, do 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. This is currently the industry standard.

14.8.3 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:

$$\text{dispersion} = \frac{t_f - t_i}{x}$$

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 14.33). 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.

The earliest fibres were called *stepped index multimode* fibres. These fibres have a large core region, allowing many modes to propagate (Figure 14.34a). The ray labelled H (Figure 14.34a) is known as a *high-order mode* while 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 described below.

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 14.34b). 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 14.34c) 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.

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

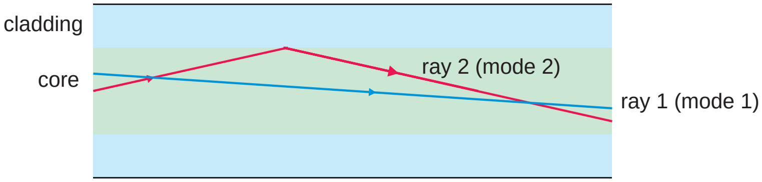


Figure 14.33 The allowed paths that light beams can take in the core region of an optical fibre are called modes. Although drawn as ray paths, in reality they are alternative light wave patterns in the core.

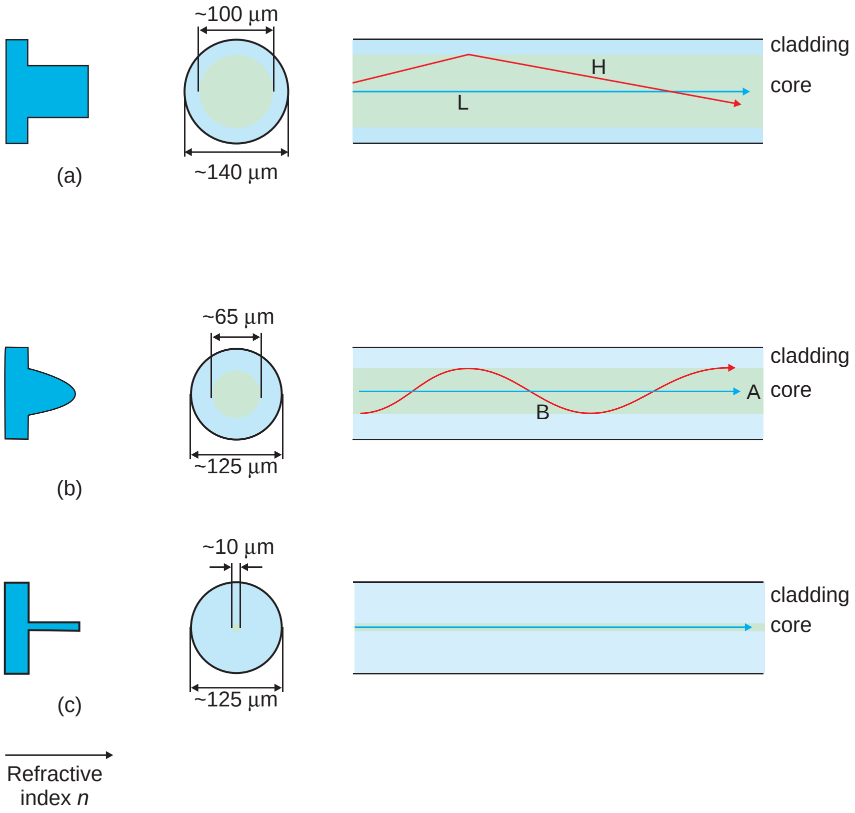


Figure 14.34 Types of optical fibre: (a) stepped index fibre; (b) graded index fibre; (c) monomode fibre.

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 research laboratories.

14.8.4 Optical amplification

The amplification of signals in fibre optic transmission is achieved in a stretch of about 30 metres of monomode fibre core containing just a few hundred parts per million of Er^{3+} (Figure 14.35a). 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 nm and 1480 nm. The erbium ions have the remarkable ability to transfer energy from the laser to the signal pulses as they traverse this section of fibre.

The energy transfer comes about in the following way. 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 14.35b). 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 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

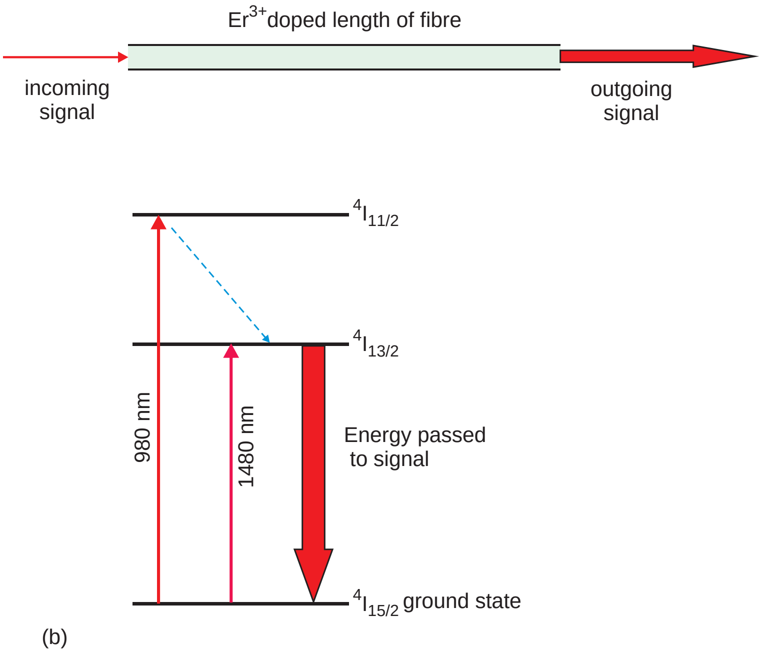


Figure 14.35 Signal amplification: (a) an incoming weak signal is amplified on passage through a length of fibre in which the core has been doped with Er^{3+} ions; (b) pump wavelengths of 980 nm and 1480 nm (upward-pointing arrows) excite the Er^{3+} ions into the $^4I_{13/2}$ state which transfers energy to the signal (downward-pointing arrow).

close to 1480 nm, can empty it via *stimulated emission* (Section 14.2.4). This achieves signal amplification while retaining the coherence of the pulse constituting the signal.

14.9 Energy conversion

14.9.1 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 meters, exposure meters and automatic shutters in cameras. The first photographic exposure meters used selenium, cadmium sulphide or silicon. In the case of selenium, the photovoltage is large enough to be measured directly and converted to an exposure value. Cadmium sulphide and silicon need voltage amplification, and these materials 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 (Figure 14.36a). The junction is not connected to any external power source. Holes and electrons produced in the junction

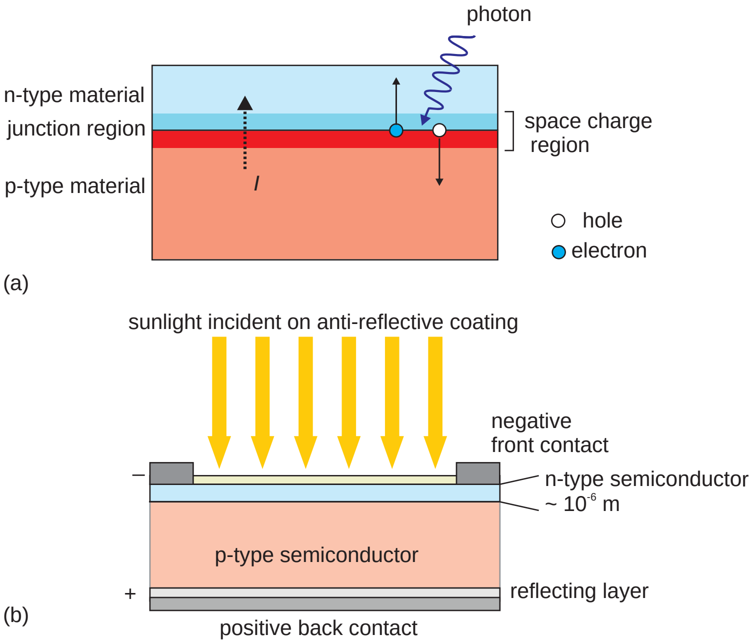


Figure 14.36 Solar cells, schematic: (a) sunlight falling on a p-n junction creates electron–hole pairs that are swept into the external circuit by the field in the junction region; (b) operating cells use a thin antireflection coating on the front surface, a thin n-type layer, a junction region near to the front surface and a reflecting layer below the cell to increase efficiency.

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. Should an external load, R, be connected, some current can pass through it, and so do useful work.

To achieve high efficiency the photons need to be absorbed 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, recombination is likely. In addition, solar cells require an antireflection surface to maximise the number of photons reaching the semiconductor, a thin initial semiconductor layer, so that optimum number of photons reach the depletion layer, and an underlying reflective layer that redirects any photons that pass straight through the lower semiconductor layer back towards the junction region (Figure 14.36b).

Because impurities and defects trap mobile electrons and holes, 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 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 use silicon. Single crystal silicon, an indirect band gap solid, is unsuitable and is replaced by amorphous silicon, which behaves as a direct band gap material. Currently many other cell materials are being investigated, including cadmium telluride, cadmium sulphide, copper indium selenide and semiconductor quantum dots (Section 14.10.2).

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 590–1200 nm.

14.9.2 Dye sensitized 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 sensitized 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 sensitizer, 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 sensitizer dye, S, by a photon: $S + h\nu \rightarrow S^*$.

(ii) The excited sensitizer, S^*, loses an electron, which moves into the conduction band of the semiconductor: $S^* \rightarrow S^+ + e^-$ (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 sensitizer is regenerated by reaction with the reduced half of the redox couple: $S^+ + R^- \rightarrow S + R$ (Figure 14.37a).

A large number of different dyes have been tried in the role of sensitizer, 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 whilst 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/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 14.37b). 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.

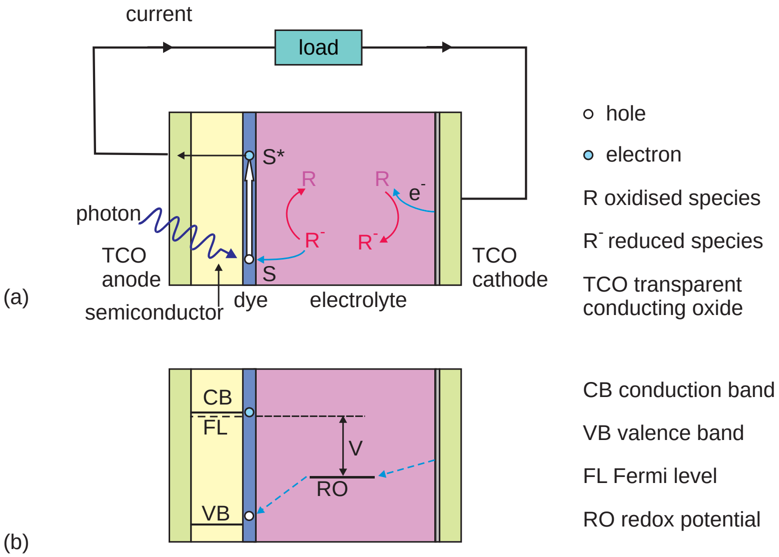


Figure 14.37 Dye sensitized solar cell schematic. (a) Sunlight absorbed by the dye liberates an electron into the semiconductor. The dye is regenerated by interaction with an internal redox couple. (b) Energy levels in a cell.

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). Active research in this area means that new cell specifications are continually appearing in the literature.

14.10 Nanostructures

14.10.1 The optical properties of quantum wells

In a quantum well, the electrons and holes occupy electron and hole sub-bands (Section 13.5.1, Figure 14.38). When electrons in the upper energy levels drop to the lower levels in *interband transitions*, a photon is emitted. The energy separation of the sub-bands 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 with the bulk.

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 sub-bands. Using equation 13.6:

$$E(\text{photon}) = E_g + \left(\frac{h^2}{8a^2}\right)\left(\frac{n^2}{m_e^*} + \frac{n^2}{m_h^*}\right)$$

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:

$$E(\text{photon}) = E_g + \frac{n^2h^2}{4a^2m^*}$$

The selection rule for the transition is $\Delta n = 0$; that is, transitions only take place between levels with

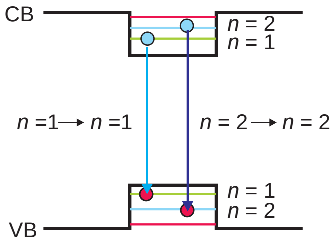


Figure 14.38 Interband transitions (schematic) between electron (upper) and hole (lower) sub-bands.

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 sub-band*. Holes can make similar transitions between levels in the *hole sub-band*. These transitions, which give rise to extra peaks in the emission spectrum, are known as *intersub-band transitions*.

Because the dimensions of the quantum well can be changed, the emission spectrum can be varied or *tuned*. This feature is called *bandgap 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 14.39). A change in the composition of the SQW active layer allows the colour emission to vary between 450 nm blue to 600 nm yellow.

14.10.2 The optical properties of nanoparticles

The optical properties of photoluminescent nanoparticles, which behave as quantum dots, have been extensively investigated because they emit fluorescent light that is a precise function of the dimensions of the quantum dot. 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 14.40).

The colour variation comes about because the conduction and valence bands of the bulk semiconductor are transformed into a closely lying set of 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 14.41a). To produce fluorescent light, electrons are excited from the lower band to the upper band with ultraviolet radiation. The electrons in higher energy levels subsequently lose energy by non-radiative

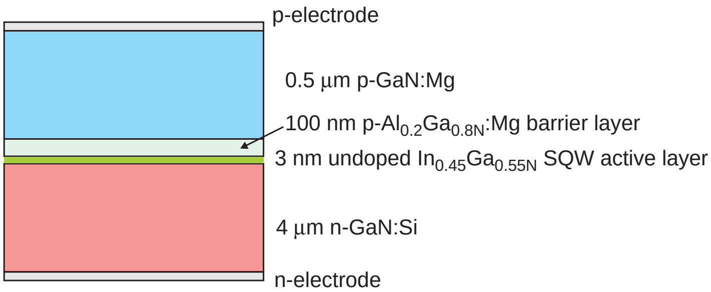

p-electrode
0.5 µm p-GaN:Mg
100 nm p-$Al_{0.2}Ga_{0.8N}$:Mg barrier layer
3 nm undoped $In_{0.45}Ga_{0.55N}$ SQW active layer
4 µm n-GaN:Si
n-electrode

Figure 14.39 Green-emitting single quantum well (SQW) active layer LED (schematic).

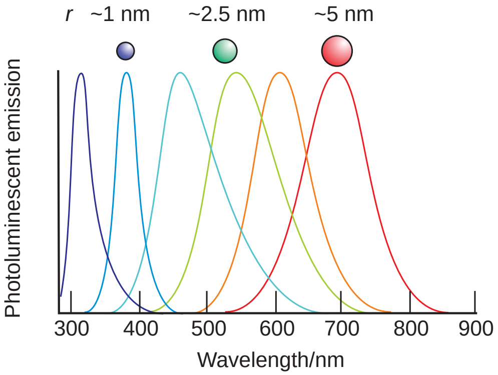


Figure 14.40 Photoluminescent colours emitted by CdS quantum dots.

transitions to end in the lowest energy level of the upper set. A photon is then emitted as the electron drops to the topmost energy level of the lower set (Figure 14.41b).

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, 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.

Besides roughly spherical dots, many other dot geometries are being created, including rods, dipods, tetrapods and so-called flowers. All are being tested for applications in medicine and biology, photovoltaics and optical computing.

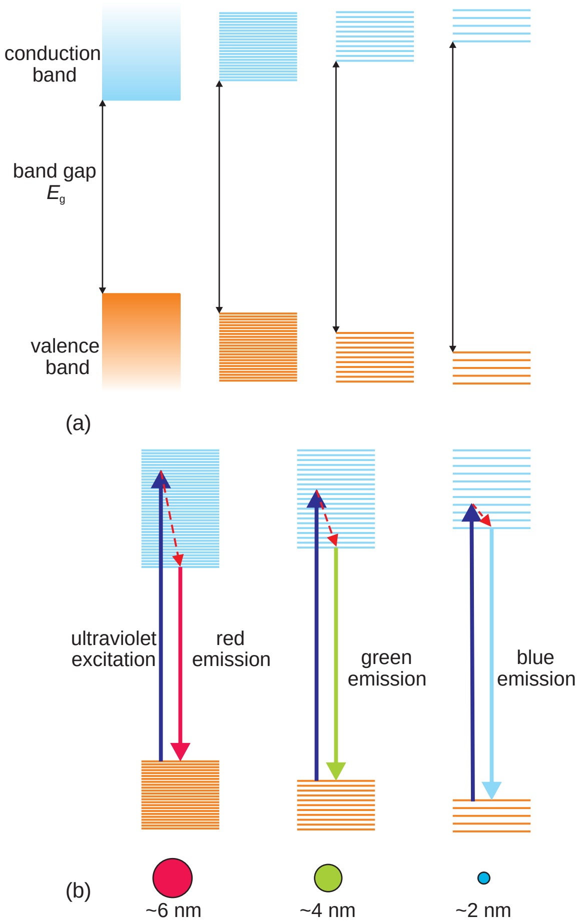


Figure 14.41 Quantum dot colours: (a) the change in band structure of a quantum dot as the diameter falls; (b) fluorescence colours of different diameter dots (schematic).

Further reading

General

Heavens, O.S. and Ditchburn, R.W. (1993) *Insight into Optics*. John Wiley & Sons, Ltd., Chichester.

M. Fox, M. (2001) *Optical Properties of Solids*. Oxford University Press, Oxford.

The properties of light with respect to colour are found in:

Nassau, K. (2001) *The Physics and Chemistry of Colour*, 2nd edn. John Wiley & Sons, Ltd., New York.

Tilley, R.J.D. (2011) *Colour and the Optical Properties of Materials*, 2nd edn. John Wiley & Sons, Ltd., Chichester.

The engineering aspects of optical fibres are described by:

Hecht, J. (1999) *Understanding Fiber Optics*, 3rd edn. Prentice Hall, Upper Saddle River, New Jersey.

A series of articles on Photovoltaics is to be found in *Materials Research Society Bulletin*, XVIII (October) 1993.

A series of articles on Photonic Materials for Optical Communications is to be found in *Materials Research Society Bulletin*, 28, May 2002.

Problems and exercises

Quick quiz

1 The long wavelength part of the electromagnetic spectrum is associated with:
(a) Radiowaves.
(b) X-rays.
(c) Infrared radiation.

2 The short wavelength region of the visible spectrum is associated with the colour:
(a) Red.
(b) Violet.
(c) Green.

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.

4 The emission of light by a solid at high temperatures is called:
(a) Luminescence.
(b) Incandescence.
(c) Phosphorescence.

5 The light produced by spontaneous emission is:
(a) Coherent.
(b) Incoherent.
(c) Partly coherent.

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.

7 Electron transitions involved in colour (optical transitions) are:
(a) Spin-allowed transitions.
(b) Phonon assisted transitions.
(c) Radiationless transitions.

8 The four-level neodymium laser utilises:
(a) p electron transitions.
(b) d electron transitions.
(c) f electron transitions.

9 Mixing the three additive primary colours gives:
(a) Black.
(b) White.
(c) Colourless.

10 Mixing the three subtractive primary colours gives:
(a) Black.
(b) White.
(c) Colourless.

11 Which of the following subtractive primary colours absorbs *green*:
(a) Cyan.

(b) Yellow.

(c) Magenta.

12 A ray of light passes from 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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

22 The cladding of an optical fibre consists of:

(a) Air.

(b) Plastic.

(c) Glass.

23 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.

24 The ‘spreading out’ of a pulse in an optical communications fibre is called:

(a) Pulse dispersion.

(b) Modal dispersion.

(c) Wavelength dispersion.

25 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.

Calculations and questions

14.1 Calculate the frequency and energy of photons associated with wavelengths 425 nm, 575 nm and 630 nm. What colours are attributed to these wavelengths?

14.2 Light of wavelength 400 nm is shone through a gas of absorbing molecules. Calculate the energy absorbed by one mole of gas if each molecule absorbs one photon.

14.3 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.

14.4 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 of light that could cause this decomposition to occur?

14.5 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?

14.6 Calculate the wavelength at which the rates of spontaneous and stimulated emission become equal at 300 K.

14.7 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?

14.8 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.

14.9 What is the separation of the energy levels in neodymium ions that give rise to laser lines at (a) 0.914 μm and (b) 1.06 μm.

14.10 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? In order to make an orange LED with an emission at a wavelength of 600 nm, it is proposed to make a solid solution $Ga_xAl_{1-x}As$. (b) Taking the variation in band gap with composition as linear, what value of x is required?

14.11 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? In order to make a green LED with an emission at a wavelength of 525 nm, it is proposed to make a solid solution $Ga_xIn_{1-x}N$. (b) Taking the variation in band gap with composition as linear, what value of x is required?

14.12 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? The band gap of InP is 1.27 eV and that of AlP is 2.45 eV.

14.13 A solution is quoted as having a 22% transmittance. What is the absorbance?

14.14 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 is 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?

14.15 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 at.%.

14.16 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?

14.17 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?

14.18 Estimate the refractive indices of the ceramics (a) barium titanate, $BaTiO_3$, (b) lead titanate, $PbTiO_3$, using the Gladstone–Dale formula. Densities: $BaTiO_3$, 6017 kg m^{-3}; $PbTiO_3$, 8230 kg m^{-3}.

14.19 Estimate the refractive index of the minerals (a) spinel, $MgAl_2O_4$, (b) akermanite, $Ca_2MgSi_2O_7$, using the Gladstone–Dale formula. Densities: spinel, 3600 kg m^{-3}; akermanite, 2940 kg m^{-3}.

14.20 Estimate the refractive index of the minerals (a) beryl, $Be_3Al_2(SiO_3)_6$, (b) garnet, $Mg_3Al_2Si_3O_{12}$, using the Gladstone–Dale formula. Densities: beryl, 2640 kg m^{-3}; garnet, 3560 kg m^{-3}.

14.21 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.

14.22 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, $n = 2.15$ (average).

14.23 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, $n = 1.85$, $k = 3.27$.

14.24 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.

14.25 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.

14.26 Describe the differences between the reflectivity of a soap film, thickness $\lambda/4$, in air, compared with 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.

14.27 (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, and (c) by transmission.

14.28 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.

14.29 Determine the reflectivity of a $\lambda/4$ film of (a) silicon oxide, $n = 2.0$, and (b) titanium dioxide, $n = 2.775$, on glass, $n = 1.504$, in air.

14.30 (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)?

14.31 (a) Plot a graph of reflectivity versus 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%?

14.32 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?

14.33 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.

14.34 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.

14.35 Determine the 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.

14.36 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.

14.37 The linear scattering coefficient of limestone dust in the air is $0.0002194\,\text{m}^{-1}$. Over what distance will the irradiance of a light beam diminish to 10% of its initial value?

14.38 What is the relative amount of light scattered (a) in the incident direction, (b) at 45° to the incident direction, (c) perpendicular to the incident direction, (d) in the reverse direction, for Rayleigh scattering.

14.39 (a) What slit width is needed in a transmission diffraction grating to cause red light, $\lambda = 700\,\text{nm}$, to be deviated by 7.5° to the normal?

(b) What will be the deviation of violet light, $\lambda = 400\,\text{nm}$?

14.40 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?

14.41 A photonic crystal is made by laying down 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) close-packed layers, how will the result change?

14.42 Estimate the attenuation of an optical fibre in which 0.5% of the initial power is lost in a distance of 1 km.

14.43 The attenuation of ordinary window glass is of the order of $10^6\,\text{dB km}^{-1}$. What thickness of glass would cause the incident light intensity to fall to 50%?

14.44 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), 1.46265; n(700 nm) 1.45530.

14.45 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.

14.46 A film of zinc sulphide 4 nm in thickness forms on metallic zinc. Treating the film as a quantum well, what is the wavelength of the transition $\Delta n = 2$? The bandgap of ZnS is 3.54 eV and the effective mass of both electrons and holes is 0.4 $\mathrm{m_e}$.

15

Thermal properties

- 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.

15.1 Heat capacity

15.1.1 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:

$$C = \frac{\Delta Q}{\Delta T}$$

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°. The heat capacity is frequently quoted as the *molar heat capacity*: the amount of heat that increases the temperature of 1 mole of substance by 1°, or the *specific heat capacity*, the amount of heat that increases the temperature of 1 g of a substance by 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 15.1). 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:

$$C_p = 106.6 + 0.0178T - \frac{2.85 \times 10^6}{T^2}$$

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

Table 15.1 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

in the temperature range from 300–1800 K (Figure 15.1a).

15.1.2 Classical theory of heat capacity

The heat capacity of most metals was found to be approximately 25 J mol^{-1} K^{-1} as long ago as 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 $^1/_2 k_B T$, and as there are N_A atoms per mole, the molar heat capacity is:

$$C_v(\text{classical}) = 6N_A \frac{1}{2} k_B T = 3R \approx 25\ \text{J mol}^{-1}\ \text{K}^{-1}$$

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.

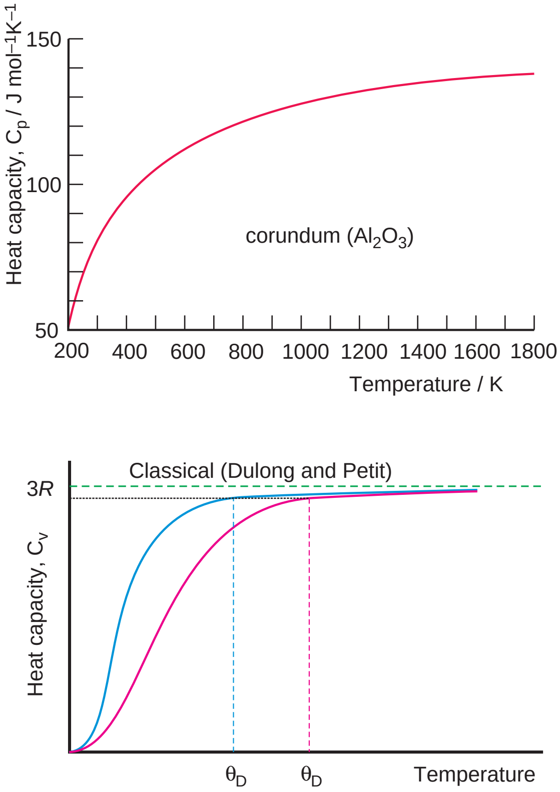


Figure 15.1 The variation of heat capacity of a solid with temperature: (a) experimental C_p data for corundum (Al_2O_3); (b) Debye theory C_v for two elements with different Debye temperatures (schematic).

15.1.3 Quantum theory of heat capacity

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:

$$E = (n + \frac{1}{2})h\nu \qquad (15.1)$$

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 equation (15.1) shows that the lowest energy possible is not zero, as it is for a classical system, but is equal to $^1\!/_2 h\nu$ 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 statistics 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 2.3.2). 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 the 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:

$$C_\mathrm{v} = 9R\left(\frac{T}{\theta_\mathrm{D}}\right)^3 \int_0^{\theta_\mathrm{D}/T} \frac{x^4 e^x}{(e^x - 1)^2}\,dx$$

$$x = \frac{h\nu}{k_\mathrm{B}T}$$

$$\theta_\mathrm{D} = \frac{h\nu_\mathrm{D}}{k_\mathrm{B}}$$

where θ_D is the *Debye temperature*, ν_D is the *Debye frequency*, R is the gas constant and T the temperature (Figure 15.1b)

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 15.1b). Note that the Debye temperature drops systematically on moving down a periodic table group (Table 15.1).

At temperatures well *below* the Debye temperature, the value of C_v is given by:

$$C_\mathrm{v} = \frac{12\pi^4}{5} R\left(\frac{T}{\theta_\mathrm{D}}\right)^3$$

That is, $C_\mathrm{v} = AT^3$ where A is a constant.

15.1.4 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 (Section 8.2), order–disorder changes (Sections 8.4, 11.3.5, 11.3.6), or when a ferromagnetic solid becomes paramagnetic (Sections 12.1.2, 12.3.1). In all of these cases the curve of heat capacity versus 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 versus temperature curve will show a discontinuity (Figure 15.2a).

A second-order transformation is similar, but in these cases no latent heat of transformation appears. The heat capacity curve will show a discontinuity,

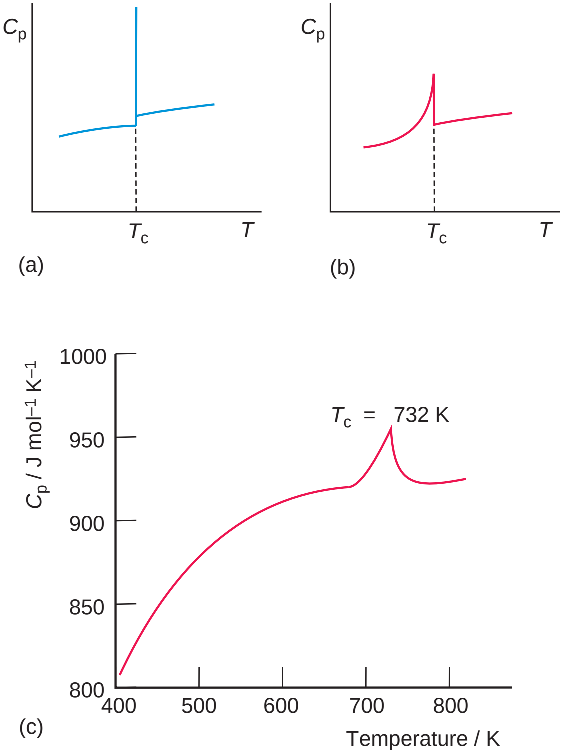


Figure 15.2 Variation of heat capacity (C_p) of a solid at a phase transition: (a) first-order (schematic); (b) second-order (schematic); (c) data for $SrFe_{12}O_{19}$ with a second-order transition from ferrimagnetic to paramagnetic at 723 K. Part of (c) is drawn from data in Rakshit *et al.* (2007) *J. Solid State Chem.*, 180: 523–32.

but it will not be infinite (Figure 15.2b). These breaks are often used to characterize second-order transitions such as those involved in ferroelectric and ferromagnetic transformations (Figure 15.2c).

15.2 Thermal conductivity

15.2.1 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. The equations describing heat transfer, initially formulated by Fourier, pre-date and are of identical mathematical form to Fick's laws of diffusion (Sections 7.1, 7.2, 7.3).

In the case of steady-state heat transfer, the one-dimensional heat transfer equation is:

$$J_Q = -K\frac{dQ_c}{dx} \tag{15.2}$$

where J_Q is the heat flux, dQ_c/dx is the heat concentration gradient along the direction of heat flow, x, and K is the *thermal diffusivity*, defined by:

$$K = \frac{\kappa}{\rho\, c_p}$$

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 equation 15.2 are identical in form to those given in Section 7.3 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:

$$\frac{dT}{dt} = K\frac{d^2T}{dx^2}$$

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 diffusion equation (Section 7.2).

15.2.2 Thermal conductivity of solids

Thermal conductivity is generally determined under steady-state conditions using equation (15.2). When the two ends of a solid are held at different temperatures, heat flows across the body from the hot to the cold side (Figure 15.3a). (This figure is the analogue

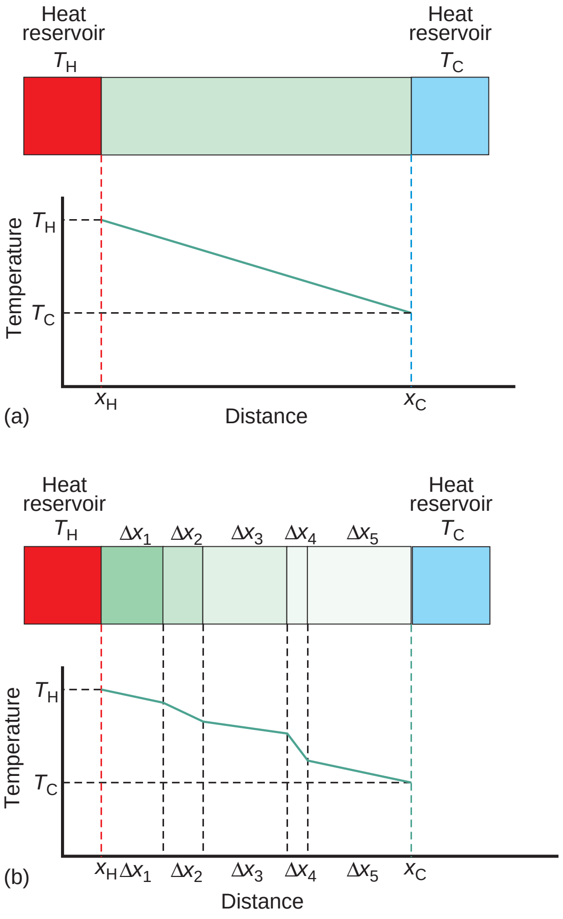


Figure 15.3 Thermal conductivity: (a) the temperature gradient across a solid is a measure of its thermal conductivity; (b) the temperature drop across a series of different materials will vary with the thermal conductivity of each.

of Figure 7.6.) 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 ($x_H - x_C$):

$$Q_t = \kappa A \left(\frac{T_H - T_C}{x_H - x_C} \right)$$

$$Q_t = \kappa A \frac{dT}{dx} \qquad (15.3)$$

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:

$$Q_t = A \sum \kappa_i \frac{\Delta T_i}{\Delta x_i}$$

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 (Figure 15.3b).

The thermal conductivity of solids varies considerably (Table 15.2, Figure 15.4). 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 hexagonal metal cadmium (A3 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 W m^{-1} K^{-1} and 6.5 W m^{-1} K^{-1} parallel to the **a**-axis.

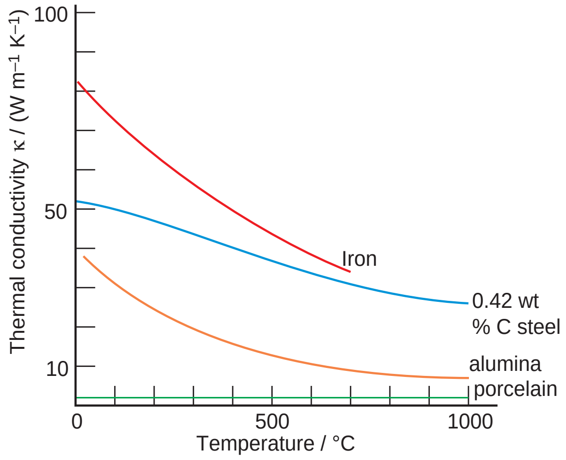


Figure 15.4 The variation of thermal conductivity with temperature of some common materials.

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 boron nitride and aluminium nitride (Table 15.2).

15.2.3 Thermal conductivity and microstructure

Thermal conductivity is mainly attributed to the mobile electrons present in the solid and the phonon spectrum:

$$\kappa = \kappa(\text{electrons}) + \kappa(\text{phonons})$$

Mobile electrons make the greatest contribution, and so metals would be expected to show a much higher thermal conductivity than insulators. At the simplest level, the electrons can be imagined as a free electron gas, moving with a velocity that is higher at the hot end of the solid than the cold end. (The same model was mentioned in Section 2.3.2 with respect to electrical conductivity.) 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 the thermal conductivity of these depends upon the defects present, which in itself is the result of the thermal and mechanical history of the sample, the equation is usually written:

$$\kappa = L_0 \, \sigma \, T + C$$

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:

$$\kappa = 2.22 \times 10^8 \, \sigma \, T + 10.5$$

Thermal conductivity in insulators, 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. Defects in a structure drastically 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 by oxidation in air. This can react with the bulk nitride to introduce substitutional defects and vacancies in the following way:

$$2SiO_2(s) \rightarrow 2Si_{Si} + 4O_N + V_{Si}$$

where Si_{Si} represents silicon atoms on normal silicon sites (not defects), O_N represents oxygen atoms on nitrogen atom sites (substitutional defects), and V_{Si} represents a vacancy on a silicon site (vacancy defects). 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.

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 defects such as 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 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 disordered 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.

In many materials thermal conductivity is hampered by the presence of 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 (equation 15.3), it is possible to define an interfacial thermal conductivity γ by:

$$Q_t = \gamma A \frac{\mathrm{d}T}{\mathrm{d}x}$$

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 and both molecular dynamics simulations and experimental studies indicate that suitable chemical bonding in the interfacial region can improve γ.

15.3 Expansion and contraction

15.3.1 Thermal expansion

Most materials increase in volume as the temperature is increased, a feature called *thermal expansion*. There are, though, an increasing number of solids known that contract as the temperature increases.

Thermal expansion is a most important property in practice. It is used in most everyday thermometers. The shattering of ordinary glass on being cooled rapidly is due to the thermal contraction of the outer layers. This can be avoided if the glass has a low thermal expansion, such as Pyrex® glass or fused silica. The thermal expansion of components in electronic devices is important, and the difference in thermal expansion of materials in building construction can lead to grave difficulties. The coincidence of the thermal expansion of steel and concrete at normal temperatures allows the use of steel-reinforced concrete in buildings.

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:

$$\alpha_\mathrm{m} = \frac{l_\mathrm{f} - l_\mathrm{i}}{(T_\mathrm{f} - T_\mathrm{i})l_\mathrm{i}} = \frac{\Delta l}{\Delta T\, l_\mathrm{i}}$$

where l_f is the final length, l_i is the initial length and Δl is the length increment (Figure 15.5a).

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:

$$\beta_\mathrm{m} = \frac{\Delta V}{\Delta T\ V_\mathrm{i}}$$

where ΔV is the volume change and V_i is the original volume at T_i. A reasonable approximation is:

$$3\alpha_\mathrm{m} \approx \beta_\mathrm{m}$$

In the case of a solid with different mean thermal expansion coefficients along **x**-, **y**- and **z**-axes:

$$(\alpha_\mathrm{x} + \alpha_\mathrm{y} + \alpha_\mathrm{z}) \approx \beta_\mathrm{m}$$

The *linear expansivity of a solid*, α, is defined as the increase in length per unit length at a given temperature:

$$\alpha = \left(\frac{1}{l}\right)\frac{\mathrm{d}l}{\mathrm{d}T}$$

The expansivity is the *slope* of the length versus temperature curve at a length l and temperature T (Figure 15.5a), 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 15.5b).

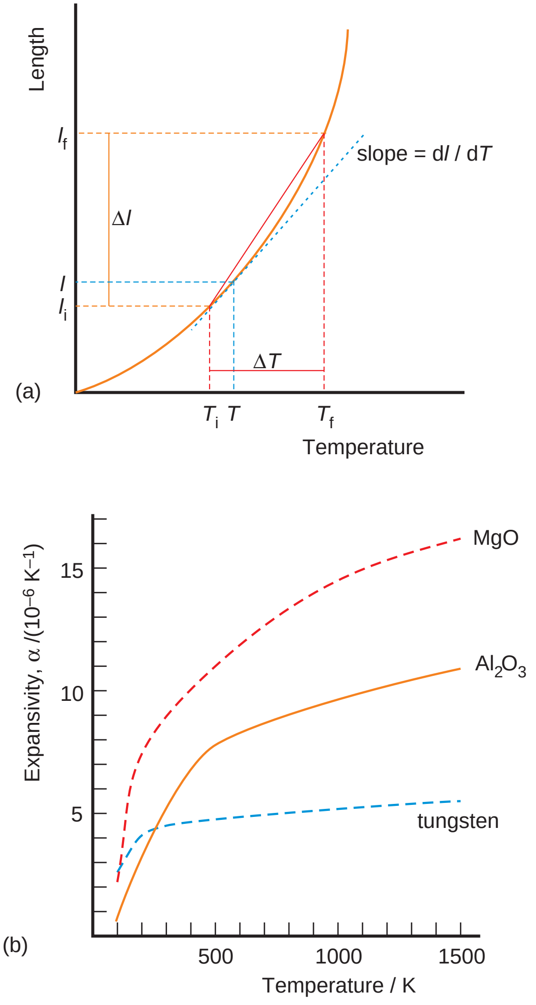


Figure 15.5 Thermal expansion: (a) The mean coefficient of thermal expansion α_m defined as $\Delta l/\Delta T\ l_i$ and the expansivity, α, defined as $(1/l)(\mathrm{d}l/\mathrm{d}T)$; (b) the approximate variation of expansivity with temperature for a metal, tungsten and two ceramics, MgO and Al_2O_3.

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 cookingware, for example, glass ceramics with negligible thermal expansion over the temperature ranges normally encountered are widely available.

15.3.2 *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 15.6a). The extent of the vibrational potential energy V_0, at a low temperature, T_0,

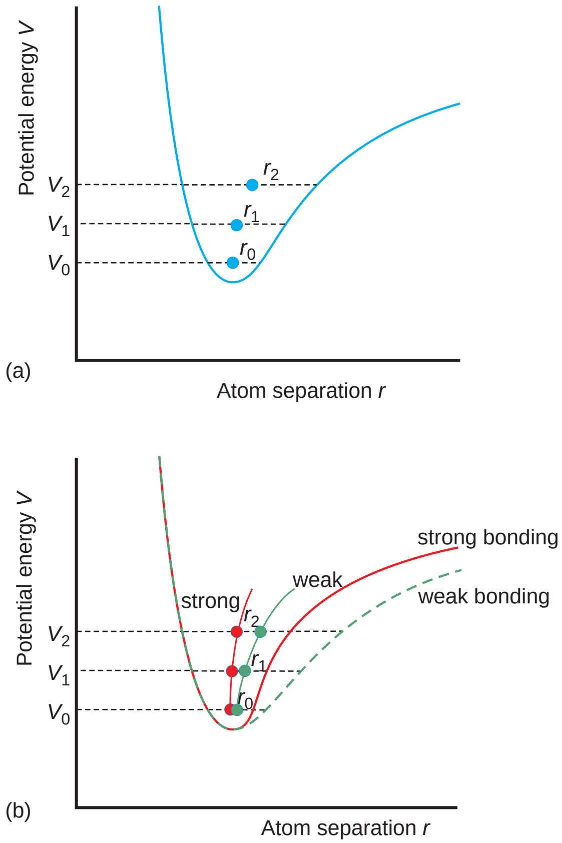


Figure 15.6 (a) The variation of the potential energy between two atoms as a function of the interatomic spacing (schematic). (b) Potential energy curves for a pair of atoms linked by either weak or strong bonds (schematic).

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.

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 15.6b). 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 15.2).

In a real solid, account has to be taken of all of the atoms in the unit cell, and the interatomic potentials between each pair of atoms has 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.

15.3.3 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 produce materials that contract as the temperature increases, sometimes called *negative thermal expansion* (NTE)

Table 15.2 Coefficients of thermal conductivity and thermal expansion

Material	Thermal conductivity/ ($\mathrm{W\,m^{-1}\,K^{-1}}$)	Thermal expansion/ ($10^{-6}\,\mathrm{K^{-1}}$)*
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
Graphite**	2000	0.6
Cubic BN	1300	–
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

*Data could not be located for cells marked –.
**Graphite thermal conductivity perpendicular to the *c*-axis.

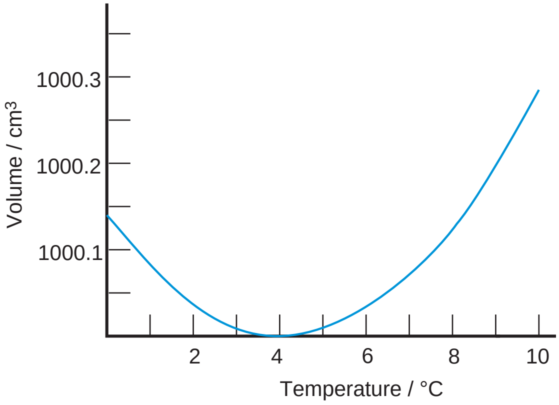


Figure 15.7 The thermal expansion/contraction of water close to 0 °C.

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 15.7). 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.

Many anisotropic crystals are known that show a contraction along one or two crystallographic axes as the temperature increases, even though other directions may show normal thermal expansion (Figure 15.8). Still others, including cubic ZrW_2O_8 and a silica polymorph with the faujasite structure, show contraction along all axes as the temperature rises. Some of the most important compounds that show thermal contraction are cordierite, $Mg_2Al_4Si_5O_{12}$, β-eucryptite, $LiAlSiO_4$, β-spodumene, $LiAlSi_2O_6$ and $NaZr_2P_3O_{12}$, (NZP). One reason for their importance is that all of these structures can be maintained over a wide composition range. For

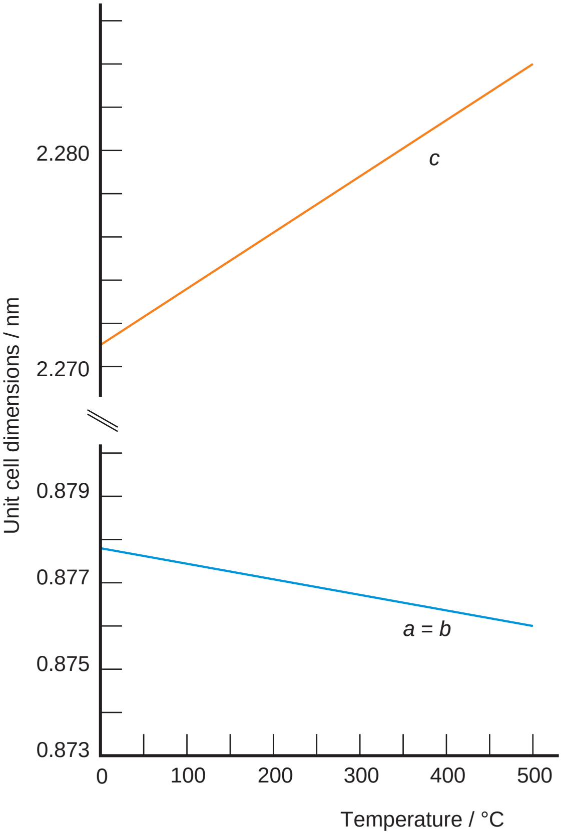


Figure 15.8 The thermal expansion of the **c**-axis and contraction of the **a**- and **b**-axes for $CaZr_4P_6O_{24}$, redrawn from data given by Agrawal (1994) (see Further reading).

example, β-spodumene can take compositions $Li_2Al_2Si_nO_{4+2n}$, in which n can take values from 4 to 9. Similarly, NZP 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, for example, $Ca_{1-x}Sr_xZr_4P_6O_{24}$. This ability allows the thermal expansion and contraction to be carefully tailored, and materials with almost zero thermal expansion can be fabricated.

Thermal contraction is not the result of a single mechanism. However, the changes can be related to the cation polyhedra that build up the structure. For example, in many ferroelectric perovskite-structure compounds, a distortion of the metal–oxygen BO_6 octahedra is responsible for the ferroelectric effect (Section 11.3.7). As

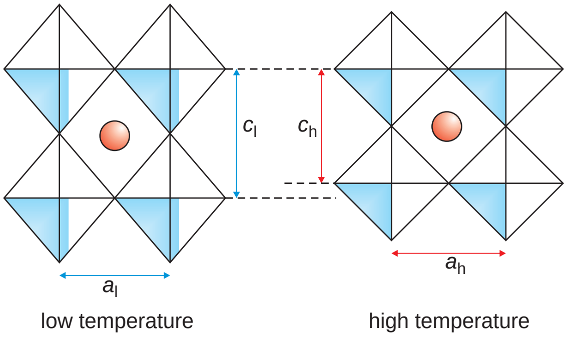


Figure 15.9 Thermal contraction due to the removal of octahedral distortion in the perovskite structure.

the temperature increases, the distortion tends to decrease, because of changes in vibrational energy and a decrease in anion–anion repulsion, so that the long diagonal shortens. At the same time, the undistorted diagonals expand normally (Figure 15.9). When the contraction outweighs the expansion, the material shows overall thermal contraction: behaviour exhibited by $PbTiO_3$.

In the families of cordierite, β-eucryptite, β-cordierite and NZP, a mechanism similar to that giving rise to auxetic (negative Poisson's ratio) materials seems to occur (Section 10.5.2). The structure is built from inflexible layers, similar to those found in clay minerals linked by Si-O-Si and O-Si-O bonds (Figure 15.10). As the temperature rises the layers expand, mostly laterally. The Si-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.

Contraction can also be achieved by 'rocking' or cooperative rotation of the polyhedra, found in cubic Zr_2WO_8, $NbOPO_4$ and $ThO(PO_4)_2$. In this group of materials, the polyhedra remain the same shape, but as the temperature rises, increased vibration allows the polyhedra to rotate, thus producing a contraction of the unit cell. In effect the polyhedra tilt in a cooperative fashion as the temperature rises, so that one or more unit cell edges contract while others may expand or also contract (Figure 15.11).

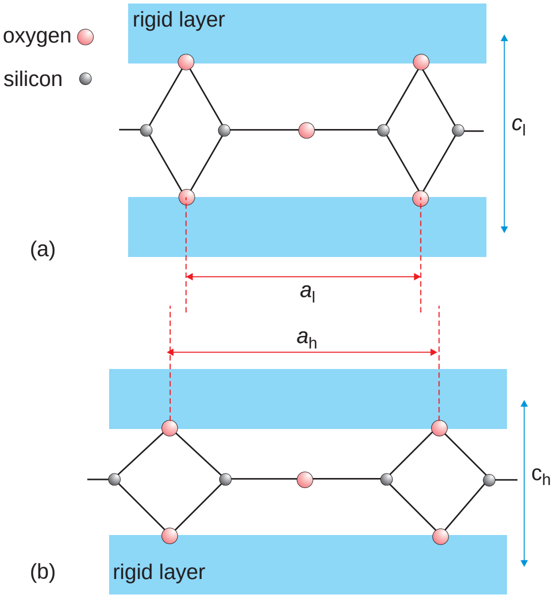


Figure 15.10 Thermal contraction brought about by silicon–oxygen 'hinges' connecting rigid sheets within a crystal structure: (a) the low-temperature structure; (b) the high-temperature structure.

15.3.4 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 $Fe[Co(CN)_6].nH_2O$ and $Ni(CH_3)_4CuZn(CN)_6.nH_2O$, 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 the antiperovskite structure in which the Cu and Ge atoms (or their equivalents) occupy the A positions, the N atoms take the

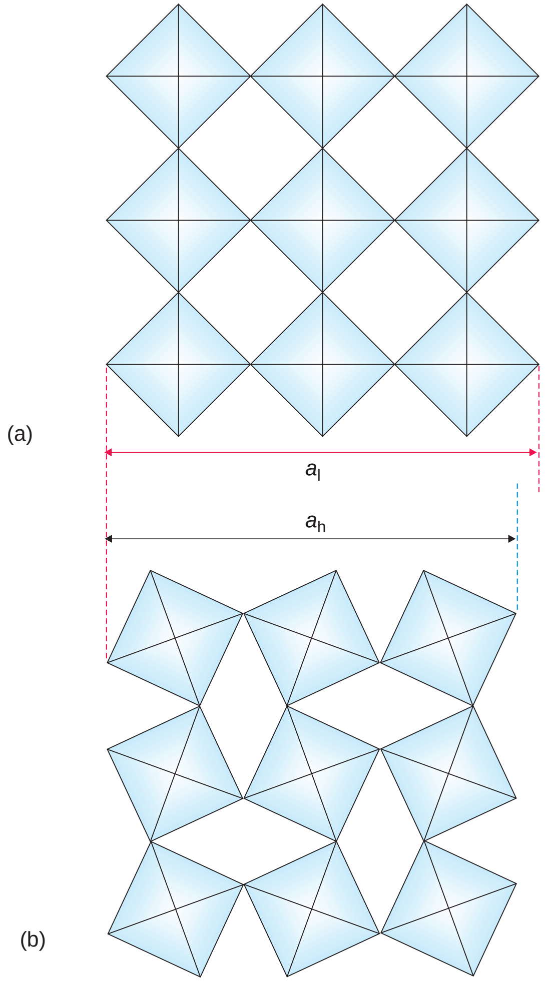


Figure 15.11 Thermal contraction due to the 'rocking' (rotation) of polyhedra: (a) the low-temperature structure; (b) the high-temperature structure.

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 normal thermal expansion of the structure on the one hand, and 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.

15.4 Thermoelectric effects

15.4.1 Thermoelectric coefficients

The conduction of electricity and the conduction of heat are closely linked. These have been treated separately in earlier chapters 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 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 15.12a). 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*, 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 $10\,\mu V\,K^{-1}$ for metals and $200\,\mu V\,K^{-1}$ for semiconductors.

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 15.12b). 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:

$$\Delta Q = \Pi_{AB}\, It$$

where Π_{AB} is the *Peltier coefficient*, I is the current flowing in the circuit and t is the time.

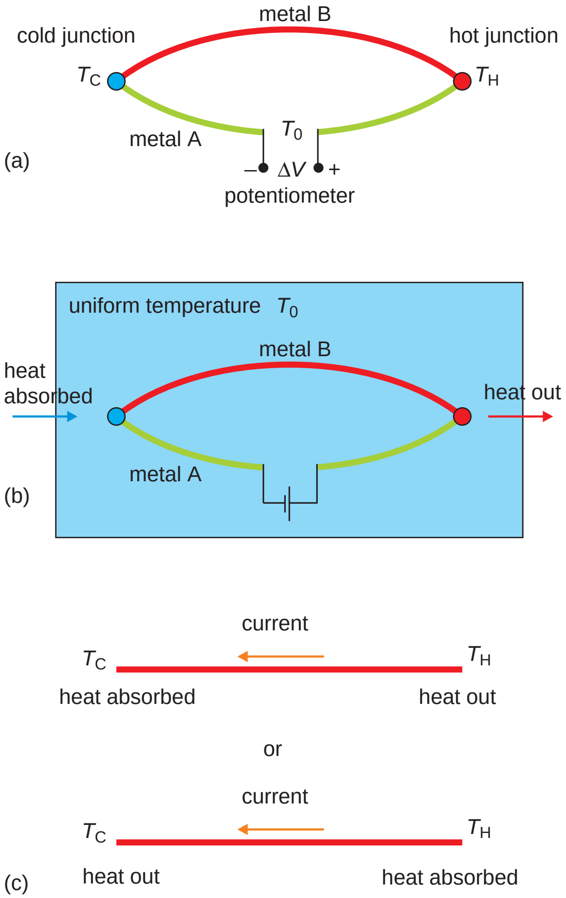


Figure 15.12 Thermovoltaic effects: (a) the Seebeck effect; (b) the Peltier effect; (c) the Thompson effect.

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 to the other (Figure 15.12c). The amount of heat energy absorbed or given out, ΔQ, is given by:

$$\Delta Q = \tau\, It\, \Delta T$$

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. Peltier or Thomson heat production is 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*, σ_S (Figure 15.13), and an *absolute Peltier coefficient* π. (The Thomson coefficient, τ, only refers to a single material.) As would be expected, these coefficients vary with direction in non-cubic crystals.

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 15.12a), the Seebeck coefficient of two materials Σ_{AB}, at a temperature T, is given by the difference between the absolute Seebeck coefficients of the components:

$$\Sigma_{AB} = \sigma_S(A) - \sigma_S(B)$$

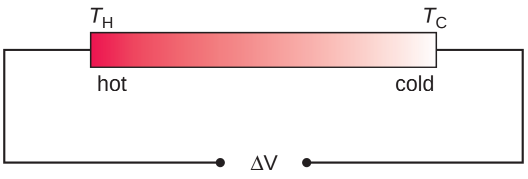


Figure 15.13 The Seebeck effect: a single phase sample with one end maintained at a high temperature T_H and the other at a low temperature T_C will develop a potential difference ΔV across its length.

15.4.2 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:

$$\sigma_e = -\left(\frac{k_B}{e}\right)\left[{}^5\!/_2 - \ln\left(\frac{N_c}{n}\right) + a_e\right]$$

and for hole transport σ_h is given by:

$$\sigma_h = \left(\frac{k_B}{e}\right)\left[{}^5\!/_2 - \ln\left(\frac{N_v}{p}\right) + a_h\right]$$

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:

$$n = N_c \exp\left(\frac{E_F - E_c}{k_B T}\right)$$

$$p = N_v \exp\left(\frac{E_v - E_F}{k_B T}\right)$$

(Section 13.2.1) and ignoring corrections we find:

$$\sigma_e = -\left(\frac{k_B}{e}\right)\left[{}^5\!/_2 - \left(\frac{E_F - E_c}{k_B T}\right)\right]$$

$$\sigma_h = \left(\frac{k_B}{e}\right)\left[{}^5\!/_2 - \left(\frac{E_v - E_F}{k_B T}\right)\right]$$

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 which 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:

$$\pi_e = \frac{E_c - E_F + \frac{3}{2}k_B T}{-e}$$

and for hole transport, π_h is given by:

$$\pi_h = \frac{E_F - E_v + \frac{3}{2}k_B T}{e}$$

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.

15.4.3 The Seebeck coefficient of solids containing point defect populations

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 complex materials such as doped cobaltites and manganites (Section 12.17.2) and cuprate superconductors (Section 13.6.5). 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.

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:

$$\alpha = \pm\left(\frac{k_\mathrm{B}}{e}\right)\ \ln\left(\frac{n_0}{n_\mathrm{d}}\right) \tag{15.4}$$

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.

This equation is formally equivalent to the *Heikes equation*:

$$\alpha = \pm\left(\frac{k}{e}\right)\ \ln\left(\frac{1-c}{c}\right)$$

where c is the fraction of defects (or mobile charge carriers) present.[1]

The usefulness of this approximate relationship can be illustrated with reference to the perovskite structure cobaltite $\mathrm{LaCoO_3}$, which is generally slightly oxygen rich when prepared in air. Any additional oxygen ions will be present as a point defect population, and in order to maintain charge neutrality some compensating 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 $\mathrm{Co^{3+}}$ ions, which have an easily variable valence. In this case each added $\mathrm{O^{2-}}$ ion will generate two $\mathrm{Co^{4+}}$ ions to maintain neutrality. Now a $\mathrm{Co^{4+}}$ ion can be regarded as $\mathrm{Co^{3+}}$ together with a hole, but this might be strongly bound to the $\mathrm{Co^{3+}}$, that is, the $\mathrm{Co^{4+}}$ state is very stable, or it could be weakly bound and jump from one $\mathrm{Co^{3+}}$ 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 $+700\ \mu\mathrm{V\,K^{-1}}$, indicating indeed that weakly bound holes are present. Using equation (15.4), the value of n_d, the number of holes (or $\mathrm{Co^{4+}}$ ions) present, is $3 \times 10^{-4}\, n_0$ (or $\mathrm{Co^{3+}}$ ions). Each $\mathrm{Co^{4+}}$ ion contributes $\tfrac{1}{2}\mathrm{O^{2-}}$ to the composition so that the material has a formula $\mathrm{LaCoO_{3.00015}}$. This population of $\mathrm{Co^{4+}}$ ions will also change the magnetic properties of the phase, which can now be estimated from the known composition.

15.4.4 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 (Figure 15.14). This is because the potential

[1] Note that the form of the equation for holes is often written:

$$\alpha = \mp\left(\frac{k}{e}\right)\ \ln\left(\frac{c}{1-c}\right)$$

which is identical to that given, as $\ln x = (-\ln(1/x)$.

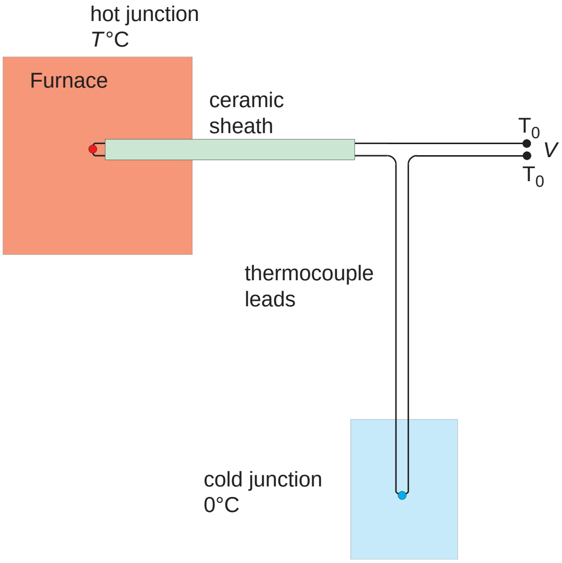


Figure 15.14 A thermocouple consists of a loop made of two different metals with one junction at 0 °C and the other at the temperature to be measured.

generated in the circuit is easily measured, and metal thermocouples capable of operating up to temperatures of almost 2000 °C are available. In practice, one junction is maintained at 0 °C. 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 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.

A series of thermocouples linked in series forms a *thermopile* (Figure 15.15). 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 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 radioactive sample.

Thermoelectric materials are also used for heating, refrigeration and for electricity generation utilising the Peltier effect. In both cases two thermoelectric materials 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 15.16a). 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 15.16b). 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.

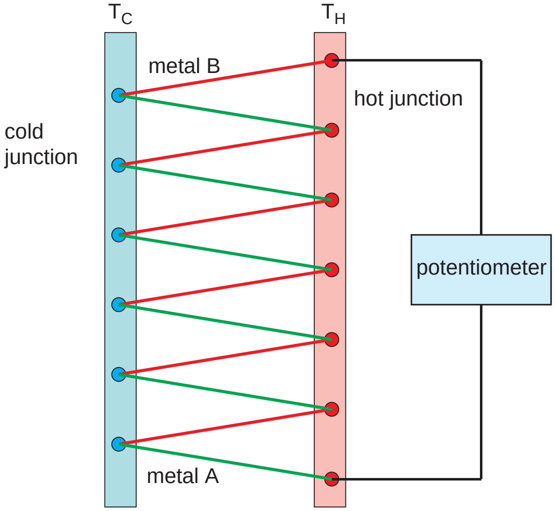


Figure 15.15 A thermopile, consisting of a number of thermocouples connected in series (schematic).

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 and 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:

$$ZT = \frac{T\,\Sigma^2}{\kappa\,\rho}$$

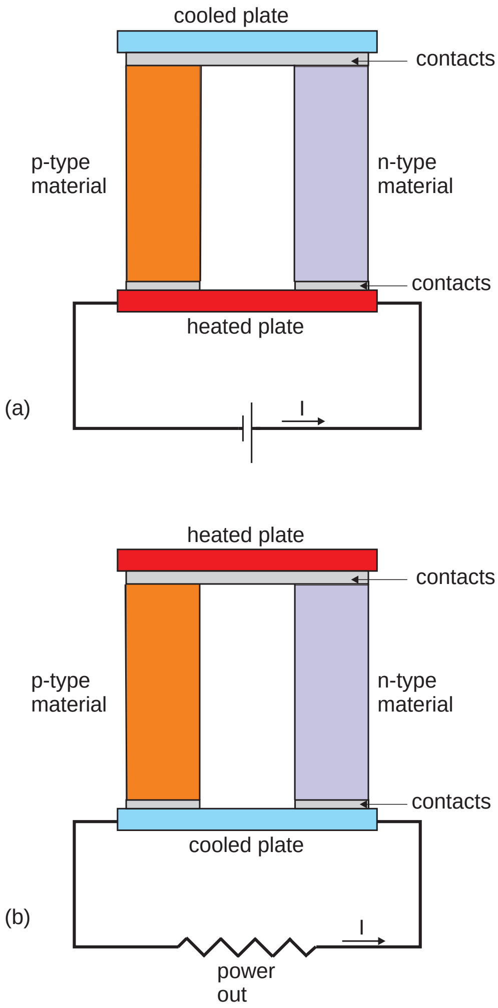


Figure 15.16 (a) Use of the Peltier effect for thermoelectric heating or cooling: a current passed through a circuit containing n-type and p-type thermoelectric materials will cause one plate to become warmer and one to become cooler. (b) Use of the Peltier effect for thermoelectric power generation: current will flow in a circuit containing a heated plate and a cooled plate connected by n-type and p-type thermoelectric material.

where Σ is the relative Seebeck coefficient of the thermoelectric elements, κ is the thermal conductivity and ρ the electrical resistivity. 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}.

15.5 The magnetocaloric effect

15.5.1 The magnetocaloric effect and adiabatic cooling

Magnetic and thermal properties are linked via the *magnetocaloric effect* (MCE), 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 15.17a).

(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 $\mathbf{H}$ (Figure 15.17b).

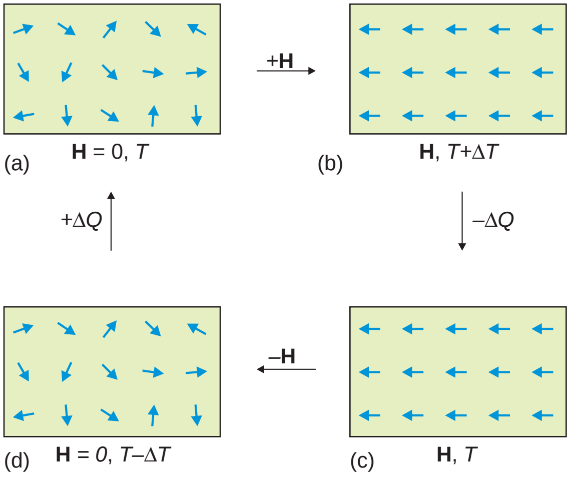


Figure 15.17 Adiabatic cooling: a paramagnetic working material (a) is placed in a magnetic field in adiabatic conditions leading to a temperature increase ΔT (b). A small quantity of heat ΔQ is then removed while the magnetic field remains in place to drop the temperature to its initial value (c); the magnetic field is then reduced adiabatically, producing a temperature drop ΔT (d). Addition of a small amount of heat ΔQ from the material to be cooled regenerates the initial state (a).

(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 $\mathbf{H}$ (Figure 15.17c).

(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 $T - \Delta T$ (Figure 15.17d).

(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 (Figure 15.17a).

(vi) The whole cycle is now repeated until the material to be cooled reaches the desired temperature.

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.

15.5.2 The giant magnetocaloric effect

Recently, the discovery of so-called *giant magnetocaloric effects* (GMCE) 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 most recent interest has centred on a group of *Heusler alloys* containing Ni and Mn which are also shape-memory materials (Section 8.5.4). 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 (Section 12.7.2).

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. The imposition of a magnetic field results in a conventional magnetocaloric effect that generates a small temperature increase from T to $T+\Delta T_{mag}$ under adiabatic conditions (Figure 15.18a,b). Concurrently the field also forces a crystallographic transition to take place from the martensite phase to a ferromagnetic high-symmetry austenite phase. In the alloys mentioned the austenite

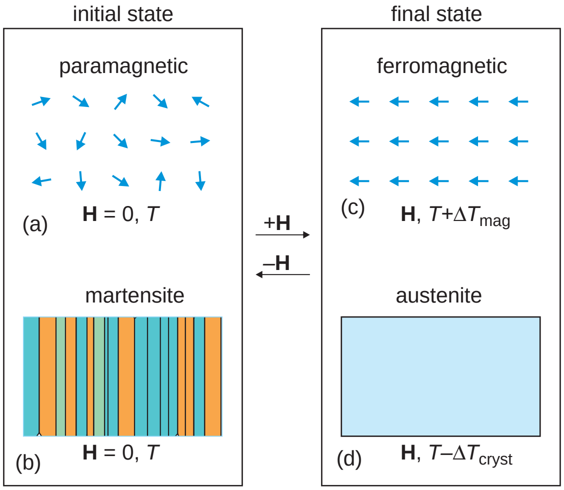


Figure 15.18 Giant magnetocaloric effect: the initial state (a, b) of the working material is paramagnetic multiply-twinned martensite; application of a magnetic field produces the final state (c, d) consisting of ferromagnetic austenite and a net temperature drop ΔT. Removal of the field regenerates the initial state.

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 $-\Delta T_{\text{cryst}}$ (Figure 15.18b,c). The net change is given by the sum of these two thermal effects:

$$\Delta T(\text{net}) = +\Delta T_{\text{mag}} + (-\Delta T_{\text{cryst}})$$

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. The martensitic start temperature in these alloys is close to 319 K and the martensite finish temperature close to 311 K. Heat exchange allows the initial martensitic phase at the initial temperature T to be regenerated. The cycle can then be repeated, as described above.

At the moment (2012) prototypes of magnetic refrigerators have been produced, but no commercial models are yet on the market as a number of technical difficulties have to be overcome, one of which is that the temperature at which the martensitic transformation takes place decreases in strong magnetic fields. New materials to use as the working substance are continually being developed to surmount this and other difficulties.

Further reading

General:

Hummel, R.E. (2001) *Electronic Properties of Materials*, 3rd edn. Springer-Verlag, New York.

Rowe, D.M. (ed.) (1995) *CRC Handbook of Thermoelectrics*. CRC Press, Boca Raton.

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.

Losego, M.D., *et al.* (2012) Effects of chemical bonding on heat transport across interfaces. *Nature Materials*, **11**: 502–6, and references therein.

Negative and zero thermal expansion:

Agrawal, D.K. (1994) [NPZ]: a new family of real materials for low thermal expansion applications. *J. Mater. Educ.*, **16**: 139–65.

Adak, S., *et al.* (2011) Thermal expansion in 3d-metal Prussian Blue analogues – a survey study. *J. Solid State Chem.*, **184**: 2854–61.

Sleight, A.W. (1998) Compounds that contract on heating. *Inorg. Chem.*, **37**: 2854–60.

Song, X., *et al.* (2011) Adjustable zero thermal expansion in antiperovskite manganese nitride. *Adv. Mater.*, **23**: 690–94.

The magnetocaloric effect in alloys:

Liu, J., *et al.* (2012) Giant magnetocaloric effect driven by structural transitions. *Nature Materials*, **11**: 620–26, and references therein.

Problems and exercises

Quick quiz

1 The heat capacity 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 equal to the heat capacity at constant pressure, C_p.

2 The low-temperature heat capacity of a solid is proportional to:
(a) T^{-4}.
(b) T^{-3}.
(c) T^{-2}.

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}.

4 The thermal conductivity of a solid is mainly due to:
(a) Phonons.
(b) Defects.
(c) Electrons.

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 same thermal conductivity as the parent metals.

6 Compared with a poorly crystalline polymer, a highly crystalline polymer has a:
(a) Lower thermal conductivity.
(b) Higher thermal conductivity.
(c) About the same thermal conductivity.

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) $2.63 \times 10^{-5}\,K^{-1}$.

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.

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.

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.

11 A thermocouple makes use of:
(a) The Seebeck effect.
(b) The Peltier effect.
(c) The Thomson effect.

12 Thermoelectric refrigerators utilise
(a) The Thomson effect.
(b) The Seebeck effect.
(c) The Peltier effect.

Calculations and questions

15.1 How much energy is needed to raise the temperature of 2.5 moles 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?

15.2 How much energy has to be extracted to lower the temperature of 15 g tungsten metal from 2500 K 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.

15.3 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.

15.4 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.

15.5 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 °C 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}?

15.6 Show that the equation for the heat transfer across a number of slabs of material with the same surface area, A, is:

$$dQ/dt = A\Delta T / \sum_{i=1}^{n} \Delta x_i/\kappa_i$$

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.

15.7 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? How does this compare with a pan of the same dimensions with (b) a solid copper bottom 4 mm thick, and (c) a solid stainless steel bottom, 4 mm thick? The thermal conductivity of copper is 403 W m^{-1} K^{-1} and that of stainless steel is 18 W m^{-1} K^{-1}.

15.8 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? The thermal conductivity of the glass is 0.96 W m^{-1} K^{-1}, and that of air is 2.41×10^{-2} W m^{-1} K^{-1}.

15.9 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}.

15.10 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}.

15.11 (a) Estimate the mean coefficient of linear expansion for the **a**- and **c**-axes of $CaZr_4P_6O_{24}$, using the data in Figure 15.8. (b) What is the mean coefficient of volume expansion of this material?

15.12 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 μV K^{-1} over the whole of this temperature range, estimate the average value of the absolute Seebeck coefficient for Au.

15.13 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 μV K^{-1} over the whole of this temperature range, estimate the average value of the absolute Seebeck coefficient for Al.

15.14 The absolute Seebeck coefficient for lead, σ_{Pb}, at 300 K is −1.047 μK^{-1}, and for platinum at 300 K is $\sigma_{Pt} = -5.05$ μV 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.

PART 5

Nuclear properties of solids

16

Radioactivity and nuclear reactions

- What is radioactivity?
- How does carbon dating work?
- What produces energy in a nuclear power station?

16.1 Radioactivity

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 that 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.

16.1.1 Naturally occurring radioactive elements

At the turn of the 20th century only the two heaviest naturally occurring elements known at that epoch, thorium and uranium, were known to be radioactive. The earliest studies indicated that the 'penetrating radiation' characteristic of uranium and thorium compounds 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 lead to the characterisation of two new elements, which were named polonium and radium. Both are far more radioactive than uranium and decay so rapidly that no ore

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

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, a radioactive gas associated with granite, result from the radioactive decay of natural uranium ores.

16.1.2 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 ${}^{A}_{Z}X$. For example, the radioactive carbon isotope used in radiocarbon dating has a symbol ${}^{14}_{6}C$. 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 ${}^{14}_{6}C$ 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 ${}^{9}_{4}Be$ 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 16.1).

Table 16.1 Nuclides and particles in nuclear reactions

Name	Nuclide	Shorthand
Proton	${}^{1}_{1}H$	p
Deuteron	${}^{2}_{1}H$	d
Hydrogen-3 (triton)	${}^{3}_{1}H$	t
Alpha particle	${}^{4}_{2}He$	α
Neutron	${}^{1}_{0}n$	n
Electron (beta particle)	${}^{0}_{-1}e$	β, β^-, e, e^-
Positron	${}^{0}_{+1}e$	β^+, e^+
Gamma ray	–	γ
Electron neutrino	–	ν_e
Electron antineutrino		$\bar{\nu}_e$

16.1.3 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 alpha particle is correctly written ${}^{4}_{2}He$ not ${}^{4}_{2}He^{2+}$, and a uranium nucleus is written ${}^{238}_{92}U$ not ${}^{238}_{92}U^{92+}$.

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:

$${}^{238}_{92}U \rightarrow {}^{234}_{90}Th + {}^{4}_{2}He$$

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:

$$^{234}_{90}\mathrm{Th} \rightarrow {}^{234}_{91}\mathrm{Pa} + {}^{0}_{-1}\mathrm{e} + \bar{\nu}_e$$

or $$^{234}\mathrm{Th} \rightarrow {}^{234}\mathrm{Pa} + \mathrm{e}^- + \bar{\nu}_e$$

or $$^{234}\mathrm{Th} \rightarrow {}^{234}\mathrm{Pa} + \beta^- + \bar{\nu}_e$$

In this decay, radioactive thorium-234 emits an electron (beta 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 below. In addition an electron antineutrino, needed to conserve energy and spin, is also emitted during normal beta 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 4 less than the parent and a proton number of 2 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 1 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:

   $$^{1}_{0}\mathrm{n} \rightarrow {}^{1}_{1}\mathrm{H} + {}^{0}_{-1}\mathrm{e}$$

   or $$\mathrm{n} \rightarrow \mathrm{p} + \mathrm{e}^-$$

   As in all beta 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 $$^{238}\mathrm{U}(\mathrm{d}, 2\mathrm{n})^{238}\mathrm{Np}$$

16.1.4 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 forms. The complete form that these series take is rather complex. Here we present only the main reaction chains, involving α-decay, in which an alpha particle is ejected from the nucleus, or β-decay, in which a beta particle is ejected from the nucleus.

16.1.4.1 The uranium series

The parent nuclide is the naturally occurring isotope uranium-238 and the series ends with the stable nuclide lead-206 (Figure 16.1). The parent nuclide is shown at the top right of the figure. An α-decay is represented by a diagonal displacement of 2 proton number units to the left. A β-decay is shown as a

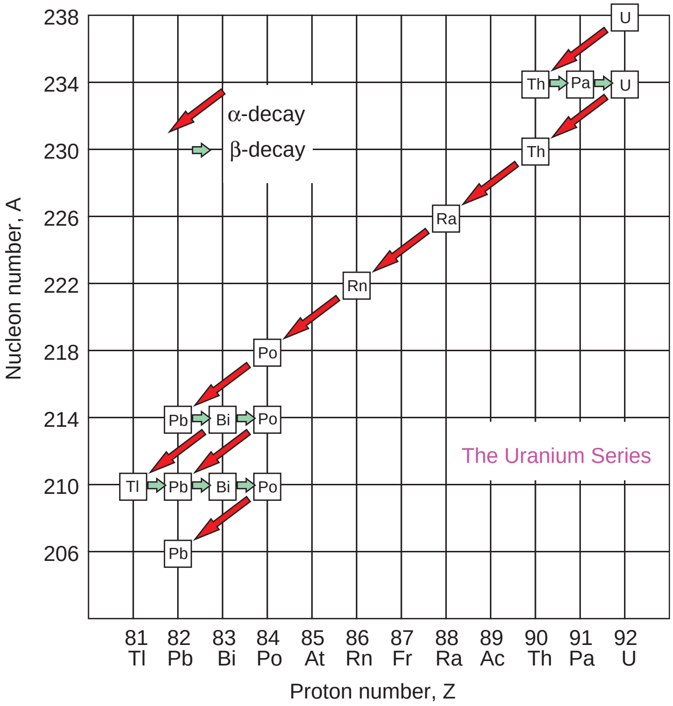


Figure 16.1 The uranium-238 decay series.

horizontal displacement of 1 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.

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:

$$^{226}_{88}\mathrm{Ra} \rightarrow {}^{222}_{86}\mathrm{Rn} + {}^{4}_{2}\mathrm{He}$$

The radon isotope formed as part of the uranium series is an α-emitter that decays according to the following reaction:

$$^{222}_{86}\mathrm{Rn} \longrightarrow {}^{218}_{84}\mathrm{Po} + {}^{4}_{2}\mathrm{He}$$

Different isotopes of radon form in the radioactive series described below.

16.1.4.2 The thorium series

This series starts with naturally occurring thorium-232, and ends with the stable isotope lead-208 (Figure 16.2). The daughter elements all have mass numbers divisible by four, and the series formula is $4x+0$. The first reaction in the series is:

$$^{232}_{90}\mathrm{Th} \longrightarrow {}^{228}_{88}\mathrm{Ra} + {}^{4}_{2}\mathrm{He}$$

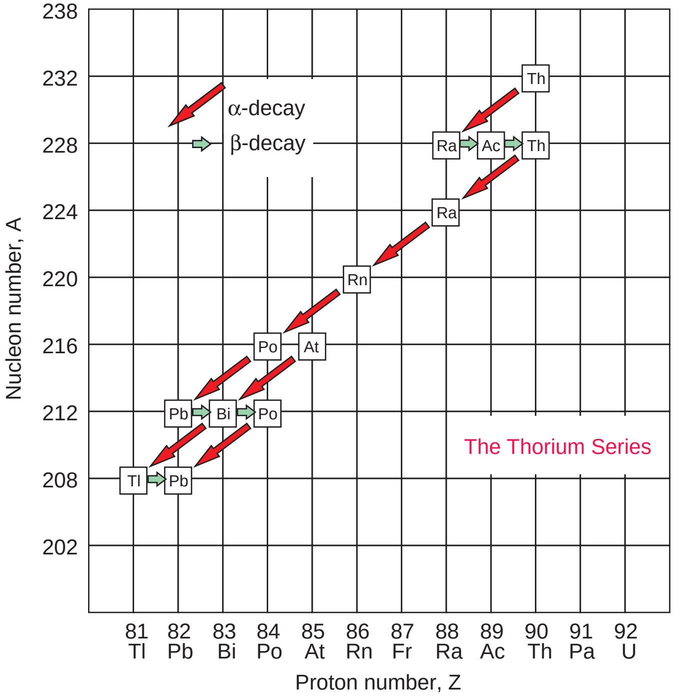


Figure 16.2 The thorium-232 decay series.

16.1.4.3 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 16.3). 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 $4x + 3$.

16.1.4.4 The neptunium series

The existence of the three naturally occurring series characterised by the formulae $4x$, $4x + 2$ and $4x + 3$, led 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 16.4). 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.

16.1.5 Nuclear stability

Although many radioactive isotopes have been discovered or prepared, most known nuclei are stable. Although it is not possible to explain nuclear stability theoretically, empirical observations allow one to guess the likely stability of any particular nuclide. A useful guide is a graph of the proton number, Z, versus the nucleon number A, or the number of neutrons present, $Z - A$ (Figure 16.5). Stable nuclei all cluster into a narrow strip called the *band of*

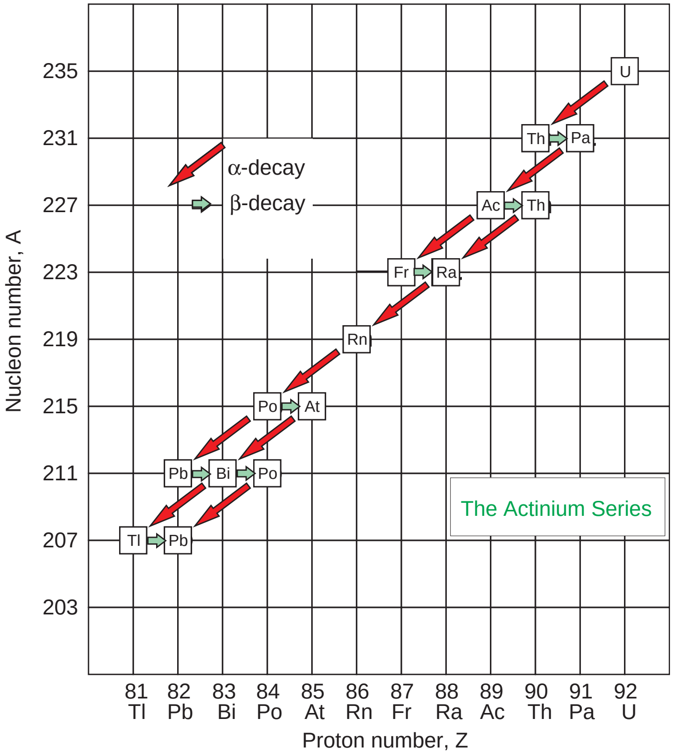


Figure 16.3 The actinium-227 (uranium-235) decay series.

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.

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. $^{2}_{1}H$, $^{6}_{3}Li$, $^{10}_{5}B$, $^{14}_{7}N$. 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 $^{4}_{2}He$, $^{16}_{8}O$, $^{40}_{20}Ca$ and $^{208}_{82}Pb$.

16.2 Artificial radioactive atoms

16.2.1 Transuranic 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. All these artificial heavy elements are radioactive, and many have very short lifetimes.

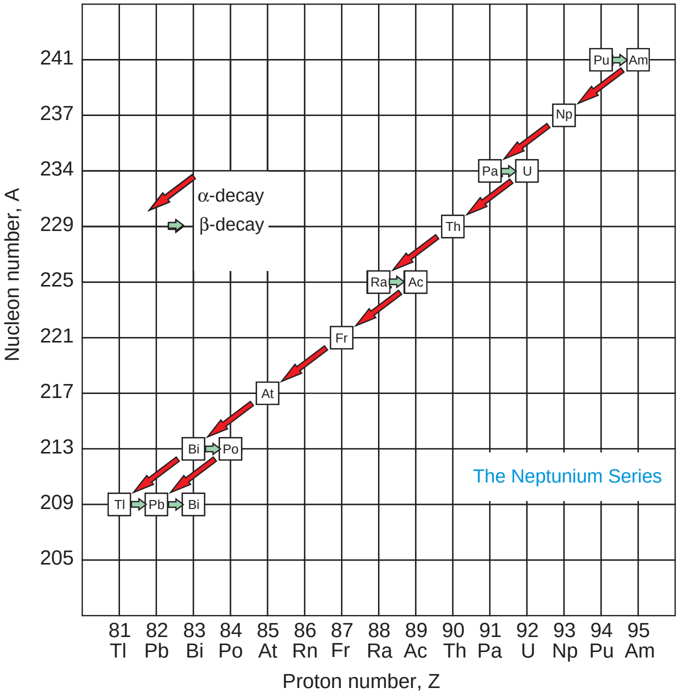


Figure 16.4 The plutonium-241 decay series.

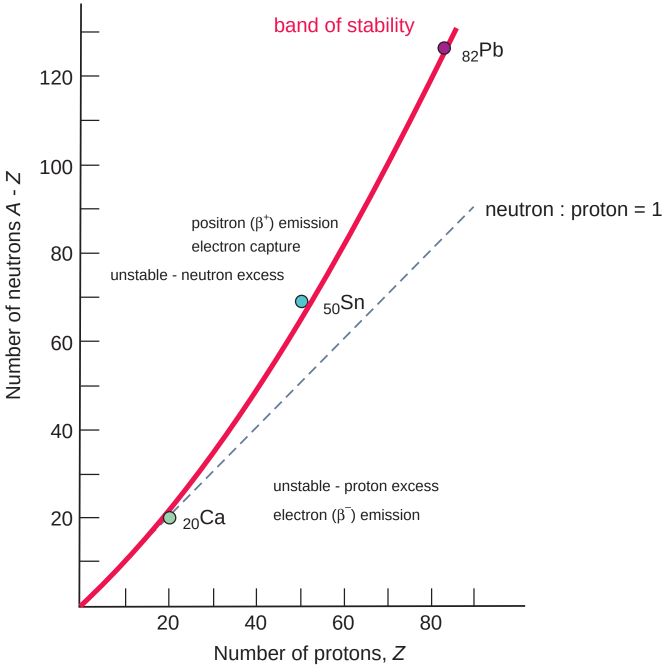


Figure 16.5 The band of stability of the elements: nuclides close to the band are stable; those above it tend to decay via positron (β^+) emission or electron capture; those below it tend to decay via electron (β^-) emission.

The *actinoids* are a series in which the 5f electron orbitals become occupied in the ground state of the elements (Table 16.2) 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. Transition metals in which the 6d orbitals are filled continue beyond the actinoids (Table 16.2). Copernicium is the last element to have been named, but other so far unnamed new elements with proton numbers up to 118 have been reported.

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

Table 16.2 The actinoid and transactinoid elements

Name	Symbol	Proton	Nucleon	Electron configuration**	Half-life***
		number Z	number A*		
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] $5f^2\ 6d^1\ 7s^2$ or [Rn] $5f^1\ 6d^2\ 7s^2$	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] $5f^4\ 6d^1\ 7s^2$ or [Rn] $5f^5\ 7s^2$	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
Nobelium	No	102	259	[Rn] $5f^{14}\ 7s^2$	58 min
Lawrencium	Lr	103	262	[Rn] $5f^{14}\ 6d^1\ 7s^2$ or [Rn] $5f^{14}\ 7s^2\ 7p^1$	3 min
Transition metals					
Rutherfordium	Rf	104	263	[Rn] $5f^{14}\ 6d^2\ 7s^2$	10 m
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] $5f^{14}\ 6d^8\ 7s^2$	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

*Most stable isotope.

**The electron configuration of these elements is uncertain and the configurations listed should be regarded as likely rather than definite.

***Of most stable isotope.

[Rn] represents the configuration of the noble gas radon.

their physical and chemical properties to be investigated experimentally.

16.2.2 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), 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:

$$^{13}_{7}\mathrm{N} \rightarrow {}^{13}_{6}\mathrm{C} + {}^{0}_{+1}\mathrm{e} + v_e$$

Release of a positron is called β^+*-decay*, and to conserve energy and spin, the positron is accompanied by an electron neutrino, v_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} sec, 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 neutron 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. An example is:

$$^{16}_{7}\mathrm{N} \rightarrow {}^{16}_{8}\mathrm{O} + {}^{0}_{-1}\mathrm{e} + \gamma$$

The production of γ-rays often takes place after an α- or β-emission because these latter processes frequently leave the daughter nucleus, $^{16}\mathrm{O}$ in this case, in an excited state. Stability is gained when the nucleus subsequently loses energy by way of γ-emission.

16.3 Nuclear decay

16.3.1 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 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 half of the sample to decay and another 29 years for half of the remaining strontium-90 to decay and so on.

Time, years	0	29	58	87
Amount of Sr-90, grams	1	0.5	0.25	0.125

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 16.3).

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 gram of that material. Radium has a specific activity $3.7 \times 10^{10}\,\mathrm{Bq\,g^{-1}}$.

The disintegration of atomic nuclei can be expressed by the *first-order* rate equation:

$$-\frac{dN}{dt} = \lambda N \qquad (16.1)$$

where dN/dt 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 exists at time $t = 0$ to the amount present at any later time, N, equation (16.1) must be integrated, to give:

Table 16.3 Half-lives of some radioisotopes

Nuclide	Half-life	Decay mode
$^{3}_{1}\mathrm{H}$	12.33 yr	β^-
$^{14}_{6}\mathrm{C}$	5730 yr	β^-
$^{22}_{11}\mathrm{Na}$	2.602 yr	β^+, γ
$^{47}_{20}\mathrm{Ca}$	4.536 d	β^-, γ
$^{59}_{26}\mathrm{Fe}$	44.496 d	β^-, γ
$^{60}_{27}\mathrm{Co}$	5.271 yr	β^-, γ
$^{90}_{38}\mathrm{Sr}$	28.5 yr	β^-
$^{131}_{53}\mathrm{I}$	8.040 d	β^-, γ
$^{133}_{54}\mathrm{Xe}$	5.245 d	β^-, γ
$^{137}_{55}\mathrm{Cs}$	30.1 yr	β^-, γ
$^{198}_{79}\mathrm{Au}$	2.6395 d	β^-, γ
$^{222}_{86}\mathrm{Rn}$	2.825 d	α, γ
$^{226}_{88}\mathrm{Ra}$	1600 yr	α, γ, X^*
$^{228}_{88}\mathrm{Ra}$	5.75 yr	β^-, γ
$^{235}_{92}\mathrm{U}$	7.037×10^8 yr	α, γ, X^*
$^{238}_{92}\mathrm{U}$	4.468×10^9 yr	α, γ

*X-ray emission.

$$N = N_0 e^{-\lambda t} \quad \text{or} \quad N_0 = N e^{\lambda t} \qquad (16.2)$$

(Figure 16.6). Equation (16.2) is often written in the logarithmic form:

$$\ln N - \ln N_0 = -\lambda t$$

$$\text{or} \quad \ln\left(\frac{N_0}{N}\right) = \lambda t \qquad (16.3)$$

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 $N = {}^1\!/_2 N_0$ into equation (16.3) shows that the half-life is given by:

$$t_{1/2} = \frac{0.693}{\lambda} \qquad (16.4)$$

The fraction of the radio-isotope remaining after n half-lives have elapsed is:

$$\frac{N}{N_0} = \left({}^1\!/_2\right)^n$$

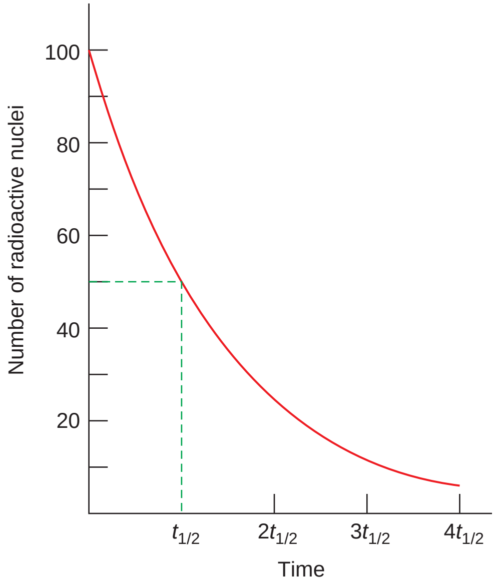


Figure 16.6 The exponential rate of radioactive decay.

The specific activity of a material can be related to its half-life in the following way.

$$\text{activity} = -\frac{dN}{dt} = \lambda N$$

The specific activity is the activity divided by the mass in grams, m:

$$\text{specific activity} = \frac{\text{activity}}{m} = \frac{\lambda N}{m}$$

Substituting for λ from equation (16.4) gives:

$$\text{specific activity} = \frac{0.693\,N}{t_{1/2}\,m}$$

The number of atoms, N, remaining at any time, and their mass, m, are related in the following way. The number of moles of material present is given by:

$$\text{moles of material} = \frac{N}{N_{\text{A}}}$$

where N_{A} is the Avogadro constant. The mass of this quantity is:

$$m = M\left(\frac{N}{N_{\text{A}}}\right)$$

where M is the isotopic mass of the isotope. Hence:

$$\frac{N}{m} = \frac{N_{\text{A}}}{M}$$

and:

$$\text{specific activity} = \frac{0.693 N_{\text{A}}}{t_{1/2} M} = \frac{4.2 \times 10^{23}}{t_{1/2} M}$$

16.3.2 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 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 is measured and the age of the sample is then obtained from equations (16.2) and (16.3), 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:

$$N_r = N_0 - N$$

and equations (16.2) and (16.3) become:

$$N_r = N(e^{\lambda t} - 1)$$

$$\ln\left(\frac{N + N_{\text{r}}}{N}\right) = \lambda t$$

that is:

$$^{87}\mathrm{Sr} = {}^{87}\mathrm{Rb}\,(e^{\lambda t} - 1)$$

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, another 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:

$$\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}} = \frac{^{87}\mathrm{Rb}}{^{86}\mathrm{Sr}}(e^{\lambda t} - 1)$$

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:

$$\left(\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}}\right) - \left(\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}}\right)_0 = \frac{^{87}\mathrm{Rb}}{^{86}\mathrm{Sr}}(e^{\lambda t} - 1)$$

or

$$\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}} = \frac{^{87}\mathrm{Rb}}{^{86}\mathrm{Sr}}(e^{\lambda t} - 1) + \left(\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}}\right)_0$$

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) versus (^{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 16.7). The age of the material can then be determined from the slope of the graph.

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 16.1, 16.3).

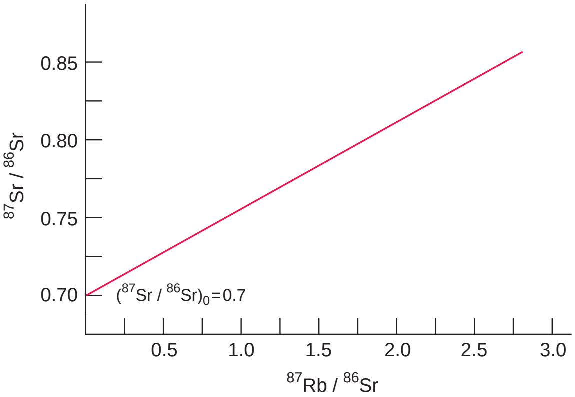


Figure 16.7 A schematic plot of isotope ratios ^{87}Sr/^{86}Sr versus ^{87}Rb/^{86}Sr. The slope of the line gives a value for the age of the material.

$$^{238}\mathrm{U} \rightarrow {}^{206}\mathrm{Pb},\ \text{measuring}\ {}^{206}\mathrm{Pb}/{}^{204}\mathrm{Pb}$$

$$^{235}\mathrm{U} \rightarrow {}^{207}\mathrm{Pb},\ \text{measuring}\ {}^{207}\mathrm{Pb}/{}^{204}\mathrm{Pb}$$

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, $\mathrm{ZrSiO_4}$, 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 above.

For $^{238}\mathrm{U} \rightarrow {}^{206}\mathrm{Pb}$:

$$\left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right) - \left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0 = \frac{^{238}\mathrm{U}}{^{204}\mathrm{Pb}}\,(e^{\lambda 1t} - 1)$$

For $^{235}\mathrm{U} \rightarrow {}^{207}\mathrm{Pb}$:

$$\left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right) - \left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0 = \frac{^{235}\mathrm{U}}{^{204}\mathrm{Pb}}\,(e^{\lambda 2t} - 1)$$

where $\lambda 1$ and $\lambda 2$ are the relevant decay constants for the two decomposition chains. Combining these two equations:

$$\frac{\left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)-\left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0}{\left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)-\left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0}=\frac{^{235}\mathrm{U}(e^{\lambda 2t}-1)}{^{238}\mathrm{U}(e^{\lambda 1t}-1)}$$

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:

$$^{40}_{19}\mathrm{K} \rightarrow {}^{40}_{18}\mathrm{Ar} + {}^{0}_{-1}\mathrm{e}\ ; \quad t_{1/2} = 1.3 \times 10^9 \text{ years}$$

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-bearing 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:

$$^{14}_{7}\mathrm{N} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{14}_{6}\mathrm{C} + {}^{1}_{1}\mathrm{H}$$

Carbon-14 subsequently degrades by β-decay:

$$^{14}_{6}\mathrm{C} \rightarrow {}^{14}_{7}\mathrm{N} + {}^{0}_{-1}\mathrm{e}; \quad t_{1/2} = 5730 \text{ years}$$

The 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 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 can then be used to determine the age of the sample itself using equation (16.2). The data obtained in this way is found to be wrong by 10–20% due to naturally occurring fluctuations in the amount of carbon-14 present in the atmosphere. A calibration curve, based upon materials that can be accurately dated by independent means, is available to allow the necessary corrections to be made if they are deemed to be significant.

16.4 Nuclear energy

16.4.1 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 pseudochemical reaction:

$$\mathrm{Z\,protons} + (\mathrm{A}-\mathrm{Z})\,\mathrm{neutrons} \rightarrow {}^{A}_{Z}\mathrm{X}$$

the difference between the mass of the products and the mass of the reactants Δm is given by:

$$\varDelta m = \mathrm{mass}\,[^{A}_{Z}\mathrm{X}] - \mathrm{mass}[\mathrm{Z\,protons}+(\mathrm{A}-\mathrm{Z})\,\mathrm{neutrons}]$$

This difference mass is converted into energy using the Einstein equation:

$$E = \Delta mc^2$$

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 one twelfth of the mass of the isotope carbon-12. As the mass of one atom of carbon-12 is 1.9926×10^{-26} kg,

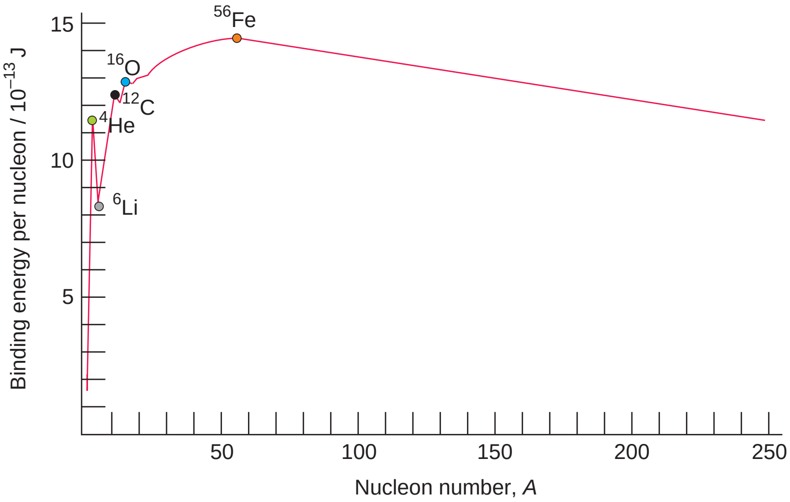


Figure 16.8 The isotopic binding energy per nucleon. The most stable isotope is ^{56}Fe.

$1u = 1.6605 \times 10^{-27}$ kg. Moreover, it is usual to use the mass of the hydrogen isotope $^{1}_{1}$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 $^{1}_{1}$H isotopes cancels neatly with the mass of the electrons on the $^{A}_{Z}$X isotope.

The isotopic binding energy is often quoted per nucleon (Figure 16.8). 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, and decreases slightly thereafter. Isotopes of lower nucleon number than ^{56}Fe will release energy in fusion reactions (Section 16.4.6). Isotopes of higher nucleon number than ^{56}Fe will release energy upon fission (Section 16.4.2).

16.4.2 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.

$$^{238}_{92}\text{U} + ^{1}_{0}\text{n} \rightarrow ^{239}_{93}\text{Np} + ^{0}_{-1}\text{e}$$

Neptunium has a half-life of about 2 days and decays to plutonium-239 by way of β-decay:

$$^{239}_{93}\text{Np} \rightarrow ^{239}_{94}\text{Pu} + ^{0}_{-1}\text{e}$$

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:

$$^{235}_{92}\mathrm{U} + ^{1}_{0}\mathrm{n} \rightarrow ^{92}_{36}\mathrm{Kr} + ^{141}_{56}\mathrm{Ba} + 3^{1}_{0}\mathrm{n} \qquad (16.5)$$

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 equation (16.5) 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, one gram of coal burnt in air produces about 3×10^{4} J. That is, uranium-235 fission produces about 1 million times more energy than burning fossil fuels.

The second important feature of equation (16.5) 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 16.9). 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.

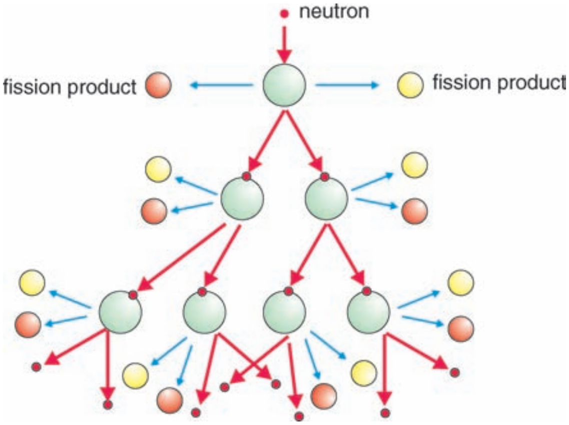


Figure 16.9 Schematic illustration of a fission chain reaction. In the chain shown, each nucleus emits two neutrons that cause further fission.

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, so the arrangement of the uranium-235 must incorporate a *moderator*, which is a material for slowing down the energetic neutrons released by the fission.

16.4.3 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 16.10). 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

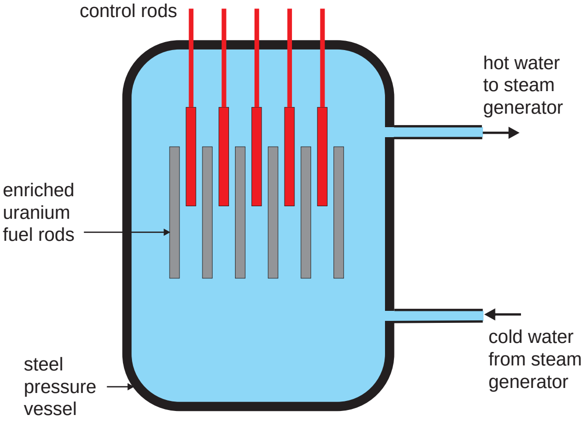

control rods
hot water to steam generator
enriched uranium fuel rods
cold water from steam generator
steel pressure vessel

Figure 16.10 Layout of a pressurised water nuclear reactor for power generation, schematic. The moderator is pressurised water, which also acts as a coolant, transferring heat to steam generators.

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.

Uranium dioxide has a number of properties that make it suitable for a fuel. The crystal structure, the fluorite type, is similar to that of calcia-stabilised zirconia and 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 Zircalloy 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.)

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. (See also Section 16.4.5.)

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.

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 whist being able to withstand intense neutron bombardment. Clearly the search for new safe materials cannot always be made experimentally, and in order to evaluate material properties in extreme environments, much use is made of computer simulations and thermodynamic modelling tools such as CALPHAD (Section 4.5).

16.4.4 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 15.4.4). 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 on 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 watts per gram. 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 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.

16.4.5 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:

$$^{238}_{92}\mathrm{U} + ^{1}_{0}\mathrm{n} \rightarrow ^{239}_{92}\mathrm{U} \quad (t_{1/2} = 24\ \text{minutes})$$

$$^{239}_{92}\mathrm{Th} \rightarrow ^{239}_{93}\mathrm{Np} + ^{0}_{-1}\mathrm{e} \quad (t_{1/2} = 2.35\ \text{days})$$

$$^{239}_{93}\mathrm{Np} \rightarrow ^{239}_{94}\mathrm{Pu} + ^{0}_{-1}\mathrm{e} \quad (t_{1/2} = 24{,}000\ \text{years})$$

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 2 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.

16.4.6 Fusion

The binding energy curve (Figure 16.8) 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 limitless clean power will continue to be a goal for research.

Uncontrolled fusion reactions form the basis of the hydrogen bomb. The temperatures needed to initiate fusion are brought about by a fission bomb. A number of competing and cooperative reactions take place during fusion. Typical reaction schemes for fusion reactions are:

$$^{2}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} \rightarrow {}^{3}_{2}\mathrm{H} + {}^{1}_{0}\mathrm{n}$$

$$^{2}_{1}\mathrm{H} + {}^{3}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}$$

These involve the hydrogen isotopes deuterium, ^{2}H, and tritium, ^{3}H, as reactants. Tritium has a low natural abundance, and is generated in the reaction by use of the isotope lithium-6.

$$^{6}_{3}\mathrm{Li} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{3}_{1}\mathrm{H} + {}^{4}_{2}\mathrm{He}$$

The reacting material is lithium-6 deuteride, ^{6}Li^{2}H, which forms the core inside a fission bomb. The energy released by such reactions can be calculated by the methods described above.

16.4.7 Solar cycles

The process that powers the Sun and all 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:

$$4\,^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He} + 2\mathrm{e}$$

The energy production is enormous, about $10^{26}\,\mathrm{J\,s^{-1}}$.

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 (1 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 raises 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 16.8) indicates that at this point 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.

16.5 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 severe problems, though, if radioisotopes were short-lived, or the decay process could be accelerated. Unfortunately, some of the more important byproducts 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.

16.5.1 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 Zircalloy 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 (but reaction with other materials including reaction products can lower melting points considerably, see Chapter 4).

At Fukushima Daiichi the reactors shut down automatically following the earthquake, and emergency power supplies took over the control electronics and cooling water-pumping system. Unfortunately the 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 Zircalloy and water subsequently exploded and damaged buildings, with the result that multiple fires broke out. Full meltdown of three reactors occurred resulting in a 'lava flow' of molten fuel assemblies, combined with concrete, steel, sand and so on. One product of this reaction mixture was zircon, $U_xZr_{1\text{-}x}SiO_4$, the same crystals that are used in uranium dating (Section 16.3.2).

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 easily 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 caesium-134 and caesium-137 pose a particular problem. The half-life of caesium-134 is 2.06 yr, decaying via β-emission. The half-life of caesium-137 is 30.17 yr, 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, and so on, by way of grazing animals, and so enter the human food chain.

The only practical solution in cases of this nature is to isolate the area as far as possible, and to prevent contaminated products from reaching market.

16.5.2 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 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 ten years or so, to cool down. During this period, many of the radioactive elements decay, as most have short half-lives, for example:

$$^{142}_{56}\mathrm{Ba} \rightarrow {}^{142}_{57}\mathrm{La} + \beta^- \quad t_{1/2} = 11 \text{ mins}$$

After ten 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 a 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 material.

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 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 is an ongoing area of intensive materials research.

Further reading

The history of the birth and development of atomic physics and radiochemistry, from the original discovery of radioactivity to the production of the hydrogen bomb, makes fascinating reading. A unique insight to this epoch can be gained by reading the lectures given by the key Nobel prizewinners, in book form and also at www.nobel.se.

Also of general interest are:

Chown, M. (1999) *The Magic Furnace*. Jonathan Cape, London.

Corfield, R. (2001) *Architects of Eternity*. Hodder Headline, London.

Lewis, C. (2000) *The Dating Game*. Cambridge University Press, Cambridge.

McQuarrie, D.A. and Rock, P.A. (1991) *General Chemistry*, 3rd edn, Chapter 24. W.H. Freeman & Co.

The search for new heavy elements:

Clery, D. (2011) Which way to the island? *Science*, **333**: 1377–9.

Radioactive dating:

Hiess, J., *et al.* (2012) ^{238}U/^{235}U systematics in terrestrial uranium-bearing minerals. *Science*, **335**: 1610–14.

Pickering, R., *et al.* (2011) *Australopithecus sediba* at 1.977 Ma and implications for the origins of the genus *Homo*. *Science*, **333**: 1421; especially the Supplementary information. (Uranium dating.).

Pike, A.W.G., *et al.* (2012) U-series dating of Paleolithic art in 11 caves in Spain. *Science*, **336**: 1409–13.

Stirling, C.H. (2012) Keeping time with Earth's heaviest element. *Science*, **335**: *1585–6*. (Uranium dating.).

Wu, X., *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*. Academic Press, Amsterdam.

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**: 1184–8.

Yoshida, N. and Kandra, J. (2012) Tracking the Fukushima radionuclides. *Science*, **336**: 1115–16.

Problems and exercises

Quick quiz

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.

2 The number of electrons surrounding a neutral atom is the same as the:
(a) Mass number.
(b) Proton number.
(c) Nucleon number.

3 Isotopes of an element have the same numbers of:
(a) Neutrons.
(b) Nucleons.
(c) Protons.

4 An atomic species that has specified nucleon number and proton number is named:
(a) An isotope.
(b) A nuclide.
(c) A radioisotope.

5 When a radioactive element emits an α-particle the daughter element has a nucleon number:
(a) 4 less than the parent atom.
(b) 2 less than the parent atom.
(c) The same as the parent atom.

6 When a radioactive element emits a β-particle the daughter element has a nucleon number:
(a) 4 less than the parent atom.
(b) 2 less than the parent atom.
(c) The same as the parent atom.

7 The total number of different gaseous elements produced in all the different radioactive series is:
(a) One.

(b) Two.

(c) Four.

8 The only non-naturally occurring radioactive decay series, the neptunium series, has a formula:
(a) $4x + 1$.
(b) $4x + 2$.
(c) $4x + 3$.

9 The elements with a partly filled 5f electron shell are named the:
(a) Lanthanoids.
(b) Actinoids.
(c) Uranoids.

10 During β^{+}-decay, the particle ejected from the nucleus is:
(a) A neutron.
(b) An electron.
(c) A positron.

11 Radioactive nuclei below the band of stability tend to emit:
(a) Positrons.
(b) Electrons.
(c) Neutrons.

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 0.00125 g:
(a) 28.1 years.
(b) 56.2 years.
(c) 84.3 years.

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 Fissile materials undergo nuclear fission:
(a) Spontaneously.
(b) Under bombardment by not very energetic neutrons.
(c) Under bombardment by very energetic neutrons.

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.

16 Thermal reactors for the production of nuclear power use a fuel of:
(a) Uranium.
(b) Uranium dioxide.
(c) Uranium hexafluoride.

17 The purpose of the control rods in a nuclear reactor is to:
(a) Slow down neutrons.
(b) Generate neutrons.
(c) Absorb neutrons.

18 A fast breeder reactor produces:
(a) Uranium.
(b) Plutonium.
(c) Neptunium.

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.

20 The process that powers the Sun and all stars is:
(a) Fusion.
(b) Fission.
(c) A mixture of fusion and fission.

21 Intermediate-level nuclear waste constitutes how much of the total:
(a) Less than 1%.
(b) Approximately 20%.
(c) Approximately 80%.

Calculations and questions

16.1 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.

16.2 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.

16.3 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).

16.4 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).

16.5 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).

16.6 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 as given in the table.

Time/ min	0	2	4	6	8	10	12	14
Activity/ counts per min	3160	2512	1778	1512	1147	834	603	579

16.7 A sample of radioactive sodium-24 with a half-life of 15.0 h 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?

16.8 What is the specific activity of radium-226, which has a half-life of 1600 y?

16.9 What is the specific activity of plutonium-241, which has a half-life of 14.35 y?

16.10 What is the specific activity of neptunium-233, which has a half-life of 36.2 min?

16.11 A purified sample of a radioactive compound is found to have an activity of 1365 counts per minute at 10 am but only 832 counts per minute at 1 pm. What is the half-life of the sample?

16.12 A 250 mg piece of carbon from an ancient hearth showed 1530 carbon-14 disintegrations in 36 h. 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 y.

16.13 A cellar of dimensions $2 \times 3 \times 2.5$ 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 d.

16.14 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 d.

16.15 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} of 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 y.

16.16 Calculate the binding energy per nucleon for ${}^{4}_{2}\mathrm{He}$. The masses are ${}^{1}_{1}\mathrm{H}$, 1.0078 u; ${}^{1}_{0}\mathrm{n}$, 1.0087 u; ${}^{4}_{2}\mathrm{He}$, 4.0026 u.

16.17 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.

16.18 Calculate the binding energy per nucleon for $^{34}_{16}\mathrm{S}$. The masses are $^{34}_{16}\mathrm{S}$, 33.967865 u, $^{1}_{1}\mathrm{H}$, 1.0078 u; $^{1}_{0}\mathrm{n}$, 1.0087 u.

16.19 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. One mole of helium gas occupies 22.4 dm^3 at 273 K and 1 atm pressure.

16.20 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; $^{102}_{42}\mathrm{Mo}$, 101.91025 u; $^{128}_{50}\mathrm{Sn}$, 127.91047 u; $^{1}_{0}\mathrm{n}$, 1.0087 u.

16.21 Energy generation in the Sun is by way of the reaction of hydrogen to form helium. One reaction is:

$$4\,{}^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He} + 2\mathrm{e}$$

(a) Calculate the energy liberated per fusion reaction. It is estimated that the sun produces approximately 10^{26} J s^{-1}. (b) How many fusion reactions per second are required for this? The masses are $^{1}_{1}\mathrm{H}$, 1.0078 u; $^{4}_{2}\mathrm{He}$, 4.0026 u; $^{0}_{+1}\mathrm{e}$, 5.486×10^{-4} u.

Subject index

Understanding Solids The Science of Materials Index

actinoids *10*, 11, 526, **526**
  metallic radii of 142
activity 528, 529
aerogels 55, 60
aggregations 49–72
  crystalline solids 52–3
    quasicrystals 53–4, *53*
  microstructure 52, **52**, 57–60, *59*
  nanostructures 52, **52**, 55–7
  noncrystalline solids 54–5
  partly crystalline solids 55–6
  point defects 60–8
  scales of structures 52, **52**
  weak chemical bonds 49–52
  see also crystalline solids
alkali metals *10*, 11, 142
alkaline batteries 256, *256*
alkaline earth elements *10*, 11, 142
alpha particles 519, 520–1, **520**
annealing 58, 70
antibonding orbitals 31–2, *32*, 34, 36
antiferroelectrics **341**, 344
  temperature dependence 344–5
antiferromagnetism
  antiferromagnetic materials **375**
  superexchange 381–2, *382*
Arrhenius equation 157–9
  creep rate 312
  diffusion 195, *196*, 202–3, *202*
  and point defects 202–3, *202*, 207–8, *207*
atomic energy levels 11–15
  electron energy levels 11, 13–15
  terms and term symbols 11–13
atomic number (Z) 3, 520, 521
  see also proton numbers
atomic structure 3–15
  atomic energy levels 13–15, *14*
    and spectra 11
    terms and term symbols 11–13
  hydrogen atom 3–7
    orbitals 5–7, *6*
    quantum mechanical description 3–4
  many-electron atoms 7–15
    electron spin and configuration 7–9
    orbital approximation 7
    periodic table 9–11, *10*
atomistic simulation 23–4, 306
attractive energy, between atoms 290, *291*
Aufbau principle 9, 11, 31, 42
auxetic materials 313–14, *314*, *315*

Bardeen-Cooper-Schrieffer (BCS) theory 429
batteries 65–6, 247–8
  dry cells and alkaline primary batteries 255–6, *256*
  fuel cells 260–2, *260*, *261*
  lead-acid rechargeable batteries 257, *257*
  lead-acid secondary batteries 257, *257*
  lithium-air batteries 259, *259*
  lithium-ion primary batteries 256–7, *257*
  lithium-ion rechargeable batteries 257–9, *258*
  types of 255–62
Beer-Lambert law 462
bending forces 283, *283*

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

- bending forces in ceramics 283, *283*
- beta particles 519, **520**, 524, *525*
- Bethe-Slater curve 377, *377*
- binding energy of nuclides 531–2, *532*
- bioluminescence **450**
- black-body radiation 449–50, *450*
- Bloch wall 379, *379*
- body-centred cubic (A2) structure 121–2, *122*
  - tetrahedral and octahedral sites 144, *145*
- Bohr magneton 13
- Bohr radius 5
- bonds 19–45
  - covalent bonding 24–35, *25*, 44
    - and diamond hardness 288–9
    - electronegativity 34
    - molecular orbital theory 30–5
    - molecular orbitals in diatomic molecules 30–3, **32**, *33*
    - molecular orbitals for unlike atoms 33–4, *34*
    - multiple bonds 29–30, *29*, *30*
    - orbital hybridisation 26–9, *26*, *27*, *28*, **29**
    - valence bond theory 24–30
  - ionic bonding 19–24
    - atomistic simulation 23–4
    - lattice energy 21–3, **22**
    - Madelung energy 22
    - repulsive energy 22–3, *23*
  - metallic bonding 35–45
    - bands in ionic and covalent solids 43–4, *43*
    - Brillouin zones 41, 42
    - classical free-electron gas 37–8
    - computations 44–5
    - energy bands 36–42, *36*, *41*
    - Fermi energy and Fermi surface 39–40, *39*, *40*, 146
    - properties of metals 146
    - quantum free-electron gas 38–40, *38*, *39*, *40*
  - of silicate ceramics 147–9, **149**
  - of silicate glasses 155–7
  - in solid lubricants 312–13, *313*
  - weak bonds 49–52
- Bragg's law 115
- Bravais lattices 102, *103*, *104*, *105*, 119
- Brillouin function 372, 375
- Brillouin zones 41, 42
- Brinell hardness number (BHN) 287, *288*
- brittle fracture 295–8, *296*
- brittle materials 284, *284*, 301, *301*
  - Knoop test for 287
- bulk modulus (B) *292*, 293, 295
- Burgers circuit and vector 68–9, *68*, *70*
- calomel electrode 251–2
- CALPHAD 93–4, 229, 534
- capacitors 327–8, *328*, 332
- cathodoluminescence **450**
- cell potential (E_{cell}) 249–50
  - and Gibbs energy 250–1
- ceramics 129, 147–54
  - bend tests on 283, *283*
  - bonding and structure of silicate 147–9, **149**, *150*, *151*
    - clays 148, **149**, 152, *152*, *153*
  - ceramic-matrix composites 178
  - creep in 308, 310
  - glass ceramics 58, 160–1, *160*
  - high-temperature superconductors 430–7, *431*, *432*, *433*, *434*, **435**, *436*, *437*
  - non-silicate ceramics 149–51
  - phase diagram for 92–3, *92*
  - plastic deformation 301, *301*
  - preparation and processing of 152, 154
  - properties of 154
  - as solid lubricants 313
  - strengthening 305–6
  - Young's modulus of 319–20
- chain reaction, nuclear fission 533, *533*, 537
- chalcogens, anions of 20
- chemiluminescence **450**
- Chernobyl 537
- Clausius-Mossotti equation 331, 333
- clays 148, **149**, 152, *152*, *153*, 538
- Clebsch-Gordon rule 12
- coatings, protective 264–5, *265*
  - *see also* films
- cobaltites 384, 390–3, *391*
- Coble creep 310, 311, *311*
- composite materials 129, 177–81
  - cement and concrete 178–81, **179**, *180*
    - hardening of cement 179–80
    - microstructures 181
    - Portland cement 179, **179**, *180*
  - ceramic-matrix composites 178
  - fibre composites 177, *177*, *178*
    - elastic moduli 289, 318–19
  - mechanical properties 317–20
    - elastic modulus of fibre-reinforced composites 289, 318–19
    - elastic modulus of large-particle composites 317–18
    - elastic modulus of a two-phase system 319–20
  - metal-matrix composites 177–8
- compression 281, *282*, 283, *283*
  - and hardness 287, 288
- computation

  - CALPHAD 93–4, 534
    - for shape-memory alloys 229
  - of deformation and fracture 306–7
  - of electronic energy level 14–15
  - of material properties 44–5
  - of molecular orbitals 34–5
- conduction band 44
- contraction, thermal 503–6, *504*, *505*
- Cooper pairs 429, *429*, 430
- corrosion
  - aqueous 262–6, *262*
  - direct/dry air 236–40
  - dissimilar metal corrosion 264–5, *264*
    - copper and iron/steel 264, *264*
    - galvanisation 264–5, *265*
    - tin-plate 265
  - Pourbaix diagrams 270, 273–4, *274*
    - of iron 271–3, *272*
  - reaction of metals with water and aqueous acids 262–4
  - single metal electrochemical corrosion 265–6, *266*
- corrosion fatigue 308
- Coulomb potential 22
- Coulomb repulsion 376
- Coulombic forces 23
- crack propagation 295–8, *297*, 306
  - and toughness 286
- creep 308–12
  - primary creep 308–9, *310*
  - secondary creep 309–12, *310*
    - diffusion creep 310–11
    - temperature dependence 310–11, *310*
  - tertiary creep *310*, 312
- critical stress 296–8
- crystal field theory 369–71, *370*, *371*, *372*, *374*
- crystalline solids 52–3
  - atomic coordinates 118–19
  - atomistic simulation for 23–4
  - crystal lattices 101–2, *102*, **102**
    - Bravais lattices 102, *103*, *104*, *105*, 119
    - reciprocal lattices 112–14, *113*, *114*, 115
  - crystal planes and Miller indices 106–9, *107*, *108*, *109*
  - crystal structures and systems 102–4
    - atomic coordinates 119
    - body-centred cubic (A2) structure 121–2, *122*
    - cubic closed-packed (A1) structure 121, *121*
    - densities 119–21
    - diamond-type (A4) structure 122, *123*, 141
    - graphite-type hexagonal (A9) structure 123, *124*, 141
    - halite (B1) structure 120–1, 123, *124*, 129
    - hexagonal (A3) structure 122, *122*
    - interplanar spacings 112, *113*
    - nomenclature 119
    - spinel (H11) structure 125–6, *125*, 129
    - unit cells 118–19
  - partly crystalline 55
  - quasicrystals 53–4, *53*
- crystallisation
  - in glass ceramics 160
  - of polymers 55, 167–8
- crystallography 101–14
  - determination of crystal structure 114–18
    - electron diffraction 118
    - neutron diffraction 118
    - powder X-ray diffraction 115–18, *116*
    - single crystal X-ray diffraction 114–15
  - directions 106, 110–12, *111*, *112*
  - hexagonal crystals and Miller-Bravais indices 109–10, *110*
  - metals
    - allotropes 140, **141**
    - alloy solid solutions 142–5, *143*, *144*, *145*
    - metallic radii 141–2, *142*
  - parallel directions 110, *111*
  - self-diffusion 191–2, 201–2
  - structural relationships 126–31
    - ionic structures 128–9
    - polyhedral representations 129–30, *130*, *131*
    - sphere packing 126–9, *126*, *127*, *128*, **128**
  - symmetry and crystal classes 104, 106
    - and piezoelectric effect 335, *335*
- cubic close-packed (A1) crystal structure 121, *121*
- cuprate superconductors 426, 430–7, **431**
  - hole doping from charge reservoir layers 436–7
  - lanthanum cuprate 431–2, *432*
  - neodymium cuprate 432–3, *432*
  - perovskite-related structures 434, *434*, **435**, 436, *436*, *437*
  - yttrium barium copper 433–4, *433*
- Curie constant 344, 365, 373
- Curie temperature 344–5, *345*, 366, *366*, *367*, 376
- Curie-Weiss law 344, *345*, *366*, 367, *367*, 376

- d orbitals 5, 7, 9, 11
  - shapes 7, *8*
- Daniell cell 248–9, *248*, 250, 251
- defects in crystalline solids 60–8
  - Frenkel defects 63–4, *63*, **64**, 199, *199*, 203
  - line defects 68–9, *68*, *69*, *70*
  - non-stoichiometric compounds 64–6, *65*

- defects in crystalline solids (*Continued*)
    - point defects 64–6, *65*
  - partial and mixed dislocations 69, *71*
  - planar defects 69–70, *72*
  - and plastic deformation 300
  - point *see* point defects
  - and relative permittivity 333
  - Schottky defects *see* Schottky defects
  - screw dislocations 69, *69*, *70*
  - and strength of solids 297–8
  - substitutional 61, *61*, 147–8
  - volume defects 70, 72, *72*
- deformation
  - brittle and ductile materials 301–2, *301*, *302*
  - brittle fracture 295–8, *296*, 317
  - bulk modulus (*B*) *292*, 293, 295
  - computation of 306–7
  - dislocation movement and plastic deformation 298–300, *299*, *300*
  - elastic deformation
    - and interatomic bonding 289–90
    - of linear elastic materials 289
  - elastic deformation and Young's modulus 289–91, *290*, **292**, 293
  - fracture following plastic deformation 302–4, *304*
  - glass 157–9
  - hardness 287–9, *288*, *289*
  - longitudinal modulus *290*, 292, 293
  - plastic deformation of metals 298, *298*
  - plastic deformation of polymers 285–6, 302, *303*
  - Poisson's ratio 291–2, **292**
  - shear modulus (*G*) 288, *289*, 292–3, *292*, 295
  - strength of materials 281–2
  - strengthening of materials 304–6
  - stress and strain 282–3, *282*, *283*
    - stress and strain curves 283–6, *284*, *285*
  - superelasticity 286–7, *287*
  - toughness and stiffness 286, *286*
  - *see also* plastic deformation
- dehydration 60
- delocalised electrons 42, 123
- density functional theory 45, 307
- density of states function 39
- diamagnetic materials 361–2, *362*, 363, 365, *365*
- diamond-type (A4) structure 122, *123*, 141
- diatomic molecules
  - electron configuration 31–3, **32**
  - molecular orbital energy levels 30–3, *32*, *33*
- dielectrics 327–33, 334, *334*
  - piezoelectric solids 334
  - polarisability 329–30
  - polarisability and relative permittivity 330–1
    - frequency dependence of 331–2, *332*
  - relative permittivity of crystals 332–3, **333**
  - relative permittivity and polarisation 327–9
- diffraction 475–9
  - aperture diffraction 475–6, *476*
  - crystal-like structures 477–8, *478*
  - diffraction gratings 476–7, *477*
  - photonic crystals 478–9, *479*
  - X-ray diffraction 114–18, *116*, *117*, *118*
- diffusion 191–205
  - Arrhenius equation 195–6, *196*, 202–3
  - in creep 310
  - effect of impurities 196, *196*, 198–9, *199*
    - Arrhenius plots 196, *196*, 202, *202*
  - ionic conductivity 205–8
    - relationship with diffusion coefficient 208
  - non-steady-state 194, *194*
  - oxide film growth 237–9, *237*, *238*
    - alloys 240
  - penetration depth 198
  - random walk 197–8, 204
  - self-diffusion 191–3, *192*, 199–201
    - correlation factors 204–5, *204*, **205**
    - crystals 191–2, 201–2
  - in solid-state reactions 240–1, *241*
  - steady-state diffusion 195, *195*
  - temperature variation of diffusion coefficient 195
  - tracer diffusion 192–3, 204–5, *204*
- diodes
  - light-emitting diodes (LEDs) 453–4, *453*, *454*
  - p-n junction diode 413, *414*, 415–16, *415*, *416*
- dipoles 34, 49, 52
- displacive transitions 218–19, *219*, *220*
  - martensitic transformations as 223, 226
- dopants 61, *61*
- doping of ceramics 154
- ductile fracture 302–4, *304*, *305*
  - of gold nanowires 317
- ductile materials 146, 301
  - stress-strain relations 284, *284*, 285
  - tests on 302, 303
- dye-sensitized solar cells 485–6, *486*
- dynamic equilibrium 77, 230
  - *see also* equilibrium

- edge dislocations 68–9, *68*, 300, *300*
- Einstein diffusion equation 197
- elastic moduli 289–95, *290*
  - determination with ultrasonic waves 293–5, *294*
  - *see also* Young's modulus

elastomers 284, *284*, 286, 302, 306
electrochemical cells 247, 264
  titanium from 269–70, *269*
  see also galvanic cells
electrode potentials, standard 249–50, **249**
electroluminescence **450**
electrolysis 266–70
  amount of product 268–9
  electrolytic cells 267, *267*, **267**
  electroplating 267–8, *268*
  titanium via Fray-Farthing-Chen Cambridge process 269–70, *269*
electrolytic cells 267, *267*
electromagnetic spectrum *446*, *447*
  gamma rays 519, **520**, 527
  light waves 445–7, *446*, *448*
  photons *see* photons
electron configurations 8–9
  diatomic molecules 31–3, **32**
  and the periodic table 10–11, *10*
  transition metals 14, *14*
  transuranic elements **526**
electron diffraction 118
electronegativity 34, 49, 50
electronic conductivity
  of ceramics 154
  in graphite 123
  insulators 327, 399–400, *400*
    modification of 418–19
    see also dielectrics
  metals 42, 146, 399–405
  nanostructures and quantum confinement of electrons 423–6
    quantum wells 424–5, *424*, *425*
    quantum wires and quantum dots 425–6
  polymers 161–2
    conducting 420–3, *420*, *421*, *422*, *423*
  range of 327, *328*
  semiconductors 399–400, *400*, 405–16
    characterisation 411–13, *411*, *412*
    extrinsic semiconductors 408–11, *408*
    intrinsic semiconductors 405–7, *405*
    modification of insulators 418–19
    p-n junction diode 413, *414*, 415–16, *415*, *416*
  superconductivity 426–37, *426*
    ceramic high-temperature superconductors 430–7, *431*, *432*, *433*, *434*, **435**, *436*, *437*
    effect of electric currents 428, *429*
    effect of magnetic fields 427–8, *427*, *428*, *429*
    Josephson junctions 430, *430*
    nature of superconductivity 428–9
electrons
  in atoms 3, 520
  as beta particles **520**, 521
  energy levels 4–5, *4*, 14–15
  hydrogen atom
    energy levels 4–5, *4*
    orbital shapes 6, *6*
  in insulators 327
  and ions 19
  in Kröger-Vink notation **66**, 67
  lone pairs 28
  many-electron atoms
    configuration light elements 7–9
    configuration and position in the periodic table 9–11, *10*
    orbital approximation 7
  in metals 36–42
    energies of 38–40, *39*
    and ionisation 19–20
  in molecular orbital theory 31–5
  in orbitals *see* orbitals
  in redox reactions 247
  in valence bond theory 24–30
electroplating 267–8, *268*
  in cells 249
equilibrium 77, 230
  heat treatment of steel 233–6, *234*, *235*
  on Pourbaix diagrams 274
  solidification 230, *231*
    non-equilibrium 230
    systems with eutectic point 231–3, *232*, *233*
eutectic points *84*, 85–6
  solidification 231–3, *232*, *233*
expansion upon heating 501–3, *502*, **503**, *504*
  interatomic potentials 502–5, *502*

f orbitals 5, 7, 9, 11
  in energy band formation 37
face-centred cubic (A1) structure 121, *121*, *127*
Faradays laws of electrolysis 268
fatigue 307–8, *308*, *309*
Fermi energy and Fermi surface 39–40, *39*, *40*, 146
ferrimagnetism
  double exchange 384–5, *385*
  microstructures, domains 378–9, *379*
ferroelectrics 334, *334*, 340–54, **341**
  Curie temperature 344–5, *345*
  domains 341, *342*, 343–4, *343*
  doping and modification of properties 349–51, *350*
  flexoelectricity 353–4, *354*
  hydrogen bonding 345–7, *345*, *346*, *347*

- ferroelectrics (*Continued*)
  - hysteresis 341, *342*, 343
  - nanoparticles 352
  - polar groups 347–8, *348*
  - poling and polycrystalline solids 349
  - relaxor ferroelectrics 329, 351–2, *351*
  - superlattices 352–3, *352*
  - temperature dependence 344–5, *345*
  - thin films 352
  - transition-metal ceramics 348, *349*
- ferromagnetism 374–6
  - Curie temperature 366, *366*, *367*, 376
  - exchange energy 376–8, *377*
  - ferromagnetic and antiferromagnetic materials **375**
  - microstructures
    - domains 378–9, *379*
    - hysteresis 380–1, *380*, *381*
    - hysteresis loops 380–1, *380*, *381*
  - *see also* magnetic materials
- fibre optics
  - attenuation in glass fibres 479–80, *480*
  - dispersion and optical fibre design 480–2, *481*, *482*
  - optical amplification 482–3, *483*
  - optical communications 479
- Fick's first law 195, 200, 201, 239
- Fick's second law 192
- films
  - oxide films 236–40, *238*
  - protective coatings 264–5, *265*
  - protective films 238, 240, 274
  - reflection of light 467–70, *468*, *469*
    - colour in air 469–70, **470**, *471*
    - colour on a substrate 470
    - low- and high-reflectivity coatings 471
    - multiple thin films and dielectric mirrors 471–2, *472*, *473–4*
    - reflectivity in air 469
  - superparamagnetism 386, *387*
  - *see also* thin films
- finite element analysis 307
- first-order phase transitions 217, *218*, 219
- flexoelectricity 353–4, *354*
- fluorescence 450
  - fluorescent lamps 451–2, *451*
  - light-emitting diodes (LEDs) 453–4, *453*, *454*
  - solid-state lasers 450, 454–60
    - neodymium laser 458–9, *458*, *459*
    - ruby laser 456–8, *457*
    - semiconductor lasers 459–60, *460*
- fluxoids 427–8, *428*
- Fraunhofer diffraction 475–6
- Fray-Farthing-Chen Cambridge process 269–70, *269*
- free radical polymerisation 170, 173, *174*
- Frenkel defects 63–4, *63*, **64**
  - and diffusion 199, *199*, 203
  - enthalpy of formation 203, **203**
  - and ionic conductivity 207, 208
- fuel cells 260–2, *260*, *261*
- Fukushima Daiichi 537

- galvanic cells 247–51, *248*
  - cell potential and free energy 250–1
  - chemical analysis 251–5
  - concentration dependence 251
  - and electrolytic compared 267–8
  - overview 247–9, *248*
  - standard electrode potentials 249–50, **249**
  - *see also* electrochemical cells
- galvanisation 264–5, *265*
- gamma rays 519, **520**, 527
- garnet structure ferrites, magnetic properties of 383
- Gladstone-Dale formula 465
- glass transition temperature 156, *156*
  - of amorphous polymers 161
  - of elastomers 173
- glasses 54
  - bonding and structure of silicate glasses 154–61, *155*, **155**
    - deformation 157–9, *157*, *158*, **158**
    - glass ceramics 160–1, *160*
    - strengthened glass 159–60
  - glass ceramics 58, 160–1, *160*
  - metallic glasses 54, 145–6
  - recrystallisation of 72
  - silicate 54, 55, 154–61, **155**
  - stress and strain on 289
  - viscosity of silicate glasses 157–8, *157*, **158**
- Gouy balance 361, *362*
- graphite-type hexagonal (A9) crystal structure 123, *124*
- ground state 4, 8, 9
  - determining 13, *13*
  - in molecular orbital theory 31
- groups on the periodic table *10*, 11

- half-lives **526**, 528–9, **528**, *529*
  - radioactive dating 529–31, *530*
  - transuranic elements **526**
- halite (B1) crystal structure 120–1, 123, *124*, 129
- Hall effect 412–13, *412*
- halogens, anions of 20
- hardness 287–9, 307
  - of thin films 316, 317

Hartree-Fock method 15, 35, 44
heat capacity 495–8, *496*, **496**, *498*
  at phase transitions 497–8, *498*
  theory 496–7
  see also specific heat
heat transfer 498
Herring-Nabarro creep 310–11, *311*, 312
hexagonal (A3) structure 122, *122*
hexagonal close packing of spheres 126–7, *126*, 129
Hooke's law 284–5, 289
Hund's rules 9, 13–14, 31
hybridisation 26–9
  and carbon compounds 29–30, *29*, *30*
    benzene 30, *31*
hydrogen atom 3–7, 9
  covalent bonds between 25, *25*
  electron energy levels 4–5, *4*
  orbitals 5–7, *6*
  quantum mechanical description of 3–4
hydrogen bonds 50, **50**, 52
  ferroelectrics 345–7, *345*, *346*, *347*
hysteresis
  domain growth in ferroelectric crystals 341, *342*, 343–4
  hysteresis loops, ferromagnetic 380–1, *380*, *381*
  in martensitic transformations 225, *226*, 228
  and superelasticity 286–7, *287*

incandescence 449–50, *450*
insulators 327
  electronic conductivity 327, 399–400, *400*, 418–19
  modification 418–19
  see also dielectrics
interference filters 472, *473–4*
ion selective electrodes 253–4, *253*
ions 3, 19–20
  anion packing 128–9, **128**
  diffusion in solid-state reactions 240–1, *241*, 242
  in galvanic cells 247–8, *248*
  ionic bonding 19–24
    lattice energy 21–3, **22**
    Madelung energy 22
    repulsive energy 22–3, *23*
  ionic conductivity 205–8
    relationship with diffusion coefficient 208
  size and shape 20–1
    ionic radii 20–1, *21*
  total potential energy 23, *23*
isotopes 520
  radioisotopes 520, 538

Josephson junctions 430, *430*
Joule heating 507, 511

Kelvin equation 215
Knoop hardness number (KHN) 287–8, *288*
Kröger-Vink notation 66–8, **66**

lamellae
  of phases 232, *232*, *233*
    pearlite 234, *234*
  polymer 167, *168*
Landé g-value 13, 364
Langevin function 372
lanthanoids *10*, 11, 142
  magnetic moments 368–9, **369**
lasers, solid state 450, 454–60
  neodymium laser 458–9, *458*, *459*
  ruby laser 456–8, *457*
  semiconductor lasers 459–60, *460*
lattices 101–2
  Bravais lattices 102, *103*, *104*, *105*, 119
  lattice energy 21–3, 290, *291*
  lattice points 101, 102–3, *103*
  reciprocal lattices 112–14, *113*, *114*, 115
lead-acid batteries 257, *257*
Lennard-Jones potential 23, 50, *51*
light
  interaction with matter 460–1, *461*
  photons 447–9, *449*
  sources
    incandescence 449–50, *450*
    luminescence 450–2, **450**, *451*
  wave nature 445–7, *446*, *448*
light-emitting diodes (LEDs) 453–4, *453*, *454*
linear combination of atomic orbitals (LCAO) 31
lithium-air batteries 259, *259*, 260
lithium-ion primary batteries 256–7, *257*
lithium-ion secondary batteries 257–9, *258*
longitudinal modulus *290*, 292, 293
Lorentz-Lorentz equation 332, 333
lubricants, solid 312–13, *313*
luminescence 450–2, **450**, *451*

Madelung constants 22, **22**
Madelung energy 22
magnetic defects 389–93
  in semiconductors 389–90, *390*
magnetic materials 361–8
  antiferromagnetic materials superexchange 381–2, *382*
  atomic magnetism 363–5

magnetic materials (*Continued*)
  characterisation of 361–2, *362*
  cobaltites 384, 390–3, *391*
  cubic spinel ferrites 382–3
  diamagnetic materials 361–2, *362*, 363, 365, *365*
  ferrimagnetic materials double exchange 384–5, *385*
  ferromagnetic materials
    Curie temperature 366, *366*, *367*
    ferromagnetic and antiferromagnetic materials **375**
  garnet structure ferrites 383
  hexagonal ferrites 383–4, *385*
  magnetic defects 389–93
  magnetic dipoles and magnetic flux 362–3
  magnetocaloric effect in 512–14
  microstructure of ferromagnetic and ferrimagnetic materials
    domains 378–9, *379*
    hysteresis 380–1, *380*, *381*
    hysteresis loop 380–1, *380*, *381*
  nanostructures 385–9
    photoinduced magnetism 387–9, *388*
    small particles and data recording 385–6
    superlattices 386–7, *387*
    superparamagnetism and thin films 386, *387*
  and neutron diffraction 118
  paramagnetic materials 368–74
    magnetic dipole moments 368–9, **369**
  Pauli paramagnetism 373–4, *375*
  temperature dependence of paramagnetic susceptibility 371–3
    splitting of atomic energy levels *374*
  transition metals 364–5, 366, 371
    magnetic moments 368–9, **369**
  see also ferromagnetism
magnetic quantum number (*mi*) 3–4, 5
magnetocaloric effect 512–14
  and adiabatic cooling 512–13, *512*
  giant 513–14, *513*
manganites 384, 390–3, *392*, 419
Mars Curiosity rover 535
martensitic transformations 223–9
  alloys
    Ni-Ti alloys 227–8, *228*
    shape-memory alloys 227, 228–9, *229*
  austenite-martensite 223–6, *224*, *225*, *226*
  Nitinol 286–7, *286*
  stress-induced 286–7
  in zirconia 226–7, *226*, *227*
mass numbers 520, 521
  see also nucleon numbers
metals 139–47
alkali metals *10*, 11, 142
bonding 35–6
  bands in ionic and covalent solids 43–4, *43*
  Brillouin zones 41, 42
  classical free-electron gas 37–8
  and energy bands 36–42
  and properties of metals 146
corrosion *see* corrosion
creep in 308, 310
crystal structures 139–41, *140*
  allotropes 140, **141**
  alloy solid solutions 142–5, *143*, *144*, *145*
  metallic radii 141–2, *142*
dislocation pinning 302, *302*
dissimilar-metal corrosion 264–5, *264*
  copper and iron/steel 264, *264*
  galvanisation 264–5, *265*
  tin-plate 265
ductility 302–3, *304*
  and annealing 58, 70
electronic conductivity 146, 399–405
electroplating of 267–8, *268*
energy 38–40, *38*, *39*
energy bands 36–42, *36*
energy levels 38–40, *38*
fatigue of 307–8, *308*, *309*
Fermi energy and Fermi surface 39–40, *39*, *40*, 146
free electron gas 37–40
glasses 54, 145–6
heat treatment 58
high-temperature oxidation 236–40
  driving force 236
  mechanisms 238–9, *238*
  rate 236–7
metal-matrix composites 177–8
metallic glasses 54, 145–6
nanoparticles of 57
passivation 270, 274, *274*
on the periodic table *10*, 11
plastic deformation 298–300, *298*, *299*, *300*
  and dislocations 68
polycrystalline structure of 58, *59*
Pourbaix diagrams for two valence states 271–3, *272*, *273*
production of 270
  titanium 269–70, *269*
properties 41–2, 146–7
reactions with water and aqueous acids 262–4
refractive index 465
single-metal electrochemical corrosion 265–6, *266*
sintering of 60

  solidification of 58
  strengthening 304–5
  stress on 282–3
  stress-strain relations 284, *284*
  tensile tests on 283
  working 58
  yield point *284*, 285
METGLAS 146
Mie scattering 475, *475*
Miller indices 106–10, *107*, *108*, *109*, *110*
Miller-Bravais indices 109–10, *110*
Mohs scale of hardness 287
molecular dynamics simulations 24, 306
molecular orbital theory 30–5
  covalent solids 43–4
  ionic solids 43
molten carbonate fuel cells (MCFC) 261
Monte Carlo methods 24

nanoindentation 315–16, *316*
nanoparticles 52, 55–7
  optical properties of 387–8, *388*
  of zirconia 226–7, *227*
nanoscale properties of materials 312–17
  auxetic materials 313–14, *314*, *315*
  solid lubricants 312–13, *313*
  thin films 311, 315–17, *316*
nanostructures 52, **52**, 55–7
  optical properties 486–8
nanotubes 317
  carbon 57, *57*
nanowires 317
Néel temperature 367, 382
Nernst equation 251, 252, 254
  stability field of water 270–1
Nernst-Einstein equation 208
neutron diffraction 118
neutrons 520, **520**, 521, 524
noble gases 9, 11, 49
noncrystalline solids 54–5
nonstoichiometric compounds 64–6, *65*
nuclear power
  accidents 537
  binding energy 531–2, *532*
  fast-breeder reactors 535
  fission 532–3, *533*, 535
  fusion 535–6
  solar cycles 536
  space exploration 535
  thermal reactors 533–4, *534*
nuclear reactions
  decays rates 527–9, *529*
    half-lives **526**, 528–9, **528**, *529*
    radioactive dating 529–31, *530*
  equations 520–1, **520**
  nuclear stability 523–4, *525*, 526–7
  radioactive series 521–3
    actinium series 523, *524*
    neptunium series 523, *525*
    thorium series 522, *523*
    uranium series 521–2, *522*
  radioactivity
    artificial radioactivity 524, 526–7
    isotopes and nuclides 520, **520**
  transuranic elements 523, 524, 526–7, **526**
  see also radioactivity
nuclear stability 523–4, 526–7
nuclear waste 536–8
  accidents 537
  storage 537–8
nucleon numbers 520
  on band of stability graph 523–4, *525*
  see also mass numbers
nucleons 520
nuclides 520–1, **520**
  stability of 523–4, *525*

octahedral coordination 129–30, *131*
octahedral sites 128, *128*, 129, 144, *144*, *145*
Ohm's law 399, 403
optical properties
  attenuation 461–2
  colour and appearance 460–2
    Beer-Lambert law 462
    luminous solids 460
    nonluminous solids 460–1
  diffraction 475–9
    aperture diffraction 475–6, *476*
    crystal-like structures 477–8, *478*
    diffraction gratings 476–7, *477*
    photonic crystals 478–9, *479*
  energy conversion
    dye-sensitized solar cells 485–6, *486*
    photoconductivity and photovoltaic solar cells 483–5, *484*
  fibre optics
    attenuation in glass fibre 479–80, *480*
    dispersion and optical fibre design 480–2, *481*, *482*
    optical amplification 482–3, *483*
    optical communications 479
  interaction of light with matter 460–1, *461*
  light-emitting diodes (LEDs) 453–4, *453*, *454*

- optical properties (*Continued*)
  - nanostructures 486–8
    - nanoparticles 487–8, *488*
    - optical properties of quantum wells 486–7, *487*
  - of polymers, side-groups and tacticity 169
  - reflection
    - surfaces 466–7, *466*, *467*
    - thin films 467–70, *468*, *469*, **470**, *471*
  - refraction 462–4, *463*
    - dispersion 465–6, *466*
    - metals and semiconductors 465
    - refractive coefficients 465, **465**
    - refractive index 463–5, **463**, *464*
    - refractive index and structure 463–5
    - total internal reflection 464, *464*
  - scattering
    - Mie scattering 475, *475*
    - Rayleigh scattering 472–3, *474*, 475, *475*
  - solid-state lasers 454–60
    - neodymium laser 458–9, *458*, *459*
    - ruby laser 456–8, *457*
    - semiconductor lasers 459–60, *460*
- orbital angular momentum quantum number (*l*) 3–4, 5
- orbitals 5, **5**
  - antibonding orbitals 31–2, *32*, 34, 36
  - diatomic molecules 30–3, *32*, *33*
  - and energy bands 36–7, *36*, *37*
  - hybrid orbitals 26–9, *26*, *27*, *28*, **29**
  - in hydrogen molecule 25, *25*
  - in many-electron atoms 7, 9
  - in molecular orbital theory 31–3
    - ionic and covalent solids 43–4
  - multiple bonds 29–30, *29*, *30*
  - pi orbitals 31
  - shapes 5–7, *6*, *8*
  - unlike atoms 33–4, *34*
  - in valence bond theory 24–30
- oxidation 247
  - high-temperature, of metals 236–40
  - rate of 236–7
  - and reaction of metals with aqueous acids 262, 263
  - of water 271
  - *see also* redox reactions
- oxide films 236–40, *238*
- oxygen sensors 254–6

- p orbitals 5, 7, 9, 11
  - covalent bonding 25, *25*, *26*
  - shapes 6–7, *6*
- paramagnetism 368–74
  - paramagnetic materials 361–2, *362*, 363, 365–6, 368–74
    - magnetic dipole moments 368–9, **369**
  - Pauli paramagnetism 373–4, *375*
  - superparamagnetism 386, *387*
  - temperature dependence of paramagnetic susceptibility 371–3
    - splitting of atomic energy levels *374*
- passivation 270, 274, *274*
- Pauli Exclusion Principle 9, 12, 25
- Pauli paramagnetism 373–4, *375*
- Peltier effect 506–7, *507*
- periodic table of elements 9–11, *10*
  - electron configurations on 10–11, *10*
  - ionic radii *21*
- peritectic point 87, *89*, 90
- permittivity of free space 327
- permittivity, relative
  - of crystals 332–3, **333**
  - and polarisability 330–1
    - frequency dependence 331–2, *332*
  - and polarisation 327–9
- pH, measurement of 251–3, *252*
- phase diagrams 77
  - CALPHAD for 93–4
  - dynamic equilibrium 230
  - invariant points 78, 85, 87, 89–90, *89*
  - one-component (unary) systems 77–80, *78*, *79*, **82**
  - ternary systems 90–3, *91*, *92*
  - two-component (binary) systems 80–8, *81*, *82*, **82**
    - eutectic 83–7, *84*, *85*, *86*
    - intermediate compounds 87, *87*
    - iron-carbon 88–90, *89*
    - lead-tin 83–7, *84*, *85*, *86*, 231–3, *232*, *233*
    - nickel-copper 81–3, *82*, *83*
    - and solidification 230, *231*, *232*
    - steels and cast irons 89, 233–6, *234*, *235*
- phase transformations 213, 216–17
  - displacive transitions 218–19, *219*, *220*
    - martensitic 223, 226
  - first-order transitions 217, *218*, 219
  - heat capacity at phase transitions 497–8, *498*
  - martensitic transitions 223–9
  - order-disorder transitions 221–3
  - and phase diagrams 230–6
  - reconstructive transitions 219–21, *221*
  - second-order transitions 217–18, *218*, 219
  - sintering 213–16, *214*
- phases 77
  - metastable phases 213
  - on phase diagrams *see* phase diagrams

  phase rule 77
    one-component (unary) systems 79–80
    two-component (binary) systems 81
  two-phase systems, Young's modulus of 319–20
  see also phase diagrams
phonons 497, 500, 501, 508
phosphorescence 450–1
phosphoric acid fuel cells (PAFC) 260–1
photoconductivity 483–5
photoelectrochemical cells 485
photoinduced magnetism 387–9, *388*
photoluminescence 450, **450**
photonic band gap (PBG) 479
photonic crystals 478–9, *479*
photons 4
  gamma ray 519
  light 447–9, *449*
pi bonds 25–6, *26*
  in graphite 123
  and multiple bonds 29–30, *29*, *30*
piezoelectric effect 333–4, *334*
  crystal symmetry 104, 106, 335–6, *335*
  mechanisms 336–7
  polymers 338–9, *339*, *340*
  quartz oscillators 337–8, *337*
  in transducers 294–5, *294*
Pilling-Bedworth ratio 238, 274
planar defects 69–70, *72*
plastic deformation 285
  ceramics 301, *301*
  as creep 308
  dislocation movement 298–300, *299*, *300*
  leading to fracture 302–4, *304*
  in metals 298–300, *298*, *299*, *300*
  polymers 285–6, 302, *303*
point defects 60–8, *60*, *63*, **63**, **64**
  and diffusion 198–9, *198*, *199*
    Arrhenius equation 202–3
  enthalpy of formation 202, 203, *203*, **203**
  and ionic conductivity 207–8
  Kröger-Vink notation 66–8, **66**
  in non-stoichiometric compounds 64–6
  and Seebeck effect 509
  in solid solutions 61–2, *62*
  in variable composition material 120–1
Poisson's ratio 291–2, **292**
  of auxetic materials 313
  of thin films 316
polarisability 329–30
  in crystal structure evaluations 23–4
  electronic polarisability 330, *330*, 331, *331*
  ionic polarisability 330, *330*, 331, *331*
  orientational polarisability 330, *330*
  relative permittivity 330–1
    frequency dependence 331–2, *332*
  space charge polarisability 330
polarisation 329–30
  of dielectrics 328–9, *329*
  in piezoelectrics 333–4, *334*, 335
  spontaneous 334
polymers 54, 129, 161–77, **163**
  in composites 177
  conducting 420–3, *420*, *421*, *422*, *423*
  creep of 308–9
  elastomers 161, 173–5, *174*, *175*, *176*
  elongation of 303–4, *306*
  formation of 162–5, *162*, *164*
  glass transition temperature 301
  microstructures 165–70
    chain structure 167, *167*, *168*
    copolymers 169–70, *170*
    cross-linking 169
    crystal structure 167–8, *168*
    molar mass 166–7, *166*
    tacticity 168–9, *169*, 171
  piezoelectric effect 338–9, *339*, *340*
  plastic deformation 285–6, 302, *303*
  production 170–3
    free-radical polymerisation 170, 173, *174*
    initiation 170–1, 173
    metallocene catalysis 171–3, *173*
    propagation 171, 173
    termination 171, *172*
  properties 175–7
  semicrystalline 302, *303*, 313, *314*
  strengthening 306
  stress on 282–3, 302, *303*
  stress-strain relations 284, *284*, 285–6
  tacticity 168–9, *169*, 171
Portland cement 179, **179**, *180*
Pourbaix diagrams 270–4, *271*
  limitations 274
  metals with two valence states 271–3, *272*, *273*
  passivation and corrosion 270
  tendency for corrosion 273–4, *274*
precipitates 70, 72, *72*, 233, *233*
principal quantum number (*n*) 3, 4
processing of solids 58, 60
proton exchange mechanism fuel cells (PEMFC) 260
proton numbers 3, 520, 521
  on band of stability graph 523–4, *525*

protons 3, 520, **520**
  and nuclear stability 524
  transformation to neutrons 527
pyroelectric effect 334, *334*
  and crystal symmetry 335–6, *335*

quantum dots 56, *56*, 57, 425–6
  optical properties of 387–8, *388*
quantum theory
  and electrons in metals 38–40
  hydrogen atom 3–4
quantum wells 56, *56*, 424–5, *424*, *425*
  optical properties of 486–7, *487*
quantum wires 56, *56*, 57, 425–6
quasicrystals 53–4, *53*

radioactive dating 529–31, *530*
radioactivity 519–24
  artificial radioactive elements 524, 526–7
  see also nuclear reactions
radioisotopes 520, 538
radioluminescence **450**
Rayleigh scattering 472–3, *474*, 475, *475*
reconstructive transitions 219–21, *221*
redox reactions 247
  in aqueous corrosion 262–6
  in batteries 255–9
  in electrolysis 266–70
  in fuel cells 260–2
  and galvanic cells 247–55
  on Pourbaix diagrams 270–4
  see also oxidation
reduction 247
  see also redox reactions
reflection 466–72
  multiple thin films and dielectric mirrors 471–2, *472*, *473–4*
  single thin films 467–70, *468*, *469*
    colour in air 469–70, **470**, *471*
    colour on a substrate 470
    low- and high-reflectivity coatings 471
    reflectivity in air 469
  total internal reflection 464, *464*
refraction 462–4, *463*
  dispersion 465–6, *466*
  refractive index 463–5, **463**, *464*
    of crystals 332, **333**
    metals and semiconductors 465
    refractive coefficients 465, **465**
    structure 464–5
    total internal reflection 464, *464*
refractive index 463–5, **463**, *464*
  of crystals 332, **333**
refrigeration 510–12, *511*
relaxor ferroelectrics 329, 351–2, *351*
Reuss model 320
Rockwell hardness number (RHN) 287, *288*
Russell-Saunders terms 14, *14*

s orbitals 5–6, *6*, 7, 9
  covalent bonding 11, 29, *29*
Schottky defects 62–3, *63*, **63**
  and diffusion 199, *199*, 202
  enthalpy of formation 202, **203**
  and ionic conductivity 207, 208
Schrödinger wave equation 3, 14–15
  and many-electron atoms 7, 8, 35
  metals 38, 40–1
screw dislocations 69, *69*, *70*
secant modulus 285, *285*
second-order phase transitions 217–18, *218*, 219
Seebeck effect 506, 507, *507*, 509
self-diffusion 191–3, *192*
  coefficients 192, **196**
  correlation factors 204–5, *204*, **205**
  in crystals 191–2, 201–2
  and ionic conductivity 208
  in one-dimension 199–201
semiconductors 399–400, *400*, 405–16
  characterisation 411–13, *411*, *412*
  extrinsic semiconductors 408–11, *408*
    carrier concentrations 409–11, *410*
    n-type 409–11, *410*
    p-type 409–11, *410*
  intrinsic semiconductors 405–7, *405*
    band gap values **407**
  magnetic defects in 389–90, *390*
  modification of insulators 418–19
  nanoparticles of 57
  p-n junction diode 413, *414*, 415–16, *415*, *416*
  quantum wells 55–6, *56*
  refractive index 465
  semiconductor lasers 459–60, *460*
shape-memory alloys 227, 228–9, *229*
  superelasticity of 286, *286*
shear forces 281, 292, *292*
shear modulus (*G*) 292–3, *292*
  and hardness 288, *289*
  ultrasonics for 295
shells of electrons 5, 9
Shockley equation 415
sigma bonds 25, *25*, 26

  and multiple bonds 29–30, *29*, *30*, *31*
silver/silver chloride electrode 252, *252*, *253*
sintering 60, 213–16, *214*
  driving force 215–16
  kinetics of neck growth 216, *216*
slip 299, *299*, 300, 301, *301*
  slip planes 300, *300*
solar cells 483–6, *484*, *486*
  dye-sensitized 485–6, *486*
solar cycles 536
solid oxide fuel cell (SOFC) 261–2, *261*
solid solutions 61–2, *62*
  alloys 62, *62*, 142–5
    interstitial solid solutions 62, *62*, 144–5, *144*, *145*
    substitutional solid solutions 61–2, *62*, 143–4, *143*
  on phase diagrams 82
    iron-carbon 88, 89, *89*
    partial solid solutions 84, *84*
    ternary 92, *92*
  precipitate formation 70, 72, *72*
  Vegard's law 115–17, *117*
solid-state reactions 240–2
  spinel formation 240–2, *240*, *241*
    kinetics 241–2
solidification 58
  eutectic systems 231–3, *232*, *233*
  non-equilibrium 230, *231*
  simple binary alloys 230, *231*
sonoluminescence **450**
specific heat 38, 39–40, 217
specific strength 285
spherulites, polymer 167, *168*
spin, electron 7–8, 9
  and orbitals 25, *25*
spin quantum number (*m*s) 7–8, 11, 12
spinel (H11) crystal structure 125–6, *125*, 129
SQUID devices 430, *430*
stiffness 286, *286*, 289
  of thin films 317
strain 282–3, 289
  on composites 318
  and creep 308–12, *310*
  iso-strain condition 318, 319
  and longitudinal modulus 292
  plastic strain 298
  see also stress-strain curves
strengthening of materials 159–60, 304–6
stress 282–3, 289
  on composites 319
  and creep 308–12
  iso-stress condition 319–20
  and martensitic transformations 286–7
  and piezoelectric effect 333–4, *334*, 335
  and strength of solids 296–8
  stress cycles 307, *307*
  see also stress-strain curves
stress-strain curves 283–6, *284*, *285*
  plastic deformation 298, *298*
    fracture following 302–3, *304*
  for polymer elongation *306*
  for steels 302, *302*
  stiffness on 286, *286*
  superelasticity on 286–7, *287*
  toughness on 286, *286*
structural relationships 126–31
superconductors 426–37, *426*
  ceramic high-temperature superconductors 430–7, *431*, *432*, *433*, *434*, **435**, *436*, *437*
  cuprate *see* cuprate superconductors
  effect of electric currents 428, *429*
  effect of magnetic fields 427–8, *427*, *428*, *429*
  Josephson junctions 430, *430*
  nature of superconductivity 428–9
supercooling 79, *79*
superelasticity 286–7, *287*
superlattices 56, *56*, 352–3, *352*, 386–7, *387*
superparamagnetism 386, *387*

tacticity 168–9, *169*, 171
tangent modulus 285, *285*
tension 281, *282*, 283, *283*
terms 12–13
  Russell-Saunders 14, *14*
thermal conductivity 498–501, *499*
  of ceramics 154
  of metals 146
  and microstructure 500–1
  of solids 498–500, *499*
thermal properties
  contraction 503–6, *504*, *505*
  expansion 501–3, *502*, **503**, *504*
    interatomic potentials 502–5, *502*
  of glass-ceramics 160
  heat capacity 495–8, *496*, **496**, *498*
    theory 496–7
  thermoelectric effects 506–12
    charge carriers 508
    coefficients 506–7, *507*
    thermocouples, power generation and refrigeration 509–12, *510*, *511*
thermocouples 146, 509–10, *510*, *511*
thermoelectric effects 506–12, *507*

thermoluminescence **450**
thermopiles 510, *511*
thin films 315–17
  creep in 311
  ferroelectrics 352
  low- and high-reflectivity coatings 471
  mechanical properties of 315–17, *316*
  reflection 467–70, *468*, *469*
  superparamagnetism 386, *387*
  see also films
Thompson effect 507, *507*
time-dependent properties 307–12
  creep 308–12, *310*, *311*
  fatigue 307–8, *308*, *309*
torsion 281, *282*
tracer diffusion 192–3, 204–5, *204*
transgranular (transcrystalline) fracture 295, *296*
transition metals *10*, 11
  electron configurations 14, *14*
  interstitial solid solutions of 144
  magnetic moments 368–9, **369**
  metallic radii of 142
  Pourbaix diagrams 271–3, *272*, *273*
  transuranic 526, **526**
transuranic elements 523, 524, 526–7, **526**
triboluminescence **450**
triple point 78, *78*, 80
twin planes 70, *72*
  martensitic transformations 228–9, *229*

ultrasonic waves 293–5, *294*
unit cell 101–2, **102**, *105*, 118–19

valence band 44
van der Waal's bonding 49–50
vectors 111
  and dipoles 328
Vegard's law 115–17, *117*, 168
Vickers hardness number (VHN) 287, 288, *288*
Vogel-Fulcher-Tamman equation 158
Voigt model 319
volume defects 70, 72, *72*
vulcanisation of rubber 58, 175, *176*

water, stability field of 270–1, *271*
wavefunctions 3, 4, 5, 6, 7
  in computations 15, 35, 45
  electrons in metals 38–41
Weiss domains 378–9, *379*

X-ray diffraction
  powder technique and crystal identification 115–18, *116*, *117*
  single crystal technique 114–15

Young's modulus (*E*) 289–91, *290*, **292**
  of composite materials 317–20
  relationship with other moduli 293
  and stiffness 286
  and strength of solids 296–8
  on stress-strain curves 285, *285*, 286
  thin films 316, 317
  ultrasonics for 295

Zeeman effect 13
zero thermal contraction 505–6
zero thermal expanding solids 505–6
zero-point energy 497

Conversion factors and other relationships

atmosphere (atm):	1(atm) = 101.325 kPa
electron volt (eV):	1(eV) = 96.485 kJ mol^{-1} (eV) × $1.60218 \times 10^{-19} \rightarrow$ (J) (J) × $6.24150 \times 10^{18} \rightarrow$ (eV)
Boltzmann constant k_B	$= 8.61739 \times 10^{-5}$ eV K^{-1}
atomic mass unit (u):	1(u) = 9.31494×10^{8} eV

electron mass = 5.48580×10^{-4} u
neutron mass = 1.00866 u
proton mass = 1.00728 u

calorie (cal):	1(cal) = 4.184 J

$RT = 2.4790$ kJ mol^{-1} at 298.15 K
$RT/F = 25.693$ mV at 298.15 K
$hc = 1.98645 \times 10^{-25}$ J m
$\ln x = 2.302585 \log x$

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

Constants

Quantity	Symbol	Value	Units
Atomic mass unit	u	1.66054×10^{-27}	kg
Avogadro's constant	N_A	6.02214×10^{23}	mol^{-1}
Bohr magneton	μ_B	9.27402×10^{-24}	$\mathrm{J\,T}^{-1}$
Bohr radius	a_0	5.29177×10^{-11}	m
Boltzmann's constant	k_B	1.38066×10^{-23}	$\mathrm{J\,K}^{-1}$
Elementary charge	e	1.60218×10^{-19}	C
Electron mass	m_e	9.10939×10^{-31}	kg
Faraday's constant	F	9.6485×10^{4}	$\mathrm{C\,mol}^{-1}$
Gas constant	R	8.31451	$\mathrm{J\,K}^{-1}\,\mathrm{mol}^{-1}$
Neutron mass	$\mathrm{m_n}$	1.67493×10^{-27}	kg

Quantity	Symbol	Value	Units
Planck's constant	h	6.62608×10^{-34}	J s
Proton mass	m_p	1.67262×10^{-27}	kg
Standard acceleration due to gravity	g	9.80665	$\mathrm{m\,s}^{-2}$
Vacuum permeability	μ_0	$4\pi \times 10^{-7}$	$\mathrm{H\,m}^{-1}$
Vacuum permittivity	ε_0	8.85419×10^{-12}	$\mathrm{F\,m}^{-1}$
Velocity of light	c	2.99792×10^{8}	$\mathrm{m\,s}^{-1}$

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

The elements

Element	Symbol	Atomic number	Molar mass/ g mol^{-1}
Actinium*	Ac	89	(227.028)
Aluminium	Al	13	26.982
Americium*	Am	95	(243.061)
Antimony	Sb	51	121.757
Argon	Ar	18	39.948
Arsenic	As	33	74.922
Astatine*	At	85	(209.987)
Barium	Ba	56	137.327
Berkelium*	Bk	97	(247.070)
Beryllium	Be	4	9.012
Bismuth	Bi	83	208.980
Bohrium*	Bh	107	–
Boron	B	5	10.811
Bromine	Br	35	79.904
Cadmium	Cd	48	112.411
Caesium	Cs	55	132.905
Calcium	Ca	20	40.078
Californium*	Cf	98	(251.080)
Carbon	C	6	12.011
Cerium	Ce	58	140.115
Chlorine	Cl	17	35.453
Chromium	Cr	24	51.996
Cobalt	Co	27	58.933
Copper	Cu	29	63.546
Curium*	Cm	96	(247.070)
Dubnium*	Db	105	–
Dysprosium	Dy	66	162.50
Einsteinium*	Es	99	(252.082)
Erbium	Er	68	167.26
Europium	Eu	63	151.965
Fermium*	Fm	100	(257.095)
Fluorine	F	9	18.998
Francium*	Fr	87	(223.019)
Gadolinium	Gd	64	157.25
Gallium	Ga	31	69.723
Germanium	Ge	32	72.61
Gold	Au	79	196.967
Hafnium	Hf	72	178.49
Hassium*	Hs	108	–
Helium	He	2	4.003
Holmium	Ho	67	164.930
Hydrogen	H	1	1.008
Indium	In	49	114.818
Iodine	I	53	126.904
Iridium	Ir	77	192.22
Iron	Fe	26	55.847
Krypton	Kr	36	83.80
Lanthanum	La	57	138.906
Lawrencium*	Lr	103	(260.105)
Lead	Pb	82	207.2
Lithium	Li	3	6.941
Lutetium	Lu	71	174.967
Magnesium	Mg	12	24.305
Manganese	Mn	25	54.938
Meitnerium*	Mt	109	–
Mendelevium*	Md	101	(258.099)
Mercury	Hg	80	200.59
Molybdenum	Mo	42	95.94
Neodymium	Nd	60	144.24
Neon	Ne	10	20.180
Neptunium*	Np	93	(237.048)
Nickel	Ni	28	58.693

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.

Element	Symbol	Atomic number	Molar mass/ g mol^{-1}
Niobium	Nb	41	92.906
Nitrogen	N	7	14.007
Nobelium*	No	102	(259.101)
Osmium	Os	76	190.23
Oxygen	O	8	15.999
Palladium	Pd	46	106.42
Phosphorus	P	15	30.974
Platinum	Pt	78	195.08
Plutonium*	Pu	94	(244.064)
Polonium*	Po	84	(208.982)
Potassium	K	19	39.098
Praseodymium	Pr	59	140.908
Promethium*	Pm	61	(144.913)
Protoactinium*	Pa	91	(231.036)
Radium*	Ra	88	(226.025)
Radon*	Rn	86	(222.018)
Rhenium	Re	75	186.207
Rhodium	Rh	45	102.906
Rubidium	Rb	37	85.468
Ruthenium	Ru	44	101.07
Rutherfordium*	Rf	104	–
Samarium	Sm	62	150.36
Scandium	Sc	21	44.956
Seaborgium*	Sg	106	–
Selenium	Se	34	78.96
Silicon	Si	14	28.086
Silver	Ag	47	107.868
Sodium	Na	11	22.990
Strontium	Sr	38	87.62
Sulphur	S	16	32.066
Tantalum	Ta	73	180.948
Technetium*	Tc	43	(97.907)
Tellurium	Te	52	127.60
Terbium	Tb	65	158.925
Thallium	Tl	81	204.383
Thorium*	Th	90	232.038
Thulium	Tm	69	168.934
Tin	Sn	50	118.710
Titanium	Ti	22	47.88
Tungsten	W	74	183.84
Uranium*	U	92	238.029
Vanadium	V	23	50.942
Xenon	Xe	54	131.29
Ytterbium	Yb	70	173.04
Yttrium	Y	39	88.906
Zinc	Zn	30	65.39
Zirconium	Zr	40	91.224

The molar mass of most elements is that of a normal terrestrial sample, containing a mixture of isotopes.

*No stable isotope. A value of a molar mass in parenthesis is that of the isotope with the longest half-life. For thorium and uranium, the terrestrial isotopic composition of long-lived isotopes is used.

SI units

Quantity	Name	Symbol	Units
Base units			
Length	metre	m	
Mass	kilogram	kg	
Time	second	s	
Electric current	ampere	A	
Thermodynamic temperature	kelvin	K	
Amount of substance	mole	mol	
Derived units			
Force	newton	N	$\mathrm{kg\,m\,s^{-2}}$
Pressure	pascal	Pa	$\mathrm{N\,m^{-2}}$, $\mathrm{kg\,m^{-1}\,s^{-2}}$
Energy	joule	J	$\mathrm{N\,m}$, $\mathrm{kg\,m^2\,s^{-2}}$
Power	watt	W	$\mathrm{J\,s^{-1}}$, $\mathrm{m^2\,kg\,s^{-3}}$
Electric charge	coulomb	C	$\mathrm{A\,s}$
Electric potential difference	volt	V	$\mathrm{W\,A^{-1}}$, $\mathrm{J\,C^{-1}}$, $\mathrm{J\,A^{-1}\,s^{-1}}$, $\mathrm{m^2\,kg\,s^{-3}\,A^{-1}}$
Capacitance	farad	F	$\mathrm{C\,V^{-1}}$, $\mathrm{s^4\,A^2\,m^{-2}\,kg^{-1}}$
Electric resistance	ohm	Ω	$\mathrm{V\,A^{-1}}$, $\mathrm{m^2\,kg\,s^{-3}\,A^{-2}}$
Electric conductance	siemens	S	Ω^{-1}, $\mathrm{AV^{-1}}$, $\mathrm{s^3\,A^2\,m^{-2}\,kg^{-1}}$
Magnetic flux density	tesla	T	$\mathrm{Wb\,m^{-2}}$, $\mathrm{V\,s\,m^{-2}}$, $\mathrm{kg\,s^{-2}\,A^{-1}}$
Magnetic flux	weber	Wb	$\mathrm{V\,s}$, $\mathrm{m^2\,kg\,s^{-2}\,A^{-1}}$
Magnetic inductance	henry	H	$\mathrm{Wb\,A^{-1}}$, $\mathrm{V\,s\,A^{-1}}$, $\mathrm{m^2\,kg\,s^{-2}\,A^{-2}}$
Frequency	hertz	Hz	$\mathrm{s^{-1}}$
Activity (radionuclide)	becquerel	Bq	$\mathrm{s^{-1}}$

SI prefixes

Submultiple	Prefix	Symbol	Multiple	Prefix	Symbol
10^{-1}	deci	d	10	deca	da
10^{-2}	centi	c	10^2	hecto	h
10^{-3}	milli	m	10^3	kilo	k
10^{-6}	micro	μ	10^6	mega	M
10^{-9}	nano	n	10^9	giga	G
10^{-12}	pico	p	10^{12}	tera	T
10^{-15}	femto	f	10^{15}	peta	P
10^{-18}	atto	a	10^{18}	exa	E

Understanding solids: the science of materials, Second Edition. Richard J.D. Tilley.